Monocular Reconstruction of Neural Face Reflectance Fields

Mallikarjun B R[1] Ayush Tewari[1] Tae-Hyun Oh[2] Tim Weyrich[3]
Bernd Bickel[4] Hans-Peter Seidel[1] Hanspeter Pfister[5]
Wojciech Matusik[6] Mohamed Elgharib[1] Christian Theobalt[1]
[1]Max Planck Institute for Informatics, Saarland Informatics Campus [2]POSTECH
[3]University College London [4]IST Austria [5]Harvard University [6]MIT CSAIL

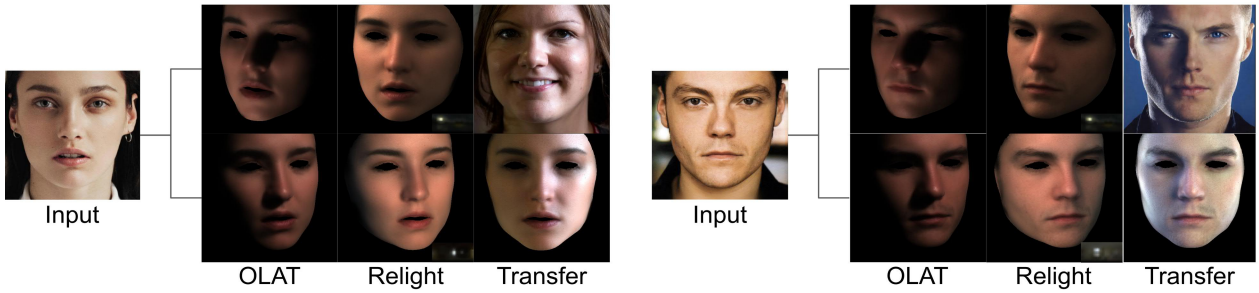

Figure 1. Given a monocular image, our method can synthesize One Light At a Time (OLAT) relit images, relight the face using any environment map (inset), and also transfer the light from another image (top) to the input (bottom). We can model view-dependent effects and can thus generate results in any head pose.

Abstract

The reflectance field of a face describes the reflectance properties responsible for complex lighting effects including diffuse, specular, inter-reflection and self shadowing. Most existing methods for estimating the face reflectance from a monocular image assume faces to be diffuse with very few approaches adding a specular component. This still leaves out important perceptual aspects of reflectance such as higher-order global illumination effects and self-shadowing. We present a new neural representation for face reflectance where we can estimate all components of the reflectance responsible for the final appearance from a monocular image. Instead of modeling each component of the reflectance separately using parametric models, our neural representation allows us to generate a basis set of faces in a geometric deformation-invariant space, parameterized by the input light direction, viewpoint and face geometry. We learn to reconstruct this reflectance field of a face just from a monocular image, which can be used to render the face from any viewpoint in any light condition. Our method is trained on a light-stage dataset, which captures 300 people illuminated with 150 light conditions from 8 viewpoints. We show that our method outperforms existing monocular reflectance reconstruction methods due to better capturing of physical effects, such as sub-surface scattering, specularities, self-shadows and other higher-order effects.

1. Introduction

Monocular face reconstruction (i.e. dense reconstruction of 3D face geometry, reflectance and illumination) has applications in visual effects, telepresence, portrait relighting, facial reenactment, and interactions in virtual environments. It has been an active area of research with tremendous progress in all aspects of reconstruction, including both geometry and reflectance [7]. Our focus is on the reconstruction of the face reflectance, which captures the interaction between the face and scene illumination, playing a very important role in perception. In the literature, one category of methods [10, 37, 40], approximates faces as a Lambertian surface. Many of them use analysis-by-synthesis optimization to estimate the face geometry, spherical harmonics lighting, and diffuse face reflectance; the latter is a stark simplification of the true face reflectance. This type of rep-

resentation fails to capture important specularities and sub-surface effects in face reflectance, which prevents truly photorealistic reconstruction. While some approaches [31, 2] use ambient occlusion and precomputed radiance transfer to model shadows in an inverse rendering framework, they still assume simple reflectance properties of the face, which limits photorealism. Another category of methods [42, 22] reconstruct diffuse and a specular face albedos from an image using machine learning methods. While being more complete, this still leaves out important components of the reflectance, such as self shadowing and other higher-order view-dependent effects and sub-surface effects.

We present the first monocular face reconstruction algorithm that estimates a full face reflectance field, representing both *view direction*- and *light direction*-dependent reflectance properties, from a single face image. We train a CNN that infers the face reflectance field from a single image, and represents it as a basis set of images showing the illuminated face in a normalized space. The images, and thus the reflectance field, are parameterized by the light direction, view direction and face geometry. This is similar to the representations used by image-based techniques for acquiring reflectance fields [6, 25, 33, 8]. However, the crucial difference to our work is that they only capture light-dependent, not view-dependent effects; they can only relight the given input camera view. While Debevec *et al*. [6] can render the face from a different viewpoint, doing so requires an assumption of the BRDF model of the face, and ignores effects such as self-shadowing in the reflectance. Our method goes significantly further by estimating the full reflectance field, including view-dependent effects. We can change both the light source and viewpoint in the image. We do this by jointly estimating the 3D face geometry from the monocular image, and representing the basis images in the UV space [4] of the template face mesh. This also offers other advantages, such as generalization outside of the training data space. Our method is trained on a light-stage dataset, which captures 300 people illuminated with 150 point light sources one at a time, and from 8 viewpoints. While all faces in the dataset are in a neutral expression with mouth closed, *our method still generalizes to real images with general facial expression*, since the training is done in the normalized expression-invariant UV space.

In summary we make the following contributions:

- A monocular method for estimating neural face reflectance fields. We show that the neural reflectance field, directly learned from real data can model complex real phenomena, unlike commonly used parametric reflectance models.

- Generalization to in-the-wild images after training on a light stage dataset. This generalization is obtained by the virtue of explicit use of a canonical space invariant to head pose, identity and expressions, i.e., UV space, as well as training with data synthesized by natural environment maps.

2. Related Work

The literature on face reflectance capture is vast, with methods varying from requiring multi-view multi-illumination images as input [25, 6, 13] to methods which can reconstruct reflectance from a single image. We focus our discussion on monocular methods.

Analysis-based Synthesis Many methods reconstruct face reflectance by solving an analysis-by-synthesis optimization problem minimizing the difference between an estimate and the input image. Since this is an under-constrained problem, methods often make simplifying assumptions, such as the skin having Lambertian reflectance [37, 12, 40, 39, 28]. This allows them to represent lighting using coarse spherical harmonic illumination [27]. Some other methods use a Phong-reflectance assumption [3, 23], which can also model specularities. Specularities using spherical harmonics have also been explored [2, 32]. These representations do not model effects such as sub-surface scattering and self-shadowing, which are important for representing realistic face appearance. Some methods model shadows using precomputed radiance transfer [31] or ambient occlusions [2]. However, due to a Lambertian or simple specular assumption, the final output lacks photorealism. Please refer to a recent survey [7] for more details on these methods.

Supervised Learning Another class of methods are based on supervised learning, where the training data is well-defined, captured from light stages featuring a dome of controlled lighting. At test time, the methods can reconstruct rich reflectance from monocular images. The common representation here is to separate the reflectance into diffuse and specular albedo [42, 22, 24]. Lattas *et al*. [22] estimate the specular albedo and normals using separate networks, using the diffuse albedo and the shape normals as input. However, other complex effects such as self-shadows and view-dependent inter-reflectance cannot be captured. A computationally expensive step of path tracing is performed to simulate shadows at test time.

Relighting Relighting methods capture only the light-dependent component of the reflectance field, without taking view-dependence into account. This makes the problem easier, and several methods can capture complex real world effects. Most approaches are trained on light-stage datasets. Sun *et al*. [33] present an encoder-decoder architecture for manipulating the lighting of an input image. Nestmeyer *et al*. [26] train a model to decouple the input image into physically-based diffuse component, with the non-diffuse components such as specularity and shadows modeled as a

residual. Unlike these approaches, Zhou *et al*. [45] train on monocular data, where the supervision is synthetically generated. Thus, they achieve lower quality compared to methods trained on light stage datasets. Please refer to a recent survey [35] for more details on relighting methods. As mentioned before, these methods cannot capture the view-dependent components of the reflectance field.

Our method, on the other hand, allows us to reconstruct the full reflectance field from a monocular image, thus allowing control over both light and viewpoint. We do not make any assumption about the reflectance properties of the face, and can thus capture all effects including sub-surface scattering, specularities and self-shadows.

3. Method

Our method takes as input an in-the-wild image of a face, a target point light source direction, and the target viewpoint. The output of the network is a mesh of the face lit by a point light from the desired direction which can be rendered from the target viewpoint. At test time, we can render the reconstructed face geometry from any viewpoint and under any environment map by projecting the environment map on a densely sampled point light basis.

3.1. Dataset

Our data-driven approach learns to predict the face reflectance field, which is a function of the face geometry, light sources and camera pose. We train our model on a light-stage dataset [41] consisting of HDR images of 350 identities, captured with 8 cameras distributed in front of the face on a hemisphere (see Fig. 2-b). The light stage also contains 150 point light sources uniformly placed on the sphere surrounding the face. 150 images are captured per person and per camera, with each of the light sources turned on one light at a time (so-called OLAT images). Every subject was captured with neutral expression with eyes and mouth closed. In order to simulate data that looks like in-the-wild images under natural illumination, we relight the light stage data using HDR environment maps. In particular we use a combination of around 205 Laval Outdoor [14] and around 2233 Laval Indoor HDR [9] images, as done in [33]. Our training dataset includes 1000 relit images each, for 300 identities. For each of the relit images, we have a randomly selected OLAT from a random camera view as target image. We use images of 10 identities for validation, and the rest 40 identities for test. Our reflectance field representation operates in a normalised UV space for facial geometry. This enables generalization of our approach to arbitrary face expressions, despite all training data showing neutral face expressions. Note that even though only a partial region in the UV space is visible in the input image, we can still compute results from different view points due to multi-view supervision.

3.2. Reflectance Field Representation

Our reflectance field is a function $\mathcal{R}(\mathcal{G}, \omega_{\mathbf{v}}, \omega_{\mathbf{l}})$, describing the reflectance of a face with geometry $\mathcal{G}$, under viewing direction $\omega_{\mathbf{v}}$ and illuminated by an input point light source direction $\omega_{\mathbf{l}}$, where $\omega_{\mathbf{v}}$ and $\omega_{\mathbf{l}}$ are unit norm vectors. We represent the face geometry using a 3D Morphable Model [3], which includes an identity model $M_{\text{id}} \in \mathbb{R}^{3N \times m_i}$ and an expression model $M_{\text{exp}} \in \mathbb{R}^{3N \times m_e}$, where N is the number of vertices. The vectors of M_{id} and M_{exp} are scaled with their corresponding standard deviations, as in [37]. This representation is well-suited for monocular reconstruction [37, 34, 38]. Mesh vertices are represented by $\mathbf{v}$, $|\mathbf{v}| = 3N$. The final geometry is defined as

$$\mathbf{v}(\alpha, \beta; M_{\text{id}}, M_{\text{exp}}) = \bar{v} + M_{\text{id}}\alpha + M_{\text{exp}}\beta \ .$$

We use the mean mesh $\bar{v}$ from [3]; $\alpha \in \mathbb{R}^{m_i}$ and $\beta \in \mathbb{R}^{m_e}$ are the identity and expression parameters. In monocular reconstruction, it is not possible to separate the effects of head and camera pose. We remove this ambiguity by assuming a camera with fixed extrinsics and intrinsics, and only modeling head pose $\omega_{\mathbf{h}} \in \text{SO}(3)$ as variable. Although the reflectance does not depend on the global translation, we need it to render the face in the correct position in the image. For any vertex $\mathbf{v_i} \in \mathbb{R}^3$, we can compute the camera space coordinates $\mathbf{v_i^c} = \omega_{\mathbf{h}}\mathbf{v_i} + \mathbf{t}$, where $\mathbf{t} \in \mathbb{R}^3$ is the global translation. The complete geometry can be represented as $\mathbf{v^c} \in \mathbb{R}^{3N}$, with $\mathbf{v_i^c}, \forall i \in \{0, \cdots, N-1\}$ stacked together. The reflectance field can then be represented as $\mathcal{R}(\mathbf{v^c}, \omega_{\mathbf{l}})$. We represent the output of this function as a 512×512 RGB image in a normalized UV parametrized space, defined using the template mesh used to represent $\mathbf{v}$, see Fig. 2-a. This allows us to easily generalize to in-the-wild images of varying identity and expression. In addition, it allows us to use image-based 2D CNN architectures, e.g., U-Net architecture [29], since the pixel correspondences required for the skip connections are valid irrespective of target head pose.

3.3. Network Architecture

Our framework consists of two neural networks, the *Geometry Network* and the *Reflectance Network*, as shown in Fig. 2-a. During training, each sample consists of two images, source (I_s) and target (I_t). I_s is an image lit by a natural environment map and I_t is the image of the same person in the same or different pose, under one of the 150 different OLAT lighting condition.

The *Geometry Network* takes both source and target face images as input and reconstructs the 3D face geometry, represented as pose, identity and expression parameters of the 3DMM. Given the reconstructed face geometry of the source image in camera-space coordinates, a differentiable renderer as explained later, produces a source texture map $T_s \in \mathbb{R}^{512 \times 512}$ in the UV space. Our goal is to generate

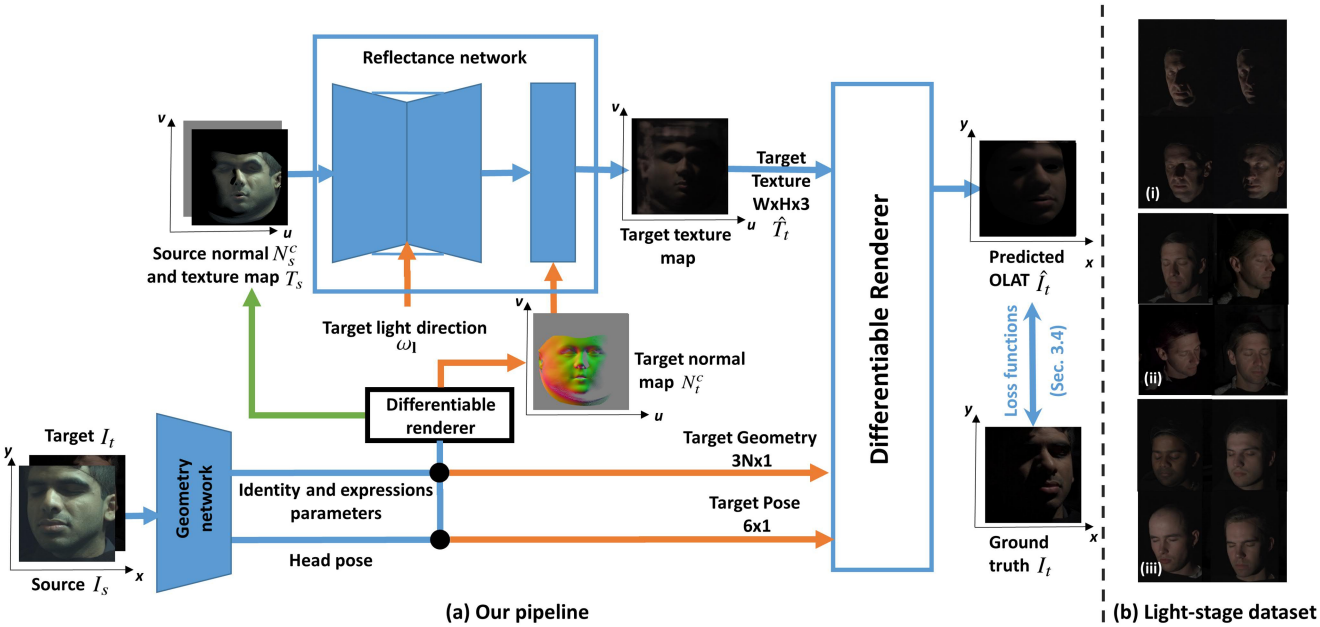

Figure 2. (a) Our approach learns the full face reflectance field by reconstructing an input image with different head poses and point-source lightings (see predicted OLAT image; One Light At a Time). At inference, this allows us to synthesize results with any environment map by linearly combining different OLAT predictions. Our solution is formulated within a normalized UV-space and minimizes for several loss functions through a differentiable renderer. The geometry network processes both the source and target images. At inference, the target normal map is computed by rotating the source normal map based on the desired pose. (b) Our solution is trained with a light-stage dataset which includes 150 lighting conditions (i), with 8 camera-views (ii) and 350 subjects (iii). We use 300 subjects for training, 10 for validation and the rest for test.

an OLAT image in the UV space, lit from a light source with direction $\omega_\mathbf{l}$ and with head pose $\omega_\mathbf{h}$. From the camera space geometries $\mathbf{v}_s^c$ and $\mathbf{v}_t^c$ of the source and target images, we also compute the source and target surface normal maps $N_s^c \in \mathbb{R}^{512\times512}$ and $N_t^c \in \mathbb{R}^{512\times512}$. The *Reflectance Network* takes as input T_s, N_s^c, $\omega_\mathbf{l}$ and N_t^c, as shown in Fig. 2-a, and outputs the target texture map $\hat{T}_t$ in a normalized UV space i.e., every pixel corresponds to a semantically well-defined structure such as eye corner or nose. The network produces an OLAT texture as output, which is rendered using the target geometry and pose to compute the final rendererd image $\hat{I}_t$. At test time, we densely generate OLAT images for each lighting direction, and linearly combine them to relight a new image according to a target environment map.

The *Geometry Network* is based on AlexNet [20, 37], while the *Reflectance Network* is based on a U-Net architecture [30]. The U-Net consists of 8 down and up convolution layers with skip connections and kernels of spatial dimensions 3×3. This is followed by 5 convolutional layers with a stride 1, which takes the output features, as well as the target normal map as input (see Fig. 2-a). Note that the target lighting is fed to the U-Net bottleneck.

Our differentiable renderer renders a 2D image from a 3D face mesh and is similar to Laine *et al*. [21]. We estimate the visible triangles using a z-buffering algorithm. Texture mapping is used to compute the color values. Interpolation (both on the mesh and the texture map) is done using barycentric coordinates. The differentiable renderer offers means for backpropagating the gradients through our normalized representation and thus allows our loss functions to be defined in image space (Sec. 3.4) Our differentiable renderer is implemented as a data-parallel custom TensorFlow layer.

3.4. Loss Functions

We enforce several loss functions to enable the learning of the face reflectance field. Our method concurrently learns to estimate the geometry and head pose as well:

$$\mathcal{L}(I_s, I_t, \omega_\mathbf{l}, \theta_\mathbf{n}) = \lambda_\mathrm{l}\mathcal{L}_\mathrm{l}(I_s, I_t, \theta_\mathbf{n}) + \lambda_\mathrm{r}\mathcal{L}_\mathrm{r}(I_s, I_t, \theta_\mathbf{n}) + \lambda_\mathrm{p}\mathcal{L}_\mathrm{p}(I_s, I_t, \omega_\mathbf{l}, \theta_\mathbf{n}) + \lambda_\mathrm{f}\mathcal{L}_\mathrm{f}(I_s, I_t, \omega_\mathbf{l}, \theta_\mathbf{n}) \quad . \tag{1}$$

Here, $\theta_\mathbf{n}$ are the trainable network parameters for both geometry and reflectance networks, $\mathcal{L}_\mathrm{l}$ is a landmark alignment term, $\mathcal{L}_\mathrm{r}$ is a geometry regularization term, $\mathcal{L}_\mathrm{p}$ is a photometric alignment term and $\mathcal{L}_\mathrm{f}$ is a deep feature alignment term.

Landmark loss This loss provides a strong geometric cue for the 3D geometry reconstruction task:

$$\mathcal{L}_{\text{l}}(I_s, I_t, \theta_{\mathbf{n}}) = \|L(\mathbf{v}_{\mathbf{s}}^{\mathbf{c}}(I_s, \theta_{\mathbf{n}})) - L_s\|_2^2 + \|L(\mathbf{v}_{\mathbf{t}}^{\mathbf{c}}(I_t, \theta_{\mathbf{n}})) - L_t\|_2^2 \ . \quad (2)$$

We use 66 automatically detected landmarks [5] from the source and target images, L_s and L_t as the ground truth. The landmarks from the reconstructions, $L(\mathbf{v}_{\mathbf{s}}^{\mathbf{c}})$ and $L(\mathbf{v}_{\mathbf{t}}^{\mathbf{c}})$ are computed by projecting the annotated landmarks on the mesh to the image plane using the fixed camera parameters. Contour landmarks cannot be fixed since they slide on the mesh, so we compute these landmarks as the closest mesh vertices from the estimated 2D landmarks [36].

Geometry Regularization We use common regularizers [11] used in monocular geometry reconstruction:

$$\mathcal{L}_{\text{r}}(I_s, I_t, \theta_{\mathbf{n}}) = \sum\nolimits_{i=\{s,t\}} \lambda_\alpha \|\alpha_i(I_i, \theta_{\mathbf{n}})\|_2^2 + \lambda_\beta \|\beta_i(I_i, \theta_{\mathbf{n}})\|_2^2 \ . \quad (3)$$

This loss ensures that the final geometry is plausible.

Photometric loss This loss ensures that the final relit images are close to the ground truth:

$$\mathcal{L}_{\text{p}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) = \|M_t(\mathbf{P}) \odot (\hat{I}_t(\mathbf{P}) - I_t)\|_1 \ , \quad (4)$$

where $\odot$ is an element-wise multiplication operator. As explained earlier, the final rendered image $\hat{I}_t$ is parametrized using the source texture map T_s, the normal maps N_s^c and N_t^c, and the light direction ω_{l}. Thus, $\mathbf{P} = (T_s(I_s, \theta_{\mathbf{n}}), N_s^c(I_s, \theta_{\mathbf{n}}), N_t^c(I_t, \theta_{\mathbf{n}}), \omega_{\text{l}})$ We only evaluate the loss in a masked interior face region $M_t(\omega_{\mathbf{h}}(I_t))$, computed using the renderer. The supervision for our UV space reflectance field is thus indirect through the final rendered image using differentiable rasterization.

Feature loss The ℓ_1 loss is known to oversmooth details [15]. To preserve the high-frequency details in the output, we introduce a deep feature loss [16] with two terms:

$$\mathcal{L}_{\text{f}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) = \mathcal{L}_{\text{l}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) + \mathcal{L}_{\text{L}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) \ . \quad (5)$$

To extract features and compute $\mathcal{L}_{\text{l}}$, we use the layers F={`conv1_2`, `conv2_2`, `conv3_3`} of a VGG network V_f pretrained on ImageNet [16] to constrain the output texture map and image as follows:

$$\mathcal{L}_{\text{l}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) = \sum\nolimits_{f \in F} \Big(\big\|V_f(M_t(\mathbf{P}) \odot \hat{I}_t(\mathbf{P})) - V_f(M_t(\mathbf{P}) \odot I_t)\big\|_2^2 + \big\|V_f(\hat{T}_t(\mathbf{P})) - V_f(T_t(I_t, \theta_{\mathbf{n}}))\big\|_2^2 \Big) \ . \quad (6)$$

We use another feature loss from features of a VGG network S_f trained to predict the light direction from images [25]. Specularities depend on light direction, thus the features learned for predicting the latter encode the necessary information:

$$\mathcal{L}_{\text{L}}(I_s, I_t, \omega_{\text{l}}, \theta_{\mathbf{n}}) = \sum\nolimits_{f \in F} \big\|S_f(\hat{T}_t(\mathbf{P})) - S_f(T_t(I_t, \theta_{\mathbf{n}}))\big\|_2^2 \ . \quad (7)$$

	Si-MSE (std. dev.)
Same Pose	0.00070 (σ=0.00059)
Different Pose	0.00084 (σ=0.00088)

Table 1. Reflectance reconstruction errors of our method, under the same and different head poses.

Training We minimize our loss function summed over all samples in the training dataset using mini-batch of size 1 with Adadelta optimizer [43] with a learning rate of 0.05 in order to obtain the network weights $\theta_{\mathbf{n}}$. We implement our method in Tensorflow [1]. We set $\lambda_\alpha = 0.4$, $\lambda_\beta = 0.002$, $\lambda_{\text{l}} = 25$, $\lambda_{\text{p}} = 5$, $\lambda_{\text{r}} = 1$ and $\lambda_{\text{f}} = 1$. To improve generalization of geometry reconstruction, we also include monocular images from FFHQ [18] in our training. FFHQ is only used for the geometry losses, $\mathcal{L}_{\text{l}}$ and $\mathcal{L}_{\text{r}}$, in this case. Overall 20% of our batches are sampled from FFHQ, and the rest from the light-stage data. The *reflectance network* is only trained on the light stage images.

3.5. Relighting

Our network is trained on the light stage data with discrete 150 light directions. However, it allows us to continuously sample light directions at test time, see Sec. 3.2. Since light transport is additive, the final image under any arbitrary environment map can be written as $\sum_{l=0}^{N-1} \lambda_l \hat{I}_t(T_s, N_s^c, N_t^c, \omega_{\text{l}})$. With the abuse of notation, N is the number of light sources, which determines the resolution for the environment map. A larger value of N allows for representing the illumination at a high resolution, at the cost of computational efficiency, since we need a forward pass of the network to compute each $\hat{I}_t$. The weights $\lambda_l \in \mathbb{R}^3$ are color values of the environment map at the pixel corresponding to light direction ω_{l}.

Light Estimation We can also estimate the environment map from an in-the-wild image. Given our reflectance field, we can optimize for the final reconstruction as follows:

$$\lambda^* = \arg\min_{\{\lambda\}} \Big\| \sum\nolimits_{l=0}^{N-1} \lambda_l M_t \odot \hat{I}_t(\omega_{\text{l}}) - M_t \odot I_t \Big\|_2^2 \ , \quad (8)$$

where I_t is an in-the-wild image and $\{\lambda\} = \{\lambda_i | i \in \{0, \cdots, N-1\}\}$. We minimize this term using least-squares. In order to get more detailed reconstruction, we further optimize the light using the feature loss as $\lambda^* = \arg\min_{\{\lambda\}} \|V_f(\hat{T}_t(\omega_{\text{l}})) - V_f(T_t(I_t))\|_2^2$, where T_t is the texture map computed from the input image I_t. We use Adadelta solver [43] to minimize this term and use the solution of Eq. 8 as the initialization.

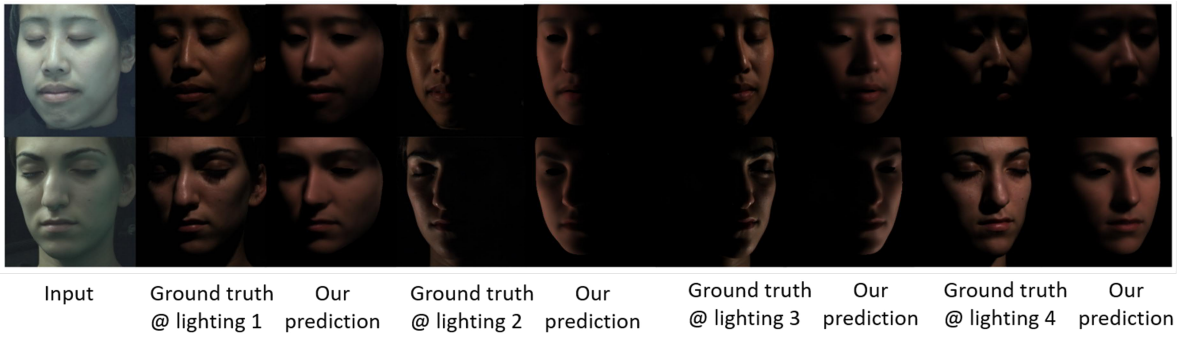

Figure 3. Input image (left) and renderings under different point source lights and with different head poses. Our results resemble ground truth with accurate shadows. Input is taken from the light stage dataset where ground truth is available.

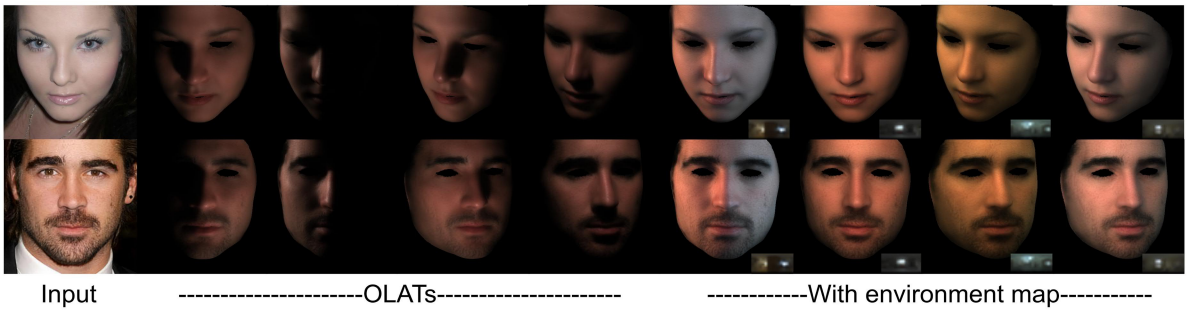

Figure 4. Input image (left) and its OLATs with same pose (2nd and 3rd) and different pose (4th and 5th). Similarly, we have the input image relighted using random environment map (bottow right inset) with same pose (6th and 7th) and different pose(8th and 9th). The scene illumination is identical in each column, allowing us to observe the view dependent effects. For example, see the change of the dominant specularity spots on the nose in the 4th column.

4. Results

We perform experiments on in-the-wild images from CelebA-HQ [17] as well as on our controlled light stage data with ground truth available. Since all images in our training data include an eye-closed expression, we cannot learn the reflectance of open eyes; thus, we remove this region from results. For quantitative evaluations, we use the scale-invariant mean square error (Si-MSE) [45] and face dissimilarity metric (Face dis.). Face dissimilarity is obtained by measuring euclidean distance between features of ground truth and predicted images using a facial recognition tool [19].

4.1. Qualitative Results

We perform several experiments to qualitatively evaluate our approach. Fig. 3 shows results from the light stage test data (identity not included in training), with the corresponding ground truths. We can synthesize different OLATs with different head poses, closely resembling ground-truth. We can capture strong shadows, specularities and sub-surface scattering effects. Fig. 4 additionally shows relighting results on natural images with different environment maps. Here, we add the results of many light sources. Our approach can synthesize results with photorealistic pose-dependent illumination effects, as can be seen in results of faces in different poses. In Fig. 5 we compare our reconstructions with the monocular reconstruction methods of Smith *et al*. [32], Schneider *et al*. [31] and Tewari *et al*. [37]. These methods also estimate the scene illumination. Tewari *et al*. assume faces to be diffuse, Smith *et al*. add a specular component, while Schneider *et al*. use precomputed radiance transfer to model shadows with a diffuse surface assumption. We train the approach of Tewari *et al*. [37] on our training data. Thus, it can be considered as a baseline result where the reflectance model is contrained to be diffuse. Smith *et al*. [32] and Schneider *et al*. [31] are analysis-by-synthesis methods. Our approach clearly produces more photorealistic reconstructions that better capture specularities, subsurface scattering and shadows. The comparison with Smith *et al*. specifically shows the advantages of our representation since their model is also trained on a light stage dataset. Fig. 6 shows further relighting results where the target environment map is computed from another reference image. Results show that our reflectance is well disentangled from illumination, even under strong directional colored illumination. Our results outperform the state-of-the-art both in terms of the quality of reflectance as well as the quality of scene illumination captured. All com-

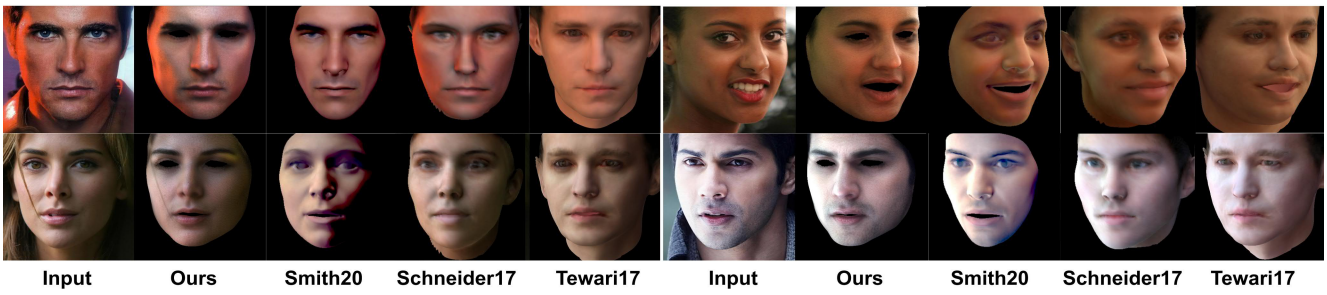

Figure 5. Comparing our face reconstruction to the approaches of Smith *et al*. [32], Schneider *et al*. [31] and Tewari *et al*. [37]. Our approach better captures specularities, sub-surface scattering, hard-shadows and overall produces more photorealistic results.

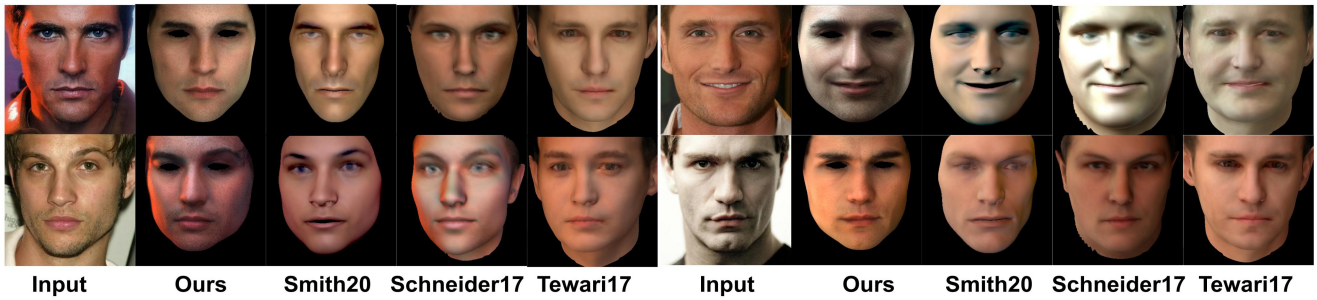

Figure 6. Light transfer results between 2 different images. Each row shows the results of relighting the input image with the light estimated from the other row. Our approach relights an image and edits its head pose, all while maintaining its identity and facial integrity.

peting approaches use a spherical harmonic light assumption, which would be incapable of handling high-frequency light conditions, which often lead to strong shadows. We also project the OLAT lights to the spherical harmonic space and perform a comparison of relighting results. We provide these results in supplemental document. Methods such as [42, 22] do not estimate the scene illumination. This makes it difficult to objectively compare to these approaches, especially since every method assumes a different coordinate system making it difficult to visualize the results under the same lighting. Please refer to the supplemental for qualitative comparisons with Yamaguchi *et al*. [42]. Finally, even though we train our method with 150 light sources, we can synthesize OLAT image for arbitrary continuous light positions. Please refer to the supplemental video for results.

4.2. Quantitative Evaluations

We evaluate our approach quantitatively through a number of experiments. Table 1 summarizes our OLAT reflectance reconstruction results on the light stage data, on a subset of the test set (40 identities, 8 poses). The input images were synthesized using 160 natural environment maps, see Sec. 3.1. A total of 3900 input images are reconstructed with a target pose same as in the input, and 8100 images with a different target pose. Table 1 shows that while our approach produces a lower scale invariant MSE (Si-MSE) for results synthesized with the same pose, the errors only slightly increase with a different pose. Table 2 compares our monocular reconstruction on in-the-wild images with that of different approaches [32, 37, 31]. We use 1774 images from CelebA-HQ [44] as a test set and report the Si-MSE [45] and face identity dissimilarity (Face dis.) [19]. While Si-MSE only looks at pixel-level similarities between images, Face dis. uses a face recognition network to compute distances between facial identity embeddings. The publicly available implementation of the method of Schneider *et al*. [31] cannot reconstruct images with non-neutral expressions. Thus, we do not compare with them on CelebA-HQ. Our approach significantly outperforms existing approaches as reported by the lower Si-MSE error and Face dis. metrics. We also evaluate the quality of reflectance under a "reflectance transfer" operation. Here, we take two images of the same person in different poses and different natural light conditions from the light stage data. We reconstruct the reflectance of both images, and then exchange them before evaluating the reconstruction error. This evaluation tests the quality of reflectance under different poses and light conditions. We also compare to other methods [32, 31, 37] in the same manner. Table 2 shows that our approach outperforms these methods over 2022 images from our test set. As our dataset only contain images with neutral expressions, we also compare with Schneider *et al*.

4.3. Ablative Study

We evaluate the different components of our method using several ablative studies.

	Ours	Smith *et al*. [32]	Schneider *et al*. [31]	Tewari *et al*. [37]
Reconstruction (Si-MSE)	**0.0060** (σ=0.0027)	0.0155 (σ=0.0124)	——×——	0.0073 (σ=0.0037)
Transfer (Si-MSE)	**0.0026** (σ=0.0015)	0.0195 (σ=0.0124)	0.0364 (σ=0.0219)	0.0147 (σ=0.0069)
Reconstruction (Face dis)	**0.5820** (σ=0.0749)	0.6642 (σ=0.0709)	——×——	0.7794 (σ=0.0699)
Transfer (Face dis)	**0.4623** (σ=0.0808)	0.6077 (σ=0.0822)	0.6427 (σ=0.0757)	0.6936 (σ=0.0880)

Table 2. Reconstruction and reflectance transfer errors (in Si-MSE and Face dis with std. dev. σ) of our method, compared with the approaches of Smith *et al*. [32], Schneider *et al*. [31] and Tewari *et al*. [37]. Evaluation is performed on 1774 images from CelebA-HQ [44] for reconstruction, and on 2022 images from our test set for reflectance transfer.

Figure 7. Removing both source and target surface normals from our reflectance learning leads to blurry results and weaker specularities.

	w/o normals (std. dev.)	w/ normals (std. dev.)
Same Pose	0.0011 (σ=0.0009)	**0.0007** (σ=0.0005
Different Pose	0.0012 (σ=0.0011)	**0.0008** (σ=0.0008)

Table 3. Reflectance reconstruction errors of our method, under the same and different input head poses. Removing the normal maps (source and target) from our network design clearly degrades performance.

	Only mean face (std. dev.)	With all (std. dev.)
Si-MSE	0.011 (σ=0.005)	**0.004** (σ=0.002)
Face dis.	0.550 (σ=0.073)	**0.550** (σ=0.080)

Table 4. Reflectance reconstruction errors of our method (in Si-MSE and Face dis with std. dev. σ) with and without face geometry learning. Performance degrades when only the mean face mesh is used (middle column), as opposed to learning the face geometry (last column).

Surface normals We assess the importance of providing surface normals as input in the network. For this, we trained a model without providing the source and target surface normals as input to the reflectance network. The network in this case would not have access to the face geometry and head pose. Table 3 summarizes the results of this experiment. Here, we evaluate OLAT reflectance reconstruction on the light stage data, on a subset of the test set (40 identities, 8 poses). The input images were synthesized using 160 different natural environment maps. A total of 3900 input images are reconstructed with a target pose same as in the input, and 8100 images with a different target pose. This is the same test data used in Table 1. We report Si-MSE for renderings with same and different input pose. Results show that removing normal maps degrades results noticeably, showing that geometry and pose information is important for the task. This reduction in performance is also reflected visually in Fig. 7 where removing surface normals leads to blurry results and weak specularities.

Impact of accurate geometry To assess the importance of accurate geometry in our solution, we train a network which only uses the mean template face mesh. The geometry network here only predicts the head pose, without the identity and expression geometry parameters. We use 130 images from CelebA-HQ [44] as a test set and report the Si-MSE and Face dis. in Table 4. Not learning the face geometry and using a fixed mean mesh instead leads to clear degradation in performance in terms of Si-MSE.

5. Conclusion and Discussion

We presented a method for monocular reconstruction of reflectance fields. Our results are not limited by any parametric reflectance model, and can capture complex real phenomena such as specularities, sub-surface scattering, and self shadowing. While we show results which allow for the estimation of reflectance fields from monocular images for the first time, our method still has some limitations. As mentioned before, we cannot estimate the reflectance of open eyes, since the training dataset does not include such images. However, our method successfully generalizes to in-the-wild images for the visible regions, even for different expressions. Our method in general is limited to the face region, because of geometry reconstruction. With advances in more complete monocular geometry reconstruction, including hair and body, our method should be able to estimate more complete reflectance fields. Although our approach can reconstruct all aspects of the reflectance, strong effects such as specularities and strong shadow boundaries can still be a bit blurred, see Fig. 3. This could be due to inaccuracies in monocular geometry estimation, leading to misalignments between estimate and ground truth during training. Nevertheless, we believe that our method takes an important step towards learning and rendering the full reflectance field of a face.

Acknowledgements: We thank Tarun Yenamandra and Duarte David for helping us with the comparisons. This work was supported by the ERC Consolidator Grant 4DReply (770784). We also acknowledge support from InterDigital.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Oswald Aldrian and William AP Smith. Inverse rendering of faces on a cloudy day. In *Proc. ECCV*, 2012.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, 1999.

[4] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. *Polygon mesh processing*. CRC press, 2010.

[5] 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 *Proc. ICCV*, 2017.

[6] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[7] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future, 2019.

[8] Graham Fyffe. Cosine lobe based relighting from gradient illumination photographs. In *Proc. SIGGRAPH*. 2009.

[9] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *Proc. SIGGRAPH Asia*, 2017.

[10] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *Proc. SIGGRAPH*, 2016.

[11] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics*, 2016.

[12] Pablo Garrido, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt. Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 2016.

[13] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. on Graphics (Proc. SIGGRAPH 2011)*, 2011.

[14] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proc. CVPR*, 2019.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*, 2017.

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. ECCV*, 2016.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, 2019.

[19] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[21] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

[22] 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 *Proc. CVPR*, 2020.

[23] Guannan Li, Chenglei Wu, Carsten Stoll, Yebin Liu, Kiran Varanasi, Qionghai Dai, and Christian Theobalt. Capturing relightable human performances under general uncontrolled illumination. In *Computer Graphics Forum (Proc. Eurographics)*, 2013.

[24] R. Li, K. Bladin, Y. Zhao, C. Chinara, O. Ingraham, P. Xiang, X. Ren, P. Prasad, B. Kishore, J. Xing, and H. Li. Learning formation of physically-based face attributes. In *Proc. CVPR*, 2020.

[25] Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. Deep reflectance fields - high-quality facial reflectance field inference from color gradient illumination. In *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, 2019.

[26] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. Learning physics-guided face relighting under directional light. In *Proc. CVPR*, 2020.

[27] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH*, 2001.

[28] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proc. CVPR*, 2017.

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

[30] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[31] A. Schneider, S. Schönborn, B. Egger, L. Frobeen, and T. Vetter. Efficient global illumination for morphable models. In *Proc. ICCV*, 2017.

[32] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proc. CVPR*, 2020.

[33] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 2019.

[34] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, 2019.

[35] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. *arXiv preprint arXiv:2004.03805*, 2020.

[36] 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 *Proc. CVPR*, 2018.

[37] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017.

[38] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

[39] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. arXiv:1808.09560, 2018.

[40] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proc. CVPR*, July 2017.

[41] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. on Graphics (Proc. SIGGRAPH 2006)*, 2006.

[42] Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics*, 2018.

[43] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.

[44] Dan Zhang and Anna Khoreva. Progressive augmentation of gans. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6249–6259. Curran Associates, Inc., 2019.

[45] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *Proc. ICCV*, 2019.