

Monocular Reconstruction of Neural Face Reflectance Fields

Mallikarjun B R¹ Ayush Tewari¹ Tae-Hyun Oh² Tim Weyrich³
 Bernd Bickel⁴ Hans-Peter Seidel¹ Hanspeter Pfister⁵
 Wojciech Matusik⁶ Mohamed Elgharib¹ Christian Theobalt¹

¹ Max Planck Institute for Informatics, Saarland Informatics Campus ² POSTECH
³ University College London ⁴ IST Austria ⁵ Harvard University ⁶ MIT CSAIL

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 as higher-order global illumination effects and self-shadowing are not modeled. 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 single 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 training dataset, which captures 300 people illuminated with 150 light conditions from 8 viewpoints. We show that our method outperforms existing monocular reflectance reconstruction methods, in terms of photorealism due to better capturing of physical primitives, 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 vast 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. 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 [9, 32, 35], approximates faces as a Lambertian surface. Many of them use analysis-by-synthesis optimization to estimate face geometry, spherical harmonics lighting, and diffuse face reflectance; the latter is a stark simplification of true face reflectance. This type of representation fails to capture important specularities and sub-surface effects in face reflectance, which prevents truly photorealistic reconstruction. While some approaches [27, 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 limit photorealism. Another category of methods [37, 19] reconstruct diffuse and a specular face albedo 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 light direction, view direction and face geometry. This is similar to the representations used by image-based techniques for acquiring reflectance fields [5, 21, 29, 7]. 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 [5] 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 signif-

icantly 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 also 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. 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 a deep face reflectance field. Our method is trained with a large set of light-stage data. Reconstructed faces model complex pose/view- and scene illumination dependent appearance, beyond diffuse and specular reflectance.
- A new deep representation for face reflectance fields, allowing us to generalize to real-world images after training on a light stage dataset. This generalization is obtained by virtue of the explicit use of a canonical space invariant to pose, identity and expression, 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 [21, 5, 11] 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 model and single input image. Since this is an under-constrained problem, methods often make simplifying assumptions, such as the skin having Lambertian reflectance [32, 10, 35, 34, 23]. This allows them to represent lighting using coarse spherical harmonic illumination [22]. Some other methods use a Phong-reflectance assumption [3, 20], which can also model specularities. Specularities using spherical harmonics have also been explored [2, 28]. These representations do not model effects such as sub-surface scattering and self-shadowing, which are important for face appearance. Some methods model shadows using precomputed radiance transfer [27] or ambient occlusions [2]. However, due to a Lambertian or simple specular assumption, the final output lacks photo-realism. Please refer to a recent survey [6] for more details on these methods.

Supervised Learning Another class of methods are based on supervised learning. Here, 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 [37, 19]. In Yamaguchi *et al.* [37] the solution also infers a high-frequency displacement map representing mesoscopic surface details. In Lattas *et al.* [19] separate networks estimate the specular albedo and normals from the diffuse albedo and the 3DMM shape normals. 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.

Image-based Methods Here we review supervised methods that train either using light stage training data [29, 21] or just on monocular images [40]. Meka *et al.* [21] show that two spherical color gradient images are capable of modeling a 4D reflectance field. This includes high-frequency details and specular reflections. Sun *et al.* [29] present an encoder-decoder architecture for manipulating the lighting of an input. The network is trained on light-stage data of 18 subjects captured under several light sources. Zhou *et al.* [40] utilize a spherical harmonic representation of lighting to synthesize large-scale training data for a relighting network. While image-based approaches provide a range of capabilities, they directly work on input images and not in 3D. This does not allow to capture the full reflectance field; these methods can only relight a given image, but not change the viewpoint.

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 assumptions 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 [36] consisting of HDR images of 350

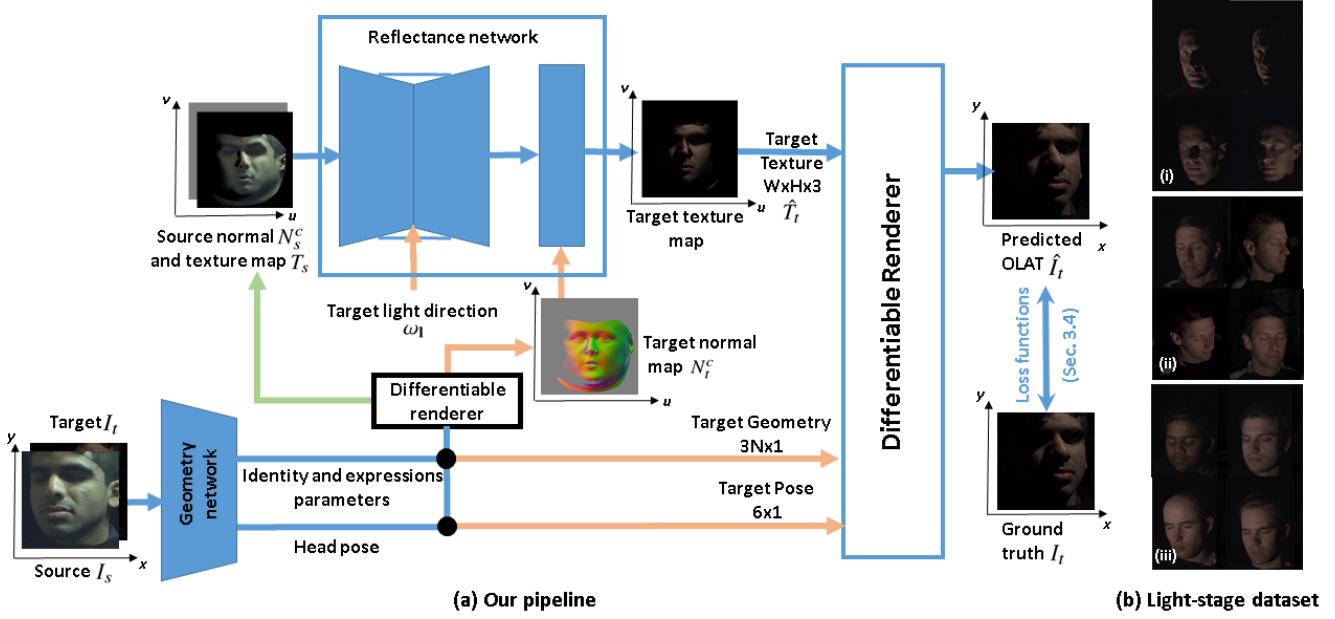


Figure 1. (a) Our approach learns the full face reflectance field by reconstructing an input image with different head poses and point-source lightnings (see predicted OLAT). At inference this allows us to synthesize results with any environment map by linearly combining different OLAT predictions. Our solution is formulated within a normalized UV-space and minimizes for several loss functions through a differentiable renderer. Note the geometry network processes both the source and target images. (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.

identities, captured with 8 cameras distributed in front of the face on a hemisphere (see Fig. 1-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 look 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 [12] and around 2233 Laval Indoor HDR [8] images, as done in [29]. 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.

3.2. Reflectance Field Representation

Our reflectance field is a function $\mathcal{R}(\mathcal{G}, \omega_v, \omega_l)$, describing the reflectance of a face with geometry \mathcal{G} , under viewing direction ω_v and illuminated by an input point light

source direction ω_l , where ω_v and ω_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 [32]. This representation is well-suited for monocular reconstruction [32, 30, 33]. Mesh vertices are represented by $\mathbf{v}, |\mathbf{v}| = 3N$. The final geometry is defined as

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

We use the mean mesh $\bar{\mathbf{v}}$ 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 $\omega_h \in SO(3)$ 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 $\mathbf{v}_i^c = \omega_h \mathbf{v}_i + \mathbf{t}$, 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 $\mathbf{v}_i^c, \forall i \in \{0, \dots, N\}$ stacked together. The reflectance field can then be represented as $\mathcal{R}(\mathbf{v}^c, \omega_l)$. 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. 1-a. This is a pose, expression and identity deformation-invariant representation, allowing us to easily generalize to in-the-wild images of varying identity and expression. In addition, it allows us to use a U-Net architecture [24], 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. 1-a. Each sample in our training 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 produces a source texture map $T_s \in \mathbb{R}^{512 \times 512}$ in the UV space. Our goal is to generate an OLAT image in the UV space, lit from a light source with direction ω_1 and with head pose ω_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 $N_s^c \in \mathbb{R}^{512 \times 512}$ and $N_t^c \in \mathbb{R}^{512 \times 512}$. The *Reflectance Network* takes as input T_s , N_s^c , ω_1 and N_t^c , as shown in Fig. 1-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 rendered image \hat{I}_t .

The *Geometry Network* is based on AlexNet [18, 32], while the *Reflectance Network* is based on a U-Net architecture [25]. 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. 1-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. 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.

$$\begin{aligned} \mathcal{L}(I_s, I_t, \omega_1, \theta_n) = & \lambda_l \mathcal{L}_l(I_s, I_t, \theta_n) + \lambda_r \mathcal{L}_r(I_s, I_t, \theta_n) + \\ & \lambda_p \mathcal{L}_p(I_s, I_t, \omega_1, \theta_n) + \lambda_f \mathcal{L}_f(I_s, I_t, \omega_1, \theta_n) . \end{aligned} \quad (1)$$

Here, θ_n are the trainable network parameters for both geometry and reflectance networks, \mathcal{L}_l is a landmark alignment term, \mathcal{L}_r is a geometry regularization term, \mathcal{L}_p is a photometric alignment term and \mathcal{L}_f is a deep feature alignment term.

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

$$\begin{aligned} \mathcal{L}_l(I_s, I_t, \theta_n) = & \|L(\mathbf{v}_s^c(I_s, \theta_n)) - L_s\|_2^2 + \\ & \|L(\mathbf{v}_t^c(I_t, \theta_n)) - L_t\|_2^2 . \end{aligned} \quad (2)$$

We use 66 automatically detected landmarks [26] 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 $L(\mathbf{v}_t^c)$ 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 [31].

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

$$\mathcal{L}_r(I_s, I_t, \theta_n) = \sum_{i=\{s,t\}} \lambda_\alpha \|\alpha_i(I_i, \theta_n)\|_2^2 + \lambda_\beta \|\beta_i(I_i, \theta_n)\|_2^2 . \quad (3)$$

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.

$$\mathcal{L}_p(I_s, I_t, \omega_1, \theta_n) = \|M_t(\mathbf{P}) \odot (\hat{I}_t(\mathbf{P}) - I_t)\|_1 . \quad (4)$$

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_1 . Thus, $\mathbf{P} = (T_s(I_s, \theta_n), N_s^c(I_s, \theta_n), N_t^c(I_t, \theta_n), \omega_1)$ We only evaluate the loss in a masked interior face region $M_t(\omega_h(I_t))$, computed using the renderer. \odot is an element-wise multiplication operator. 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 [13]. To preserve the high-frequency details in the output, we introduce a deep feature loss [14] with two terms.

$$\mathcal{L}_f(I_s, I_t, \omega_1, \theta_n) = \mathcal{L}_l(I_s, I_t, \omega_1, \theta_n) + \mathcal{L}_L(I_s, I_t, \omega_1, \theta_n) . \quad (5)$$

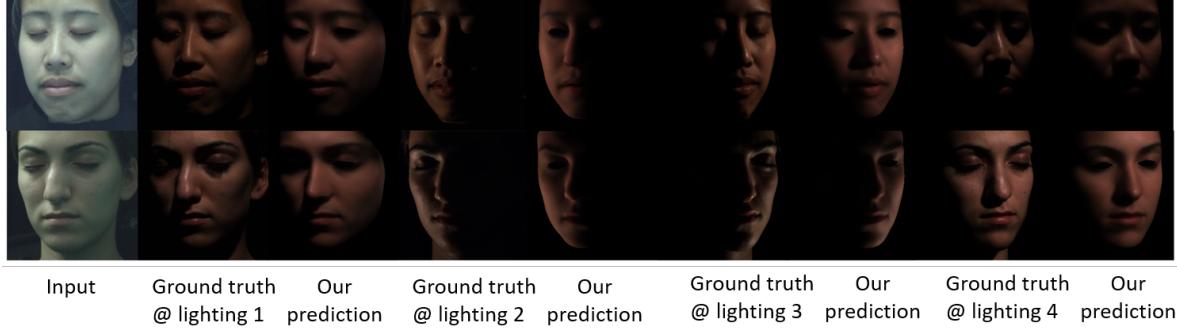


Figure 2. Input image (left) and renderings under different point source lights and with different head poses. Our results resemble ground truth with accurate shadows. Input is taken from the light stage data-set where ground truth is available.

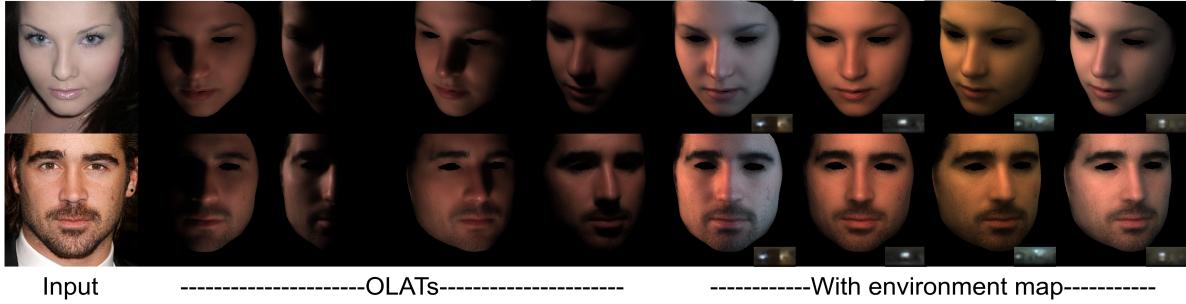


Figure 3. Input image (left) and its OLATs with same pose (2nd and 3rd) and different pose (4th and 5th). Similarly, we have the input image relighted using random environment map (bottom right inset) with same pose (6th and 7th) and different pose(8th and 9th). The scene illumination is identical in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

To extract features and compute \mathcal{L}_1 , we use the layers $F = \{\text{conv1_2}, \text{conv2_2}, \text{conv3_3}\}$ of a VGG network V_f pretrained on ImageNet [14] to constrain the output texture map and image as follows:

$$\begin{aligned} \mathcal{L}_1(I_s, I_t, \omega_1, \theta_n) = & \\ \sum_{f \in F} & \left(\|V_f(M_t(\mathbf{P}) \odot \hat{I}_t(\mathbf{P})) - V_f(M_t(\mathbf{P}) \odot I_t)\|_2^2 \right. \\ & \left. + \|V_f(\hat{T}_t(\mathbf{P})) - V_f(T_t(I_t, \theta_n))\|_2^2 \right). \end{aligned} \quad (6)$$

We use another feature loss from features of a VGG network S_f trained to predict the light direction from images [21]. Specularities depend on light direction, thus the features learned for predicting the latter encode the necessary information:

$$\mathcal{L}_L(I_s, I_t, \omega_1, \theta_n) = \sum_{f \in F} \|S_f(\hat{T}_t(\mathbf{P})) - S_f(T_t(I_t, \theta_n))\|_2^2. \quad (7)$$

Training We minimize our loss function summed over all samples in the training dataset using mini-batch of size 1 with Adadelta Optimizer [38] with a learning rate of 0.05 in order to obtain the network weights θ_n . We implement our method in Tensorflow [1]. We set $\lambda_\alpha = 0.4$, $\lambda_\beta = 0.002$, $\lambda_l = 25$, $\lambda_p = 5$, $\lambda_r = 1$ and $\lambda_f = 1$. To improve

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

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

generalization of geometry reconstruction, we also include monocular images from FFHQ [16] in our training. FFHQ is only used for the geometry losses, \mathcal{L}_l and \mathcal{L}_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. Given an input image, we can also estimate the scene illumination using OLAT predictions of these 150 light directions. Since light transport is additive, the final image under any arbitrary environment map can be written as $\sum_{l=0}^N \lambda_l \hat{I}_l(T_s, N_s^c, N_t^c, \omega_1)$. N is the number of light sources, which determines the resolution for the environ-

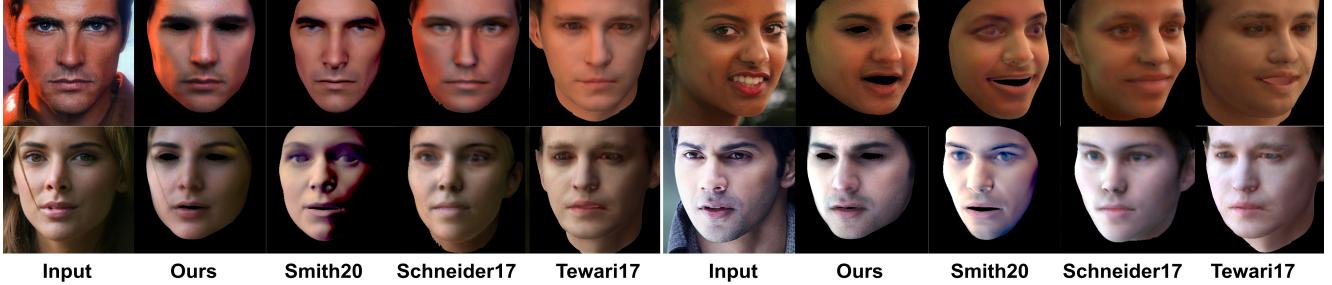


Figure 4. Comparing our face reconstruction to the approaches of Smith *et al.* [28], Schneider *et al.* [27] and Tewari *et al.* [32]. Our approach better captures specularities, sub-surface scattering, hard-shadows and overall produces more photorealistic results.

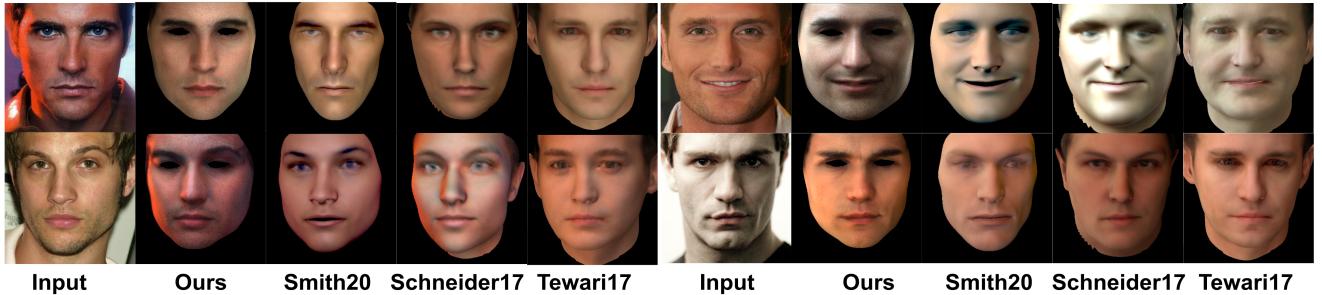


Figure 5. Light transfer results between 2 different images. Each row shows the results of relighting the input image with the light estimated from the other row. Our approach relights an image and edits its head pose, all while maintaining its identity and facial integrity.

ment. 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:

$$\lambda^* = \arg \min_{\{\lambda\}} \left\| \sum_{l=0}^{N-1} \lambda_l M_t \odot \hat{I}_t(\omega_l) - M_t \odot I_t \right\|_2^2. \quad (8)$$

Here, I_t is an in-the-wild image and $\{\lambda\} = \{\lambda_i | i \in \{0, \dots, N-1\}\}$. We minimize this term using least-squares. In order to get more detailed reconstruction, we further optimize the light using the feature loss as $\lambda^* = \arg \min_{\{\lambda\}} \|V_f(\hat{T}_t(\omega_l)) - V_f(T_t(I_t))\|_2^2$, where T_t is the texture map computed from the input image I_t . We use Adadelta solver [38] to minimize this term and use the solution of Eq. 8 as the initialization.

4. Results

We perform experiments on in-the-wild images from CelebA-HQ [15] as well as on 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, and we remove this region from results. For quantitative evaluations, we use the

scale-invariant mean square error (Si-MSE) [40] 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 [17].

4.1. Qualitative Results

We perform several experiments to qualitatively evaluate our approach. Fig. 2 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. 3 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. 4 we compare our reconstructions with the monocular reconstruction methods of Smith *et al.* [28], Schneider *et al.* [27] and Tewari *et al.* [32]. 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.* [32] on our training data. Thus, it can be considered as a baseline result where the reflectance model is constrained

	Ours	Smith <i>et al.</i> [28]	Schneider <i>et al.</i> [27]	Tewari <i>et al.</i> [32]
Reconstruction (Si-MSE)	0.004 ($\sigma=0.002$)	0.017 ($\sigma=0.015$)	0.010 ($\sigma=0.008$)	0.007 ($\sigma=0.003$)
Transfer (Si-MSE)	0.002 ($\sigma=0.001$)	0.018 ($\sigma=0.011$)	0.007 ($\sigma=0.005$)	0.003 ($\sigma=0.001$)
Reconstruction (Face dis)	0.550 ($\sigma=0.080$)	0.650 ($\sigma=0.076$)	0.685 ($\sigma=0.067$)	0.762 ($\sigma=0.075$)
Transfer (Face dis)	0.482 ($\sigma=0.084$)	0.605 ($\sigma=0.092$)	0.644 ($\sigma=0.077$)	0.689 ($\sigma=0.076$)

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.* [28], Schneider *et al.* [27] and Tewari *et al.* [32]. Evaluation is performed on 130 images from CelebA-HQ [39] for reconstruction, and on 86 images from our test set for reflectance transfer.

	Without normal maps (std. dev.)	With normal maps (std. dev.)
Same Pose	0.00113 ($\sigma=0.00093$)	0.00070 ($\sigma=0.00059$)
Different Pose	0.00126 ($\sigma=0.00116$)	0.00084 ($\sigma=0.00088$)

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.

to be diffuse. Smith *et al.* [28] and Schneider *et al.* [27] 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. 5 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 competing approaches use a spherical harmonic light assumption, which would be incapable of handling high-frequency light conditions, which often lead to strong shadows. Methods such as [37, 19] 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.

4.2. Quantitative Evaluations

We evaluate our approach quantitatively through a number of experiments. Tab. 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. Tab. 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. Tab. 2 compares our monocular reconstruction on in-the-wild images with that of different approaches [28, 32, 27]. We use 130 images from CelebA-HQ [39] as a test set and report the Si-MSE [40] and face identity dissimilarity (Face dis) [17].

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. Input images were selected to cover a rich variety in terms of pose and illumination. 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 [28, 27, 32] in the same manner. Tab. 2 shows that our approach outperforms these methods over 86 images from our test set.

4.3. Ablative Study

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

4.3.1 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. Tab. 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 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 Tab. 1 of the main paper. We report scale invariant MSE (Si-MSE) for renderings with same and different input pose.

	Only mean face (std. dev.)	With all components (std. dev.)
Reconstruction (Si-MSE)	0.011 ($\sigma=0.005$)	0.004 ($\sigma=0.002$)
Reconstruction (Face dis)	0.550 ($\sigma=0.073$)	0.550 ($\sigma=0.080$)

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).

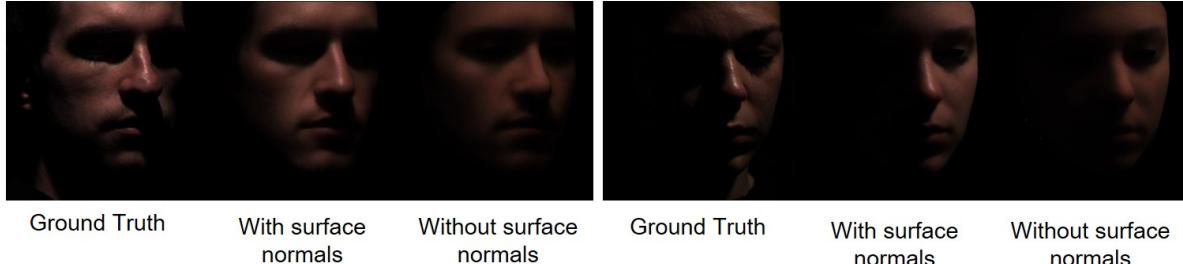


Figure 6. Removing surface normals from our reflectance learning leads to blurry results and weaker capturing of specularities.

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. 6 where removing surface normals leads to blurry results and weak specularities.

4.3.2 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 [39] as a test set and report the Si-MSE [40] and face identity dissimilarity (Face dis) [17]. This is the same test-set used in Tab. 2 (main paper). Tab. 4 reports the Si-MSE and face identity dissimilarity over the test-set. Not learning the face geometry and using a fixed mean mesh instead leads to clear degradation in performance in terms of Si-MSE.

5. Discussion and Limitations

While we show results which allow for estimation of full reflectance fields from monocular images for the first time, our method still has some limitations. As mentioned before, our method 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 reflectance, strong effects such

as specularities and strong shadow boundaries can still be a bit blurred, see Fig. 2. This could again be due to inaccuracies in monocular reconstruction, leading to misalignments between views during training. Nevertheless, we believe that our method takes an important step towards learning and rendering the full reflectance field of a face.

Acknowledgments. 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 Technicolor and 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 *European Conference on Computer Vision*, pages 201–214. Springer, 2012.
- [3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.
- [4] Mario Botsch, Leif Kobbelt, Mark Pauly, Pierre Alliez, and Bruno Lévy. *Polygon mesh processing*. CRC press, 2010.
- [5] 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*, pages 145–156, 2000.
- [6] 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.
- [7] Graham Fyffe. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH’09: Posters*, pages 1–1. 2009.
- [8] 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. *arXiv preprint arXiv:1704.00090*, 2017.
- [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.
- [10] 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.*, 35(6):219:1–219:11, 2016.
- [11] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. page 1, 12 2011.
- [12] Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6927–6935, 2019.
- [13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
- [15] 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.
- [16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [17] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [19] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction “in-the-wild”. *arXiv preprint arXiv:2003.13845*, 2020.
- [20] 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)*, volume 32, pages 1–8, 2013.
- [21] 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. volume 38, July 2019.
- [22] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH*, pages 117–128. ACM, 2001.
- [23] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [24] 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*, pages 234–241. Springer, 2015.
- [25] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).
- [26] Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Real-time avatar animation from a single image. In *FG*, pages 213–220. IEEE, 2011.
- [27] A. Schneider, S. Schnborn, B. Egger, L. Frodeen, and T. Vetter. Efficient global illumination for morphable models. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3885–3893, 2017.
- [28] William A. P. Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [29] 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.*, 38(4), July 2019.
- [30] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgarib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, 2019.
- [31] 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 *The IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

- [32] 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.
- [33] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.
- [34] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. arXiv:1808.09560, August 2018.
- [35] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [36] 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)*, 25(3):1013–1024, 2006.
- [37] 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*, 37(4):1–14, Aug. 2018.
- [38] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.
- [39] 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.
- [40] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single-image portrait relighting. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.