Monocular Reconstruction of Neural Face Reflectance Fields


1Max Planck Institute for Informatics, Saarland Informatics Campus
2POSTECH
3University College London
4IST Austria
5Harvard University
6MIT 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-
representation 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. WhileDebevec 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
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 OLA T 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 OLA T 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 \( R(G, \omega_v, \omega_l) \), describing the reflectance of a face with geometry \( G \), under viewing direction \( \omega_v \) and illuminated by an input point light source direction \( \omega_l \), where \( \omega_v \) and \( \omega_l \) are unit norm vectors. We represent the face geometry using a 3D Morphable Model [3], which includes an identity model \( M_{id} \in \mathbb{R}^{3N \times m_i} \) and an expression model \( M_{exp} \in \mathbb{R}^{3N \times m_e} \), 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 \( v, |v| = 3N \). The final geometry is defined as

\[
v(\alpha, \beta; M_{id}, M_{exp}) = \bar{v} + M_{id}\alpha + M_{exp}\beta.
\]

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 \( \omega_h \in \text{SO}(3) \) 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 every vertex \( \bar{v}_i \in \mathbb{R}^3 \), we can compute the camera space coordinates \( v^c_i = \omega_h\bar{v}_i + t \), where \( t \in \mathbb{R}^3 \) is the global translation. The complete geometry can be represented as \( v^c \in \mathbb{R}^{3N} \), with \( v^c_i, \forall i \in \{0, \ldots, N-1\} \) stacked together. The reflectance field can then be represented as \( R(v^c, \omega_l) \). We represent the output of this function as a \( 512 \times 512 \) RGB image in a normalized UV parametrized space, defined using the template mesh used to represent \( v \), 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 OLA T 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 OLA T image; One Light At a Time). At inference, this allows us to synthesize results with any environment map by linearly combining different OLA T 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.

Figure 2 (a) Our pipeline

Figure 2 (b) Light-stage dataset
Landmark loss This loss provides a strong geometric cue for the 3D geometry reconstruction task:

$$\mathcal{L}_l(I_s, I_t, \theta_n) = \| L(\mathbf{v}_s^c(I_s, \theta_n)) - L_s \|^2 + \| L(\mathbf{v}_x^c(I_t, \theta_n)) - L_t \|^2$$

(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, $L(\mathbf{v}_s^c)$ and $L(\mathbf{v}_x^c)$ 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:

$$\mathcal{L}_g(I_s, I_t, \theta_n) = \sum_{i=(s,t)} \alpha_i \| \omega_i(I_s, \theta_n) \|^2 + \alpha \| \omega_i(I_t, \theta_n) \|^2$$

(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:

$$\mathcal{L}_p(I_s, I_t, \omega, \theta_n) = \| M_t(P) \odot (\hat{I}_s(P) - I_t) \|_1$$

(4)

where $\odot$ is an element-wise multiplication operator. As explained earlier, the final rendered image $\hat{I}_s$ is parametrized using the source texture map $T_s$, the normal maps $N_s^c$ and $N_s^v$, and the light direction $\omega$. Thus, $P = (T_s(I_s, \theta_n), N_s^c(I_s, \theta_n), N_s^v(I_s, \theta_n), \omega_l)$ We only evaluate the loss in a masked interior face region $M_t(\omega_h(I_t))$, 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 $\ell_1$ loss is known to be smooth details [15]. To preserve the high-frequency details in the output, we introduce a deep feature loss [16] with two terms:

$$\mathcal{L}_f(I_s, I_t, \omega, \theta_n) = \mathcal{L}_f(I_s, I_t, \omega, \theta_n) + \mathcal{L}_f(I_s, I_t, \omega, \theta_n).$$

(5)

To extract features and compute $\mathcal{L}_f$, we use the layers $F=\{\text{conv1.2, conv2.2, conv3.3}\}$ of a VGG network $V_f$ pretrained on ImageNet [16] to constrain the output texture map and image as follows:

$$\mathcal{L}_f(I_s, I_t, \omega, \theta_n) = \sum_{f \in F} \left( \| V_f(M_t(P) \odot \hat{I}_s(P)) - V_f(M_t(P) \odot I_t) \|^2 + \| V_f(T_f(P)) - V_f(T_f(I_s, \theta_n)) \|^2 \right).$$

(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:

$$\mathcal{L}_l(I_s, I_t, \omega, \theta_n) = \sum_{f \in P} \| S_f(T_f(P)) - S_f(T_f(I_s, \theta_n)) \|^2$$

(7)

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 $\theta_n$. We implement our method in Tensorflow [1]. We set $\lambda_\alpha = 0.4$, $\lambda_\beta = 0.002$, $\lambda_l = 25$, $\lambda_p = 5$, $\lambda_s = 1$ and $\lambda_r = 1$. 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, $\mathcal{L}_l$ and $\mathcal{L}_g$ 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 $\sum_{l=0}^{N-1} \lambda_l \hat{I}_l(T_s, N_s^c, N_s^v, \omega_l)$. 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 $\hat{I}_l$. The weights $\lambda_l \in \mathbb{R}^3$ are color values of the environment map at the pixel corresponding to light direction $\omega_l$.

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:

$$\lambda^* = \arg \min_{\{\lambda\}} \sum_{l=0}^{N-1} \lambda_l M_t \odot \hat{I}_l(\omega_l) - M_t \odot I_t \|_2^2$$

(8)

where $I_t$ is an in-the-wild image and $\{\lambda\} = \{\lambda_l \mid i \in \{0, \cdots, N-1\}\}$. We minimize this term using least-squares. In order to get more detailed reconstruction, we further optimize the light using the feature loss as $\lambda^* = \arg \min_{\{\lambda\}} \| V_f(T_f(\omega_l)) - V_f(T_f(I_t)) \|^2$, where $T_f$ 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.

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

<table>
<thead>
<tr>
<th></th>
<th>Si-MSE (std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Pose</td>
<td>0.00070 (σ=0.00059)</td>
</tr>
<tr>
<td>Different Pose</td>
<td>0.00084 (σ=0.00088)</td>
</tr>
</tbody>
</table>
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 constrained 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-
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 OLA T 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 OLA T 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 OLA T 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.

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.
Table 2. Reconstruction and reflectance transfer errors (in Si-MSE and Face dis with std. dev. \( \sigma \)) 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.

![Image](50x561 to 545x624)

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Ours</th>
<th>Smith et al. [32]</th>
<th>Schneider et al. [31]</th>
<th>Tewari et al. [37]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reconstruction (Si-MSE)</td>
<td>(0.0060 (\sigma=0.0027))</td>
<td>(0.0155 (\sigma=0.0124))</td>
<td>(---\times---)</td>
<td>(0.0073 (\sigma=0.0037))</td>
</tr>
<tr>
<td>Transfer (Si-MSE)</td>
<td>(0.0026 (\sigma=0.0015))</td>
<td>(0.0195 (\sigma=0.0124))</td>
<td>(0.0364 (\sigma=0.0219))</td>
<td>(0.0147 (\sigma=0.0069))</td>
</tr>
<tr>
<td>Reconstruction (Face dis)</td>
<td>(0.5820 (\sigma=0.0749))</td>
<td>(0.6642 (\sigma=0.0709))</td>
<td>(---\times---)</td>
<td>(0.7794 (\sigma=0.0699))</td>
</tr>
<tr>
<td>Transfer (Face dis)</td>
<td>(0.4623 (\sigma=0.0808))</td>
<td>(0.6077 (\sigma=0.0822))</td>
<td>(0.6427 (\sigma=0.0757))</td>
<td>(0.6936 (\sigma=0.0880))</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pose</th>
<th>w/o normals (std. dev.)</th>
<th>w/ normals (std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Pose</td>
<td>0.0011 ((\sigma=0.0009))</td>
<td>0.0007 ((\sigma=0.0005))</td>
</tr>
<tr>
<td>Different Pose</td>
<td>0.0012 ((\sigma=0.0011))</td>
<td>0.0008 ((\sigma=0.0008))</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>Only mean face (std. dev.)</th>
<th>With all (std. dev.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si-MSE</td>
<td>0.011 ((\sigma=0.005))</td>
<td>0.004 ((\sigma=0.002))</td>
</tr>
<tr>
<td>Face dis.</td>
<td>0.550 ((\sigma=0.073))</td>
<td>0.550 ((\sigma=0.080))</td>
</tr>
</tbody>
</table>

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


