End-to-End 3D Face Reconstruction with Expressions and Specular Albedos from Single In-the-wild Images

Qixin Deng
qdeng4@uh.edu
University of Houston

Binh H. Le
ble@ea.com
SEED Lab, Electronic Arts

Aobo Jin
jinA@uhv.edu
University of Houston-Victoria

Zhigang Deng*
zdeng4@central.uh.edu
University of Houston

ABSTRACT

Recovering 3D face models from in-the-wild face images has numerous potential applications. However, properly modeling complex lighting effects in reality, including specular lighting, shadows, and occlusions, from a single in-the-wild face image is still considered as a widely open research challenge. In this paper, we propose a convolutional neural network based framework to regress the face model from a single image in the wild. The outputted face model includes dense 3D shape, head pose, expression, diffuse albedo, specular albedo, and the corresponding lighting conditions. Our approach uses novel hybrid loss functions to disentangle face shape identities, expressions, poses, albedos, and lighting. Besides a carefully-designed ablation study, we also conduct direct comparison experiments to show that our method can outperform state-of-art methods both quantitatively and qualitatively.

CCS CONCEPTS

• **Computing methodologies → Reconstruction**; **Mesh models**.

KEYWORDS

3D face reconstruction, deep learning algorithms, specular albedo, facial expressions

ACM Reference Format:

Qixin Deng, Binh H. Le, Aobo Jin, and Zhigang Deng*. 2022. End-to-End 3D Face Reconstruction with Expressions and Specular Albedos from Single In-the-wild Images. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547800

1 INTRODUCTION

Faithfully recovering 3D dense models of human faces from in-the-wild images is a long-standing research challenge in computer graphics, computer vision, and multimedia communities, and such a technology has shown its power in many applications, for example, 3D avatar creation [6, 15, 29], face recognition [3, 19, 32, 45], facial video editing [2, 17, 28], and 3D facial animation [8, 9, 26]. Due to the lacking of 3D information in 2D images, inferring the 3D face shape from a single face image in the wild is particularly difficult. The emergence of 3D Morphable Models (3DMMs) [5, 18, 27], which define a space of continuous face deformations and provide low-dimensional parametric representations, makes the face reconstruction problem computationally solvable. Conventional optimization approaches [14, 15, 25, 44] search for the optimal alignment between the projected model and the image through inverse rendering in the space defined by the 3DMM models. However, such a high-dimensional optimization process is often error-prone to local minimums and thus largely depends on initialization. In addition, most of existing 3D face reconstruction methods [7, 17, 38, 41, 43] only model a diffuse albedo and diffuse lighting, and they also put limited effort on the reconstruction of facial expressions. With neither specular albedo nor specular lighting, the accuracy of 3D face reconstruction is limited.

Inspired by the above challenge, in this paper we propose a new deep learning based framework to automatically infer 3D face models with expressions and albedos from an input 2D face image. Specifically, we design novel hybrid loss functions in the deep learning framework to simultaneously infer realistic face shapes, expressions, diffuse and specular albedos, and lighting conditions from a single in-the-wild face image. We also conduct various experiments to evaluate the effectiveness and accuracy of our method and benchmark its performance.

The main contributions of this work include: (i) A CNN-based network that reconstructs the face model from a single in-the-wild image with accurate dense 3D shape, head pose, expression, diffuse albedo, specular albedo, and the corresponding lighting conditions. (ii) The proposed method is the first-of-its-kind framework to automatically entangle diffuse albedos from specular albedos, as well as their corresponding light conditions, from a single face image input. (iii) Novel loss functions are designed to accurately train the deep learning models for 3D face reconstruction tasks, including the photometric loss, inner mouth loss, and the uniform loss.

2 RELATED WORK

Generally, the goal of 3D face reconstruction is to estimate the face identity, expression, and albedo, based on various types of inputs, such as a single image, multi-view images, monocular video, or RGB-D sequences. Existing methods can generally be classified into two categories: 1) optimization based methods, and 2) deep

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547800

learning based methods. Conventional optimization approaches fit a model to given images by optimizing energy functions. In the early years, Romdhani and Vetter [36] proposed a multi-features fitting algorithm by utilizing multiple types of information in the images, such as edges and specular lighting, to reduce local minima. Garrido et al. [15] proposed a multi-layer optimization method to generate detailed 3D face rigs from monocular video. Thies et al. [44] and Ma and Deng [28] presented optimization algorithms and systems to achieve real time facial reenactment from live video. But the limitations of conventional optimization methods restrict their further improvement. It is well known that, the convergence of iterations is quite sensitive to initial conditions. Thus, the above non-linear optimization methods may be less robust to handle a variety of inputs in practice.

Deep learning based regression methods that reconstruct 3D faces from 2D images have produced many promising results. Deep learning based methods can be further classified into two groups: 1) supervised learning, and 2) unsupervised learning. Supervised learning methods [31, 35, 40] generally require ground truth data as the reference. But often it is practically difficult to collect a large amount of dense human face scans due to the expense and the time-consuming labor work. Thus, many researchers resorted to the generation of synthetic data using morphable models[11]. The synthetic data usually cannot cover high diversity such as in-the-wild images; as a result, the performance of such trained models is often limited. To tackle the above limitations, unsupervised and weak-supervised methods have been introduced in recent years. Tewari et al. [43] proposed a parametric model-based decoder, which is fully analytical and differentiable, to enable unsupervised end-to-end training. Genova et al. [17] proposed a model that is trained at the perceptual level. They only impose the similarity of identity features between the input and the learned output, which is extracted from a face recognition network. Gecer et al. [16] presented a GAN model combined with a hybrid content loss function. Sanyal et al. [38] proposed the RingNet architecture, which utilizes a controlled human face collection to enforce the shape to be clustered to the same identity. Feng et al. [12] presented a Detailed Expression Capture and Animation model to robustly produce an UV displacement map from a low-dimensional latent space.

Different from the previous works that directly regress 3DMM coordinates from input, several recent works [10, 13, 22, 46] are model-free and they directly regress 3D face voxels or meshes. Jackson et al. [22] proposed a method to directly regress the volumetric representation of the 3D facial geometry from a single image. Feng et al. [13] designed a 2D representation, named UV position map. Their method first transfers a 3D face model to a 2D UV map space, and then, a CNN is used to repress it from a single 2D image. Wei et al. [46] proposed a graph convolution network (GCN) to directly regress 3D face model coordinates. Although the above model-free methods can capture more details and variations than model-based methods, they often require an explicit 3D supervision.

Faces in in-the-wild (unconstrained) images show a high variability of poses, expressions, and illuminations. Researchers have dealt with unconstrained face images for many years, and have produced some exciting results, such as face recognition [3, 19, 32, 45], face alignment [48, 49, 51], and face segmentation [30, 50]. These methods encode rich information in a low-dimensional space to make face reconstruction from an unconstrained image possible, with the aid of deep learning algorithms.

3 PRELIMINARIES

Morphable 3D Face Models. The Basel face model [18] is one of widely-used 3DMM face models. Specifically, given the shape coefficients $\mathbf{c}_i \in \mathbb{R}^{199}$, the expression coefficients $\mathbf{c}_e \in \mathbb{R}^{100}$, and the albedo coefficients $\mathbf{c}_t \in \mathbb{R}^{199}$ with standard Normal distributions, the shape and texture of a 3D face model are represented as follows:

$$\begin{aligned} \mathbf{S}(\mathbf{c}_i, \mathbf{c}_e) &= \overline{\mathbf{S}} + \mathbf{M}_i\mathbf{c}_i + \mathbf{M}_e\mathbf{c}_e \ , \\ \mathbf{T}(\mathbf{c}_t) &= \overline{\mathbf{T}} + \mathbf{M}_t\mathbf{c}_t \ , \end{aligned} \tag{1}$$

where $\overline{\mathbf{S}} \in \mathbb{R}^{3n}$ and $\overline{\mathbf{T}} \in \mathbb{R}^{3n}$ are the mean face shape and the texture, respectively; n is the number of vertices; $\mathbf{M}_i \in \mathbb{R}^{3n\times199}$, $\mathbf{M}_t \in \mathbb{R}^{3n\times199}$ and $\mathbf{M}_e \in \mathbb{R}^{3n\times100}$ are the matrices calculated by multiplying linear PCA bases and the diagonal matrices containing the square roots of the corresponding PCA eigen-values.

However, the texture model used by current 3DMMs is tangled with illumination, which introduces many external interference factors, such as baking in shading, shadowing, specularities, and light source colors. Thus, current model-based methods consider neither specular albedo nor lighting during the face reconstruction process. Hence, applications of these methods are limited because of incomplete texture. Smith et al. [42] proposed a pipeline to acquire truly intrinsic diffuse and specular albedo, which fully factors out the effects of camera, illumination, and other interference factors. Substituting this new albedo model to the original Basel albedo model, new face albedo colors are then represented as:

$$\mathbf{T}(\mathbf{c}_t) = [(\overline{\mathbf{T}}_d + \mathbf{M}_d\mathbf{c}_t) + (\overline{\mathbf{T}}_s + \mathbf{M}_s\mathbf{c}_t)]^{\frac{1}{2.2}} \ , \tag{2}$$

where $\overline{\mathbf{T}}_d \in \mathbb{R}^{3n}$ and $\overline{\mathbf{T}}_s \in \mathbb{R}^{3n}$ are the average diffuse albedo and the average specular albedo, respectively; $\mathbf{M}_d \in \mathbb{R}^{3n\times145}$ and $\mathbf{M}_s \in \mathbb{R}^{3n\times145}$ are basis matrices for the diffuse and the specular albedo, respectively. A non-linear gamma transformation is used to fit the camera's colour space for the camera model that does not work in sRGB. In Figure 1, we show the Face Model that is used in this work.

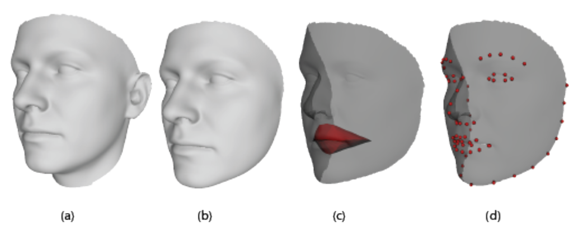

Figure 1: (a) shows the original face model with the ears and the neck. We crop the frontal face area (without the ears and the neck) and the resulting face model (b) used in our work. The red area in (c) is a pre-defined mesh which wraps around the inner mouth area. (d) shows 68 pre-defined facial landmarks (red dots) on the cropped face model.

Camera Model. We employ a weak-perspective camera model for the projection of 3D faces to 2D images. The position and direction of the camera are fixed, and its field of view (FOV) is empirically selected. The projection of each vertex $\mathbf{v}$ is represented by the following perspective transformation:

$$\text{Projection}(\mathbf{v}) = f \times \mathbf{R} \times \mathbf{T} \times \mathbf{v} \,, \tag{3}$$

where f is the scaling factor, $\mathbf{R}$ and $\mathbf{T}(t_x, t_y, t_z)$ are the 3D face rotation matrix and translation matrix, respectively. Because the rotation matrix $\mathbf{R}$ can be simply parameterized as three Euler Angles (α, β, γ), the pose of the 3D face can be regressed as 7 parameters: α, β, γ, t_x, t_y, t_z, and f.

Illumination Model. We use the Blinn-Phong reflection model for shading. To restore realistic face albedos and lighting, our illumination model includes two parts to render face albedo colors. We assume the diffuse albedo is Lambertian and use Spherical Harmonics (SH) [33, 34] to approximate the scene of diffuse illumination. The per-vertex shaded diffuse albedo color c_d is then computed as:

$$c_d = t_d \cdot \sum_{b=1}^{B^2} \lambda_b \cdot \Pi_b(\vec{\mathbf{n}}) \,, \tag{4}$$

where t_d is the vertex diffuse albedo, $\vec{\mathbf{n}}$ is the normalized vertex normal, $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}$ are SH basis functions, and λ denotes the corresponding SH coefficients.

To compute the specular albedo colors, a differentiable Blinn-Phong bidirectional reflectance distribution function (BRDF) is employed. To simplify the process, we only model a single light source and compute the specular albedo color c_s as follows:

$$c_s = t_s \cdot [\text{Nor}(\vec{\mathbf{v}}_{out} - \vec{\mathbf{v}}_{in}) \cdot \vec{\mathbf{n}}]^{\varphi} \,, \tag{5}$$

where $\text{Nor}(\cdot)$ represents the normalization process, t_s is the vertex specular albedo, $\vec{\mathbf{v}}_{in}$ is the normalized incoming light vector calculated by the predicted light source position $\mathbf{p}_l$, $\vec{\mathbf{v}}_{out}$ is the normalized outgoing light vector calculated by the pre-defined camera position, $\vec{\mathbf{n}}$ is the normalized vertex normal, and φ is the shininess coefficient. For each vertex, the final rendered color c is then defined as: $c = c_d + c_s$.

4 OUR METHOD

In this section we describe our deep learning based model with hybrid loss functions to regress the face 3DMM parameters and simulate realistic lighting conditions from an in-the-wild image. The schematic view of our deep network is illustrated in Figure 2. As shown in this figure, we choose the ResNet-50 (DCNNs) [21] as the CNN backbone to regress the shape and albedo coordinates of the face morphable model, and infer the face pose and lighting condition. The face morphable model represents a wide range of faces in a low-dimensional space, and it is particularly suitable for CNN-based methods. To form a self-surpervised training process, a differentiable renderer is introduced into our model to measure the discrepancy between the input and the learned result. Meanwhile, a face recognition network [39] is used in our model to generate face identity features during the training. For each input image, we pre-label its face segmentations, skin weights, and 2D landmarks. In particular, in this work we propose hybrid loss functions (described below) to train our model to reconstruct a realistic face from a single in-the-wild image.

4.1 Hybrid Loss Functions

The hybrid training loss functions in our method are defined as follows:

$$\begin{aligned} L = {} & \omega_{id} L_{id} + \omega_{photo} L_{photo} + \omega_{lmk} L_{lmk} + \omega_{mouth} L_{mouth} \\ & + \omega_{reg} L_{reg} + \omega_{uniform} L_{uniform} \,, \end{aligned} \tag{6}$$

where L_{id} denotes the identity loss, L_{photo} denotes the photometric loss, L_{lmk} denotes the landmark re-projection loss, L_{mouth} denotes the inner mouth loss, L_{reg} denotes the regularization loss, and L_{smooth} denotes the uniform loss. The details of these losses are described below.

In our experiments, the weights for the above losses are empirically determined as follows: $\omega_{id} = 0.2$, $\omega_{photo} = 1.92$, $\omega_{lmk} = 1.8e^{-3}$, $\omega_{mouth} = 1.3e^{-3}$, $\omega_{reg} = 5.0e^{-4}$, and $\omega_{uniform} = 5.0$.

4.1.1 Identity Loss. Several recent 3D face reconstruction methods [16, 17] utilized features extracted from face recognition networks to formulate a loss and effectively generate realistic faces. Our method takes advantage of the state-of-art face recognition network [39], $\mathcal{F}$, to extract the features of both input images and the reconstructed images. Then, we compute the cosine similarity between the two paired features as the identity loss, L_{id}, as follows:

$$L_{id} = 1 - \frac{\mathcal{F}(\mathbf{I}_{in})\mathcal{F}(\mathbf{I}_r)}{\|\mathcal{F}(\mathbf{I}_{in})\|_2 \cdot \|\mathcal{F}(\mathbf{I}_r)\|_2} \,, \tag{7}$$

where $\|\cdot\|_2$ denotes the l_2-norm, and $\mathbf{I}_{in}$ and $\mathbf{I}_r$ denote the input images and the reconstructed (rendered) images, respectively. This loss encourages the reconstruction image to be close to the input image in the low-dimensional embedding space, so that the reconstructed face can capture more fundamental and detailed identity information of the input face.

4.1.2 Photometric Loss. The above identity loss only works at the perceptual level and the used face recognition network [39] is trained to be robust with albedo colors and illuminations. Therefore, the identity loss ignores pixel level details. To enable our model to recover faithful face albedo colors, we need to introduce additional information into our networks. The challenge is that we cannot simply subtract the rendered image from the input image due to two main reasons: i) Faces in in-the-wild images often have occlusions. A face can be occluded by hair, beard, or other objects such as a hat or a pair of glasses. These occlusions could lead to errors in local albedo. ii) Lighting conditions can also have significant influence on face albedo colors. Our model aims to recover diffuse and specular albedos, and corresponding lighting conditions. Because in the rendering pipeline, the diffuse and specular albedo colors are calculated in a totally different way, direct subtraction cannot disengage them in our model.

To reduce the interference caused by occlusions, we apply a face segmentation network to put the focus of our model on face skin regions, $\mathbf{I}_{face}$, obtained through a skin detector (described below). In these skin regions, we need to further separate the diffuse albedo and the specular albedo. Inspired by the work of [7] and based on a key observation that the diffuse albedo is often smoother, more uniform, and contains much more colour information than the

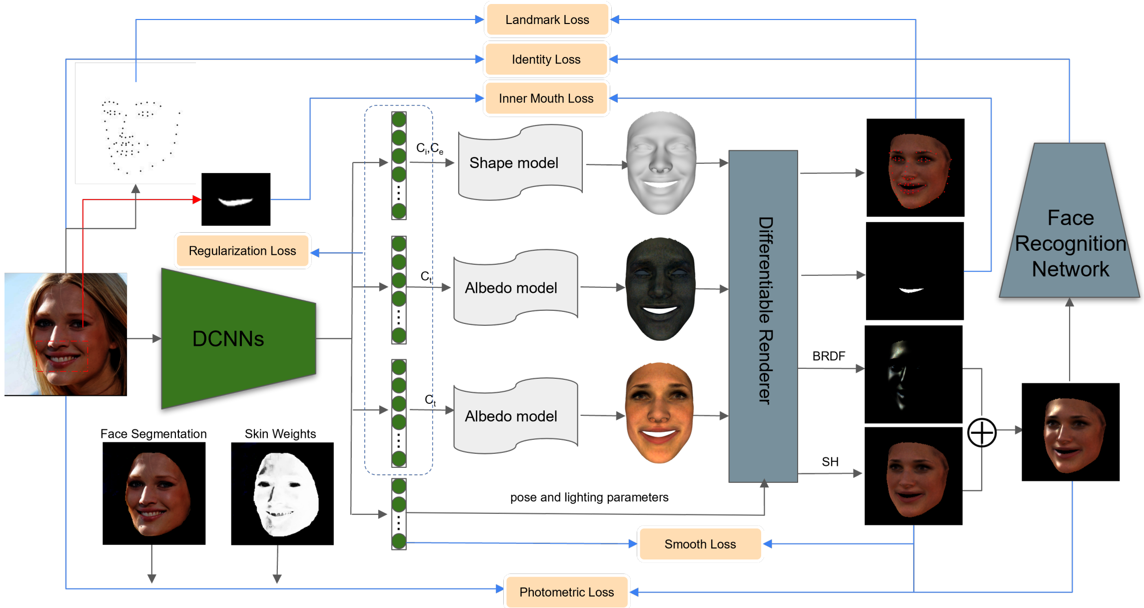

Figure 2: Schematic view of our method. We use the ResNet-50 (DCNNs) [21] as the backbone of our network.

specular albedo, we construct a weighted mask (refer to Eq. 9) based on $\mathbf{I}_{face}$, which is able to drive out the specular albedo interference and maximally capture the original diffuse albedo colors.

We let the diffuse and the specular albedos go through independent rendering processes to generate two face images $\mathbf{I}_{r\_diffuse}$ and $\mathbf{I}_{r\_specular}$. Meanwhile, a projected face region $\mathbf{I}_{project}$ is obtained. Our photometric loss L_{photo} only focuses on the intersection of $\mathbf{I}_{face}$ and $\mathbf{I}_{project}$, denoted as $\mathcal{M}$. Formally, L_{photo} is defined as follows:

$$L_{photo} = \frac{\sum_{i \in \mathcal{M}} C_{skin,i} \cdot \|\mathbf{I}_{in,i} - \mathbf{I}_{r\_diffuse,i}\|_2}{\sum_{i \in \mathcal{M}} C_{skin,i}} + \eta \frac{\sum_{i \in \mathcal{M}} \|\mathbf{I}_{in,i} - (\mathbf{I}_{r\_diffuse,i} + \mathbf{I}_{r\_specular,i})\|_2}{|\mathcal{M}|}, \quad (8)$$

$$\text{where: } C_{skin,i} = \begin{cases} 1, & \text{if} \quad p_i > 0.5, \\ p_i, & \text{otherwise,} \end{cases} \quad (9)$$

p_i is the probability of the pixel i is skin,

$|\mathcal{M}|$ denotes the size of $\mathcal{M}$.

η is a user-specified weight.

We adopt a naive Bayes classifier with Gaussian Mixture Models (GMMs) to compute p_i for each pixel i in $\mathbf{I}_{face}$. To this end, a skin pixel has a higher weight than beard, hair, and specular highlight pixels when we calculate the first term in Equation (8).

4.1.3 Landmark Re-projection Loss. Both the above identity loss and the photometric loss are sensitive to the misalignment between input and rendered images. Thus, we introduce a landmark loss in our network to improve the convergence and the effectiveness of other losses. Basically, the landmark loss L_{lmk} measures the Euclidean distance between the labeled 2D landmarks $\mathbf{k}_i$ on input images and the projections of corresponding landmarks on 3D morphable face model $\mathbf{k}_i^p$ as follows:

$$L_{lmk} = \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} \mu_i \|\mathbf{k}_i - \mathbf{k}_i^p\|_2, \quad (10)$$

where $\mathcal{K}$ is the set of landmarks, and μ_i is the importance weight of the i-th landmark. We assign large weights (e.g., 20) to the landmarks on both the outer mouth and the eyebrows, while assign small weights (e.g., 1) to the remaining ones.

4.1.4 Inner Mouth Loss. The shape of the mouth generally plays an important role for facial expressions. The accurate estimation of the inner mouth landmarks is challenging, especially for closed or slightly open mouths, because of the overlapping of the upper and the lower lip contours. In addition, the landmark loss only has limited influence on back-propagation and parameters update. Instead of only using sparse landmarks, we utilize the dense inner mouth area to formulate a loss to constrain the mouth expression. We utilize several geometry faces to enclosure the inner mouth area of the 3D morphable face model (refer to Figure 1(c)). The dense inner mouth loss L_{mouth} is computed as the difference between the projected inner mouth area A^p and the labeled inner mouth area A as follows:

$$L_{mouth} = \|A - A^p\|, \quad (11)$$

where $\| \cdot \|$ denotes the absolute value of l_1-norm. To ensure a stable training process, this loss will only be included after 50k iterations.

4.1.5 Regularization Loss. Biases could be introduced if only the above losses are used to train our model. Such biases are often considered as the domain gap between real face images and the 3D morphable face model. Hence, a regularization loss is necessary to enforce the distribution of the reconstructed faces to be close to the zero-mean standard normal distribution assumption and to prevent model degeneration. The regularization loss L_{reg} is formulated as follows:

$$L_{reg} = \sum \|\mathbf{c}_i\|^2 + \|\mathbf{c}_t\|^2 + \|\mathbf{c}_e\|^2 \tag{12}$$

where $\mathbf{c}_i$ represents shape coefficients, $\mathbf{c}_t$ represents albedo coefficients, and $\mathbf{c}_e$ represents expression coefficients in the 3D morphable face model.

4.1.6 Uniform Loss. To reward the amount of information picked up by the specular albedo, we apply a uniform loss on the 3D face diffuse albedo and Spherical Harmonics (SH) lighting parameters. The loss is designed to penalize diffuse albedo color variance, which makes the diffuse albedo colors to be smooth and uniform. the uniform loss $L_{uniform}$ is defined as follows:

$$L_{uniform} = \frac{1}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}(t_i - t_{mean}) + \sum_{b=1}^{B^2}\sum_{i=1}^{3}\lambda_{b,i} - \overline{\lambda_b}; \tag{13}$$

where $\mathcal{T}$ denotes the lit diffuse albedo colors, λ denotes SH coefficients, and B denotes the number of SH bands.

4.2 Implementation Details

Here we describe the details of setup and training of our model. We used the face recognition network (FaceNet) proposed by Schroff et al. [39] and trained it with the triplet loss. We employed a weighted U-Net [50] combined with a pixel-level cross entropy loss to train our face segmentation network. To detect and align all face images, we adopted the method proposed by Zhang et al. [49] with the image size of 224×224. To detect 68 IBUG facial landmarks [37], we employed the method in [4]. ResNet [21] is a powerful network structure to do image recognition related tasks, thus we chose it to form the main body of our network. We trained our face reconstruction network on the dataset we collected, which contains about 250 thousand images of 11 thousand identities. We randomly picked 2000 images from the dataset we collected as a test set (called the test set in this paper). Our model training used the Adam optimizer with batch size of 32, learning rate of $5e^{-5}$, and 200 thousand iterations.

5 RESULTS AND EVALUATIONS

5.1 Ablation Study

We conducted multiple ablation experiments on the test set to validate the effectiveness of each loss function (except L_{lmk}) in our model (refer to Section 4.1). Note that we did not conduct an ablation experiment to take out the landmark re-projection loss L_{lmk}, since our model cannot converge properly without L_{lmk}.

The Uniform Loss. In Figure 3, we show several comparison results between our full model and our model minus the uniform loss function $L_{uniform}$. As shown in this figure, our full model (with the introduced $L_{uniform}$) can better recover the diffuse/specular albedo, compared to the one without the uniform loss function. Specifically, comparing (b) and (e), we can see that $L_{uniform}$ helps our model to restore a uniform lit diffuse albedo. Without $L_{uniform}$, the specular parts in a face mainly blend into the diffuse albedo colours (red dotted box in Figure 3). As such, the remaining information

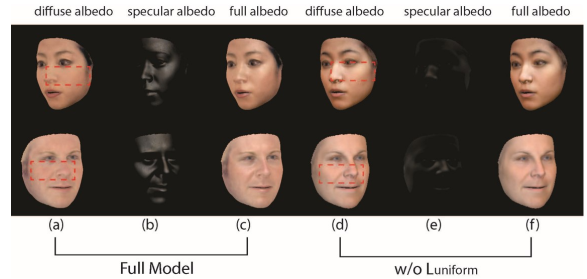

Figure 3: Example comparisons between our full model (a-c) and the model without $L_{uniform}$ (d-f). (a): the lit diffuse albedo colors produced by our full model. (b): the lit specular albedo colors produced by our full model. (c): the reconstructed faces produced by our full model. (d): the lit diffuse albedo colors produced by our model without $L_{uniform}$. (e): the lit specular albedo colors produced by our model without $L_{uniform}$. (f): the reconstructed faces produced by our model without $L_{uniform}$.

is insufficient for our model to correctly learn the specular albedo and its corresponding lighting. A more accurate specular albedo can be learned with $L_{uniform}$ (c) than the case without $L_{uniform}$ (f). Finally, the reconstructed faces by our full model with $L_{uniform}$ (d) have more realistic albedos than the case without $L_{uniform}$ (g), using the input ground-truth images (a) as the reference.

Identity/Photometric/Regularization loss. In Figure 4, we show three examples to validate the effectiveness of the identity loss L_{id}, the photometric loss L_{photo}, and the regularization loss L_{reg}. As shown in this figure, the results by our full model (b) have sharper texture and more realistic appearance than those by our model without L_{id} (c). This can be obviously observed around the eyes (within red dotted boxes). Comparing (b) and (d), we can see that our model without L_{photo} cannot correctly produce the correct face albedos and corresponding lighting conditions. In addition, comparing (b) and (e), we can see that our model without L_{reg} cannot produce accurate face shapes, since the regularization loss L_{reg} can help to prevent the degeneration of the reconstructed faces and make the reconstructed faces fall into the distribution of the trained morphable face model.

Inner Mouth Loss. In Figure 5 we show two examples to validate the effectiveness of the inner mouth loss L_{mouth}. With the introduction of L_{mouth} that can help to preserve the inner mouth shape in the reconstructed faces, especially for closed or slightly open mouth cases, our full model can reconstruct more accurate mouth shapes and thus more accurate facial expressions than our model without L_{mouth}.

Table 1 shows the albedo color accuracies on different versions of our model. We calculated the mean errors and standard deviations between the input face images and the images rendered from the corresponding reconstructed faces by different versions of our model. From this table, we can see that our full model achieves the smallest error, compared to other versions with one loss function removed. Specifically, the photometric loss L_{photo} makes the largest contribution to the reconstruction accuracy of albedo colors,

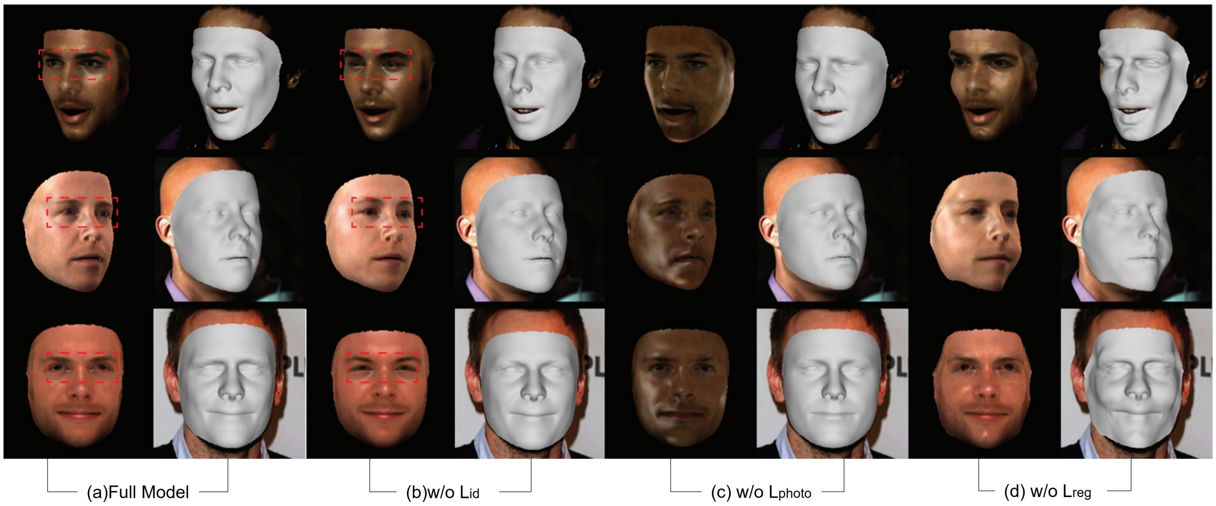

Figure 4: Example comparisons to show the effectiveness of the identity loss L_{id}, the photometric loss L_{photo}, and the regularization loss L_{reg}.

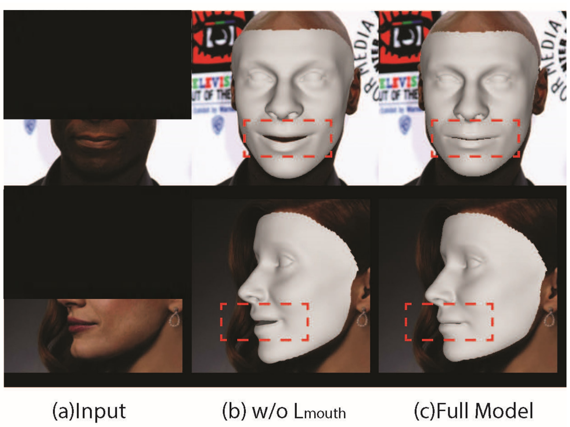

Figure 5: Example comparisons to show the effectiveness of the inner mouth loss L_{mouth}. (a): input mouth shape. (b): the reconstructed face shapes by our model without L_{mouth}. (c): the reconstructed face shapes by our full model.

Table 1: The mean errors and standard deviations of pixel colors on our testing dataset.

L_{id}	L_{photo}	L_{lmk}	L_{mouth}	L_{reg}	L_{uni}	*Mean* ± *Std*
√	√	√	√	√	√	**0.115 ± 0.169**
×	√	√	√	√	√	0.118 ± 0.172
√	×	√	√	√	√	0.228 ± 0.247
√	√	×	√	√	√	*NaN*
√	√	√	×	√	√	0.117 ± 0.170
√	√	√	√	×	√	0.148 ± 0.192
√	√	√	√	√	×	0.123 ± 0.179

followed by the regularization loss L_{reg}. Other loss terms also help to improve the accuracy.

We further evaluated the reconstructed shape accuracy of different versions of our model on the FaceScape Dataset [47]. In our experiments, we randomly chose 10 males and 10 females from the multi-view data, and calculated the point-to-plane distance using the Iterative Closest Point (ICP) algorithm with an isotropic scale to find the best alignment. The errors are presented in Table 2. From this table, we can see that the regularization loss L_{reg} makes the largest contribution to the shape accuracy, followed by the identity loss L_{id}. Other loss terms also make contributions to the accuracy of the reconstructed face shapes.

Table 2: The means and standard deviations of the reconstruction errors (in mm) on 400 face meshes in the FaceScape dataset [47]

L_{id}	L_{photo}	L_{lmk}	L_{mouth}	L_{reg}	L_{uni}	*Mean* ± *Std*
√	√	√	√	√	√	**2.05 ± 0.79**
×	√	√	√	√	√	2.39 ± 1.13
√	×	√	√	√	√	2.27 ± 1.09
√	√	√	×	√	√	2.15 ± 0.87
√	√	√	√	×	√	9.89 ± 5.81
√	√	√	√	√	×	2.16 ± 0.77

5.2 Quantitative Evaluations

We also performed quantitative evaluations by comparing our method with some state of the art face reconstruction methods, including 3DDFA_V2 [20], MGCNet [41], the weakly-supervised learning (WSL) based method [7], PRNet [13], and RingNet [38].

Quantitative shape accuracy on the MICC dataset. We first quantitatively compared the reconstructed shape accuracy among our method the above state of the art methods on the MICC Florence 3D Faces Dataset [1]. The MICC dataset contains 53 identities, where the data of each identity contain a high quality neutral face scan and three video clips in different environments (cooperative, indoor, and outdoor). The dataset only provides a single neutral face scan for each identity; however, the PRNet method [13] is based on position maps and it cannot remove expressions from input images. To make a fair comparison, we manually selected 10 to 20 frames from each video clip so that the neutral expression and various face poses can be covered. In the RingNet method [38], we defined a front face mesh covering similar area with the other methods. In this comparison, we average the resulting meshes from all the frames in each video clip and then compare the averaged mesh with the ground truth mesh/scan. Specifically, we run the ICP algorithm with an isotropic scale to find the optimal alignment and then compute the point-to-plane distances between the two meshes as the shape accuracy errors. The obtained statistical results in this quantitative comparison are presented in Table 3. As shown in this table, our method achieves the smallest average errors among all the methods for all the three video clips (corresponding to three different environments). Example comparison results are also shown in Figure 6. From this figure, we can see that the face shapes produced by our method are clearly closer to the ground-truth face than other methods.

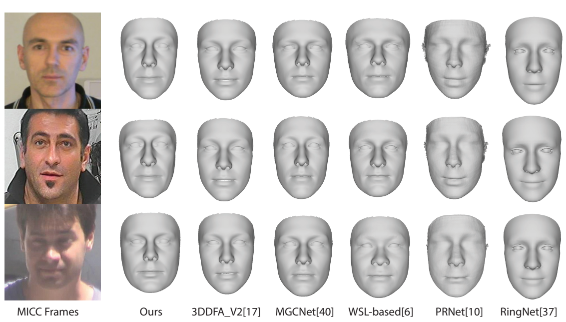

Figure 6: Example comparisons among our method and the selected state of the art methods on the MICC. The example frames (from top to bottom) are taken from cooperative, indoor, and outdoor video clips, respectively. Each face shape is the averaged face mesh for the corresponding video clip.

Table 3: The average values and standard deviations of the point-to-plane distances (in mm) between the results and the ground truths on the MICC.

Methods	Cooperative	Indoor	Outdoor
Ours	**1.60 ± 0.55**	**1.63 ± 0.52**	**1.68 ± 0.65**
3DDFA_V2 [20]	1.66 ± 0.56	1.67 ± 0.52	1.71 ± 0.66
MGCNet [41]	1.78 ± 0.55	1.78 ± 0.54	1.81 ± 0.59
WSL-based [7]	1.68 ± 0.52	1.67 ± 0.54	1.73 ± 0.62
PRNet [13]	2.01 ± 0.65	2.00 ± 0.58	2.13 ± 0.66
RingNet [38]	1.72 ± 0.58	1.73 ± 0.60	1.75 ± 0.59

Quantitative shape accuracy on the FaceScape dataset. We also quantitatively compared the shape accuracy among our method and the above state of the art methods on the FaceScape Dataset [47]. The FaceScape Dataset provides high quality expression face scans with corresponding multi-view images in an controlled environment for each identity. We randomly selected 50 males and 50 females from the multi view data with three expressions: neutral, mouth stretch, and grin. For each expression, we use the first 40 views to produce 40 meshes, and average these meshes to generate the final result for evaluation. We also run the ICP with an isotropic scale to find the optimal alignment and further compute the point-to-plane distances. We report the obtained statistical results in Table 4. As shown in this table, our method achieves the smallest average distances for all the three expression cases among all the methods in this comparison.

Table 4: The average values and standard deviations of the point-to-plane distances (in mm) between the results and the ground truths on the FaceScape Dataset

Methods	Neutral	Mouth Stretch	Grin
Ours	**2.03 ± 0.66**	**2.14 ± 0.77**	**2.16 ± 0.75**
3DDFA_V2 [20]	2.28 ± 0.70	2.25 ± 0.76	2.25 ± 0.73
MGCNet [41]	2.53 ± 0.75	2.56 ± 0.64	2.56 ± 0.69
WSL-based [7]	2.26 ± 0.66	2.26 ± 0.71	2.29 ± 0.70
PRNet [13]	2.81 ± 0.85	2.88 ± 0.89	2.87 ± 0.86
RingNet [38]	2.42 ± 0.67	2.38 ± 0.74	2.40 ± 0.75

The accuracy of mouth expression landmarks. Finally, We evaluated and compared the accuracy of mouth expression landmarks among our method and the chosen state of the art methods on the AFLW-2000 Dataset [51]. In this dataset, there are 2000 images covered by fitted 3D faces with labeled landmarks. We calculated the average errors of the inner mouth landmarks, as shown in Table 5. From this table we can see that our method also outperforms all other methods in terms of the average errors of the inner mouth landmarks.

Table 5: The average values and standard deviations of the inner mouth landmark errors (in mm) by our method and selected state of the art methods

Ours	3DDFA_V2	MGCNet	WSL-based	PRNet
3.43 ± 0.36	3.55 ± 0.32	3.64 ± 0.38	3.58 ± 0.36	3.87 ± 0.45

5.3 Qualitative Evaluation

We also conducted qualitative evaluations to visually compare the face reconstruction results between our method and the above state of the art methods (i.e., 3DDFA_V2 [20], MGCNet[41], WSL-based [7], PRNet [13], and RingNet[38]). There are several related work [16, 23, 24] we cannot do comparison because those works are claimed being commercialized and we cannot obtain enough related results. Please check our demo video for more examples.

Figure 7 shows some comparison results among our method, the MoFA method [45], and the method in [17]. The input images and the results of both the MoFA method [45] and Genova et al. [17] are directly obtained from [17]. From this figure, we can clearly observe that our method can reconstruct more accurate face shapes, expressions, and albedo colors than both the MoFA method [45] and the method in [17].

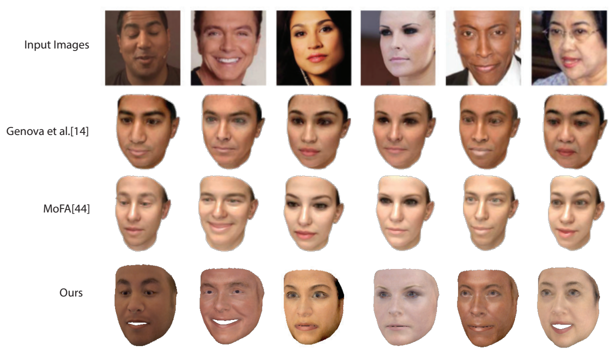

Figure 7: Comparison examples among our method, MoFA [45], and Genova et al. [17].

6 DISCUSSION AND CONCLUSION

In this paper we propose an end-to-end deep learning based method to reconstruct 3D face models from in-the-wild images. In particular, our model successfully jointly recover both diffuse and specular albedos from a single input face image. Our experiment results show that our method can outperform the state-of-the-art face reconstruction methods both quantitatively and qualitatively.

Our current method has several limitations. First, it cannot model multiple light sources in an input image. Therefore, certain specular highlights may be missed from the reconstruction if there are more than one light sources. Second, the limited spatial resolution of the 3DMM can lead to loss of details in the reconstruction. Because the 3DMM captures neither normal maps nor displacement maps of the skin, accurate skin reflectances cannot be rendered with our model. Third, our current method only uses a single global shininess parameter. Therefore, it cannot adapt to local properties on different areas. One potential solution to this problem is to learn per-vertex attributes with Graph Neutral Networks.

ACKNOWLEDGMENTS

This work was in part supported by NSF IIS-2005430, and a research grant from the SEED lab of Electronic Arts.

REFERENCES

[1] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. 2011. The Florence 2D/3D Hybrid Face Dataset. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding* (Scottsdale, Arizona, USA) *(J-HGBU '11)*. ACM, New York, NY, USA, 79–80. https://doi.org/10.1145/2072572.2072597

[2] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating faces in images and video. *Computer graphics forum* 22, 3 (2003), 641–650.

[3] V. Blanz and T. Vetter. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 9 (2003), 1063–1074. https://doi.org/10.1109/TPAMI.2003.1227983

[4] Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

[5] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425. https://doi.org/10.1109/TVCG.2013.249

[6] Qixin Deng, Luming Ma, Aobo Jin, Huikun Bi, Binh Huy Le, and Zhigang Deng. 2021. Plausible 3D face wrinkle generation using variational autoencoders. *IEEE Transactions on Visualization and Computer Graphics* (2021).

[7] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[8] Zhigang Deng and Ulrich Neumann. 2008. Data-driven 3D facial animation. Springer.

[9] Zhigang Deng and Ulrich Neumann. 2008. Expressive Speech Animation Synthesis with Phoneme-Level Controls. *Computer Graphics Forum* 27, 8 (2008), 2096–2113.

[10] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. 2017. End-to-end 3D face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*. 5908–5917.

[11] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2019. 3D Morphable Face Models - Past, Present and Future. *CoRR* abs/1909.01815 (2019). arXiv:1909.01815 http://arxiv.org/abs/1909.01815

[12] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-the-Wild Images. *ACM Trans. Graph.* 40, 4, Article 88 (July 2021), 13 pages.

[13] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 534–551.

[14] Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. 2016. Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–10.

[15] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 1–15.

[16] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. 2019. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1155–1164.

[17] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. 2018. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8377–8386.

[18] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schoenborn, and Thomas Vetter. 2018. Morphable Face Models - An Open Framework. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 75–82. https://doi.org/10.1109/FG.2018.00021

[19] Syed Zulqarnain Gilani and Ajmal Mian. 2018. Learning from millions of 3D scans for large-scale 3D face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1896–1905.

[20] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 152–168.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[22] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 1031–1039.

[23] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction "In-the-Wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. 2021. AvatarMe++: Facial Shape and BRDF Inference with Photorealistic Rendering-Aware GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[25] Martin D Levine and Yingfeng Chris Yu. 2009. State-of-the-art of 3D facial reconstruction methods for face recognition based on a single 2D training image per person. *Pattern Recognition Letters* 30, 10 (2009), 908–913.

[26] John P Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Frederic H Pighin, and Zhigang Deng. 2014. Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports)* 1, 8 (2014), 2.

[27] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17.

[28] Luming Ma and Zhigang Deng. 2019. Real-Time Facial Expression Transformation for Monocular RGB Video. *Computer Graphics Forum* 38, 1 (2019), 470–481.

[29] Luming Ma and Zhigang Deng. 2019. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. 1–10.

[30] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. 2018. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 98–105.

[31] Kyle Olszewski, Joseph J Lim, Shunsuke Saito, and Hao Li. 2016. High-fidelity facial and speech animation for VR HMDs. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–14.

[32] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.

[33] Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 497–500.

[34] Ravi Ramamoorthi and Pat Hanrahan. 2001. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 117–128.

[35] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. 2017. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1259–1268.

[36] Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 986–993.

[37] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 397–403.

[38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7763–7772.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[40] Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1576–1585.

[41] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. 2020. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 53–70.

[42] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. 2020. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5011–5020.

[43] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1274–1283.

[44] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[45] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5163–5172.

[46] Huawei Wei, Shuang Liang, and Yichen Wei. 2019. 3d dense face alignment via graph convolution networks. *arXiv preprint arXiv:1904.05562* (2019).

[47] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[48] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. 2014. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*. Springer, 1–16.

[49] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[50] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 3–11.

[51] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 146–155.