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

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

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:

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, 32], face recognition [3, 19, 35, 48], facial video editing [2, 17, 31], and 3D facial animation [8, 9, 28]. 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, 29], 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, 27, 47] 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, 41, 44, 46] 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 learning based methods. A wide variety of optimization approaches have been developed, including iterative optimization techniques, conjugate gradient methods, and those based on gradient descent. The classic work of Zollner et al. [47] forms the starting point for many others, as a widely used optimization technique for 3DMM estimation. However, these methods often suffer from initialization and ill-conditioned matrices. This work and others have proposed optimization-based methods that use 3DMMs to reconstruct faces. However, these methods are limited by the fact that 3DMMs are a limited space of continuous deformations and they are sometimes not able to capture the real-world face shape and expression variations (e.g., wrinkles, healthy skin, furrows, etc.)

Deep learning-based methods are an alternative for 3D reconstruction. These methods are generally trained in a supervised manner to learn the mapping from 2D to 3D. These methods are usually trained on widely known databases such as FFHQ [3], AFLW2000 [33], or FFHQ [3]. These methods can be categorized into optimization-based methods and learning-based methods. In general, optimization-based methods are limited in terms of the number of training datasets, while deep learning-based methods are not limited by this. Furthermore, optimization-based methods are generally not able to perform well on unseen images.

To address these limitations, we propose a novel 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 learning based methods. A wide variety of optimization approaches have been developed, including iterative optimization techniques, conjugate gradient methods, and those based on gradient descent. The classic work of Zollner et al. [47] forms the starting point for many others, as a widely used optimization technique for 3DMM estimation. However, these methods often suffer from initialization and ill-conditioned matrices. This work and others have proposed optimization-based methods that use 3DMMs to reconstruct faces. However, these methods are limited by the fact that 3DMMs are a limited space of continuous deformations and they are sometimes not able to capture the real-world face shape and expression variations (e.g., wrinkles, healthy skin, furrows, etc.)

Deep learning-based methods are an alternative for 3D reconstruction. These methods are generally trained in a supervised manner to learn the mapping from 2D to 3D. These methods are usually trained on widely known databases such as FFHQ [3], AFLW2000 [33], or FFHQ [3]. These methods can be categorized into optimization-based methods and learning-based methods. In general, optimization-based methods are limited in terms of the number of training datasets, while deep learning-based methods are not limited by this. Furthermore, optimization-based methods are generally not able to perform well on unseen images.

To address these limitations, we propose a novel 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 learning based methods. A wide variety of optimization approaches have been developed, including iterative optimization techniques, conjugate gradient methods, and those based on gradient descent. The classic work of Zollner et al. [47] forms the starting point for many others, as a widely used optimization technique for 3DMM estimation. However, these methods often suffer from initialization and ill-conditioned matrices. This work and others have proposed optimization-based methods that use 3DMMs to reconstruct faces. However, these methods are limited by the fact that 3DMMs are a limited space of continuous deformations and they are sometimes not able to capture the real-world face shape and expression variations (e.g., wrinkles, healthy skin, furrows, etc.)

Deep learning-based methods are an alternative for 3D reconstruction. These methods are generally trained in a supervised manner to learn the mapping from 2D to 3D. These methods are usually trained on widely known databases such as FFHQ [3], AFLW2000 [33], or FFHQ [3]. These methods can be categorized into optimization-based methods and learning-based methods. In general, optimization-based methods are limited in terms of the number of training datasets, while deep learning-based methods are not limited by this. Furthermore, optimization-based methods are generally not able to perform well on unseen images.

To address these limitations, we propose a novel 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.
learning based methods. Conventional optimization approaches fit a model to given images by optimizing energy functions. In the early years, Romdhani and Vetter [39] 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. [47] and Ma and Deng [31] 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 [34, 38, 43] 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. [46] 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. Sanjal et al. [41] 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, 23, 49] are model-free and they directly regress 3D face voxels or meshes. Jackson et al. [23] proposed a method to directly regress the 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. [49] 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, 35, 48], face alignment [51, 52, 54], and face segmentation [33, 53]. 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 $c_s \in \mathbb{R}^{199}$, the expression coefficients $c_e \in \mathbb{R}^{100}$, and the albedo coefficients $c_f \in \mathbb{R}^{199}$ with standard Normal distributions, the shape and texture of a 3D face model are represented as follows:

$$S(c_s, c_e) = \bar{S} + M_s c_s + M_e c_e,$$

$$T(c_e) = \bar{T} + M_e c_e ,$$

where $\bar{S} \in \mathbb{R}^{3n}$ and $\bar{T} \in \mathbb{R}^{3n}$ are the mean face shape and the texture, respectively; $n$ is the number of vertices; $M_s \in \mathbb{R}^{3n \times 199}$, $M_e \in \mathbb{R}^{3n \times 199}$ and $M_e \in \mathbb{R}^{3n \times 100}$ 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. [45] 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:

$$T(c_e) = (\bar{T}_d + M_d c_e) + (\bar{T}_s + M_s c_e)$$

where $\bar{T}_d \in \mathbb{R}^{3n}$ and $\bar{T}_s \in \mathbb{R}^{3n}$ are the average diffuse albedo and the average specular albedo, respectively; $M_d \in \mathbb{R}^{3n \times 145}$ and $M_s \in \mathbb{R}^{3n \times 145}$ 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 Basel Face Model that is used in this work.

Figure 1: (a) shows the original Basel 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 Basel 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 $v$ is represented by the following perspective transformation:

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

where $f$ is the scaling factor, $R$ and $T(t_x, t_y, t_z)$ are the 3D face rotation matrix and translation matrix, respectively. Because the rotation matrix $R$ can be simply parameterized as three Euler Angles $(\alpha, \beta, \gamma)$, the pose of the 3D face can be re-estimated as 7 parameters: $\alpha, \beta, \gamma, t_x, t_y, t_z$, 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) [36, 37] to approximate the scene of diffuse illumination. The per-vertex shaded diffuse albedo color $c_d$ is then computed as:

$$c_d = t_d \cdot \sum_{b=1}^{B^2} \lambda_b \cdot \Pi_b(\bar{n}) \cdot \varphi,$$

where $t_d$ is the vertex diffuse albedo, $\bar{n}$ is the normalized vertex normal, $\Pi : \mathbb{R}^3 \rightarrow \mathbb{R}$ are SH basis functions, and $\lambda$ 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:

$$c_s = t_s \cdot [\text{Nor}(\bar{v}_{\text{out}} - \bar{v}_{\text{in}}) \cdot \bar{n}] \cdot \varphi,$$

where $\text{Nor}()$ represents the normalization process, $t_s$ is the vertex specular albedo, $\bar{v}_{\text{in}}$ is the normalized incoming light vector calculated by the predicted light source position $p_l$, $\bar{v}_{\text{out}}$ is the normalized outgoing light vector calculated by the predefined camera position, $\bar{n}$ is the normalized vertex normal, and $\varphi$ 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) [22] 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-supervised 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 [42] is used in our model to generate face identity features during the training. For each input image, we pre-label its face segments, 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:

$$L = \omega_{\text{id}} L_{\text{id}} + \omega_{\text{photo}} L_{\text{photo}} + \omega_{\text{lmk}} L_{\text{lmk}} + \omega_{\text{mouth}} L_{\text{mouth}} + \omega_{\text{reg}} L_{\text{reg}} + \omega_{\text{uniform}} L_{\text{uniform}},$$

where $L_{\text{id}}$ denotes the identity loss, $L_{\text{photo}}$ denotes the photometric loss, $L_{\text{lmk}}$ denotes the landmark re-projection loss, $L_{\text{mouth}}$ denotes the inner mouth loss, $L_{\text{reg}}$ denotes the regularization loss, and $L_{\text{uniform}}$ 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: $\omega_{\text{id}} = 0.2$, $\omega_{\text{photo}} = 1.92$, $\omega_{\text{lmk}} = 1.8e^{-3}$, $\omega_{\text{mouth}} = 1.3e^{-3}$, $\omega_{\text{reg}} = 5.0e^{-4}$, and $\omega_{\text{uniform}} = 5.0$.

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 [42], $\mathcal{F}$, 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, $L_{\text{id}}$, as follows:

$$L_{\text{id}} = 1 - \frac{\mathcal{F}(I_{\text{in}}) \cdot \mathcal{F}(I_{\text{r}})}{\|\mathcal{F}(I_{\text{in}})\|_2 \cdot \|\mathcal{F}(I_{\text{r}})\|_2},$$

where $\| \cdot \|_2$ denotes the $L_2$-norm, and $I_{\text{in}}$ and $I_{\text{r}}$ 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 [42] 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

Figure 2: Schematic view of our method. We use the ResNet-50 (DCNNs) [22] as the backbone of our network.
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, $I_{face}$, 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 color information than the specular albedo, we construct a weighted mask (refer to Eq. 9) based on $I_{face}$, 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 $I_{diffuse}$ and $I_{specular}$. Meanwhile, a projected face region $I_{project}$ is obtained. Our photometric loss $L_{\text{photo}}$ only focuses on the intersection of $I_{face}$ and $I_{project}$, denoted as $M$. Formally, $L_{\text{photo}}$ is defined as follows:

$$L_{\text{photo}} = \frac{\sum_{i \in M} C_{\text{skin},i} \cdot \|I_{in,i} - I_{r, \text{diffuse},i}\|_2}{\sum_{i \in M} C_{\text{skin},i}} + \eta \frac{\sum_{i \in M} \|I_{in,i} - (I_{r, \text{diffuse},i} + I_{r, \text{specular},i})\|_2}{|M|},$$  

(8)

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

$p_i$ is the probability of the pixel $i$ is skin, $|M|$ denotes the size of $M$.

We adopt a naive Bayes classifier with Gaussian Mixture Models (GMMs) to compute $p_i$ for each pixel $i$ in $I_{face}$. 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 $L_{\text{lmk}}$ measures the Euclidean distance between the labeled 2D landmarks $k_i$ on input images and the projections of corresponding landmarks on 3D morphable face model $k_i^p$ as follows:

$$L_{\text{lmk}} = \frac{1}{|\mathcal{K}|} \sum_{i \in \mathcal{K}} \mu_i \|k_i - k_i^p\|_2 ,$$

(9)

where $\mathcal{K}$ is the set of landmarks, and $\mu_i$ 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 $L_{\text{mouth}}$ is computed as the difference between the projected inner mouth area $A^p$ and the labeled inner mouth area $A$ as follows:

$$L_{\text{mouth}} = \|A - A^p\|,$$

(10)

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 $L_{\text{reg}}$ is formulated as follows:

$$L_{\text{reg}} = \sum ||c_e||^2 + ||c_i||^2 + ||c_\gamma||^2$$

(11)

where $c_i$ represents shape coefficients, $c_e$ represents albedo coefficients, and $c_\gamma$ 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 $L_{\text{uniform}}$ is defined as follows:

$$L_{\text{uniform}} = \frac{1}{|T|} \sum_{t \in T} (l_t - \text{mean})^2 + \sum_{b=1}^{B^2} \sum_{i=1}^{3} l_{b,i}^2 - l_{b}^2,$$

(12)

where $T$ denotes the lit diffuse albedo colors, $\lambda$ 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. [42] and trained it with the triplet loss on the Microsoft Celeb (Ms-Celeb-1M) dataset [21], which contains about 10 millions face
images of 86 thousands identities. We employed a weighted U-Net [53] combined with a pixel-level cross entropy loss to train our face segmentation network on the CelebAMask-HQ dataset [26], which contains about 30 thousands images and 19 classes labels. To detect and align all face images, we adopted the method proposed by Zhang et al. [52] with the image size of 224 × 224. To detect 68 IBUG facial landmarks [40], we employed the method in [4] and trained it on the 300W-LP dataset [54]. ResNet [22] 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 CelebA dataset [30], which contains about 200 thousand images of 10 thousand identities. We randomly picked 2000 images from the CelebA dataset for each test set (called the CelebA test set in this paper). Our model training used the Adam optimizer with batch size of 32, learning rate of 5 × 10−5, and 200 thousand iterations.

5 RESULTS AND EVALUATIONS

Figure 3: Example comparisons between our full model (b-d) and the model without L_uniform (e-g). (a): input face images (with different face poses and expressions) selected from the CelebA test set. (b): the lit diffuse albedo colors produced by our full model. (c): the lit specular albedo colors produced by our full model. (d): the reconstructed faces produced by our full model. (e): the lit diffuse albedo colors produced by our model without L_uniform. (f): the lit specular albedo colors produced by our model without L_uniform. (g): the reconstructed faces produced by our model without L_uniform.

5.1 Ablation Study

We conducted multiple ablation experiments on the CelebA 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 colors (red dotted box in Figure 3). As such, the remaining information 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: The mean errors and standard deviations of pixel colors on the CelebA test set

<table>
<thead>
<tr>
<th>L_id</th>
<th>L_photo</th>
<th>L_lmk</th>
<th>L_mouth</th>
<th>L_reg</th>
<th>L_uni</th>
<th>Mean ± Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.115 ± 0.169</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.118 ± 0.172</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.228 ± 0.247</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>NaN</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>0.117 ± 0.170</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>0.148 ± 0.192</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>0.123 ± 0.179</td>
</tr>
</tbody>
</table>

Table 1 shows the albedo color accuracies on the CelebA test set by 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, 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 [50]. 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.

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 [44], the weakly-supervised learning (WSL) based method [7], PRNet [13], and RingNet [41].

Quantitative shape accuracy on the MICC dataset. We first quantitatively compared the reconstructed shape accuracy among our method and 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 [41], 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.

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 [50].
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 Dataset

<table>
<thead>
<tr>
<th>Methods</th>
<th>Cooperative</th>
<th>Indoor</th>
<th>Outdoor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>1.60 ± 0.55</td>
<td>1.63 ± 0.52</td>
<td>1.68 ± 0.65</td>
</tr>
<tr>
<td>3DDFA_V2 [20]</td>
<td>1.66 ± 0.56</td>
<td>1.67 ± 0.52</td>
<td>1.71 ± 0.66</td>
</tr>
<tr>
<td>MGCNet [44]</td>
<td>1.78 ± 0.55</td>
<td>1.78 ± 0.54</td>
<td>1.81 ± 0.59</td>
</tr>
<tr>
<td>WSL-based [7]</td>
<td>1.68 ± 0.52</td>
<td>1.67 ± 0.54</td>
<td>1.73 ± 0.62</td>
</tr>
<tr>
<td>PRNet [13]</td>
<td>2.01 ± 0.65</td>
<td>2.00 ± 0.58</td>
<td>2.13 ± 0.66</td>
</tr>
<tr>
<td>RingNet [41]</td>
<td>1.72 ± 0.58</td>
<td>1.73 ± 0.60</td>
<td>1.75 ± 0.59</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Neutral</th>
<th>Mouth Stretch</th>
<th>Grin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>2.03 ± 0.66</td>
<td>2.14 ± 0.77</td>
<td>2.16 ± 0.75</td>
</tr>
<tr>
<td>3DDFA_V2 [20]</td>
<td>2.28 ± 0.70</td>
<td>2.25 ± 0.76</td>
<td>2.25 ± 0.73</td>
</tr>
<tr>
<td>MGCNet [44]</td>
<td>2.53 ± 0.75</td>
<td>2.56 ± 0.64</td>
<td>2.56 ± 0.69</td>
</tr>
<tr>
<td>WSL-based [7]</td>
<td>2.26 ± 0.66</td>
<td>2.26 ± 0.71</td>
<td>2.29 ± 0.70</td>
</tr>
<tr>
<td>PRNet [13]</td>
<td>2.81 ± 0.85</td>
<td>2.88 ± 0.89</td>
<td>2.87 ± 0.86</td>
</tr>
<tr>
<td>RingNet [41]</td>
<td>2.42 ± 0.67</td>
<td>2.38 ± 0.74</td>
<td>2.40 ± 0.75</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Neutral</th>
<th>Mouth Stretch</th>
<th>Grin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>3.43 ± 0.36</td>
<td>3.55 ± 0.32</td>
<td>3.64 ± 0.38</td>
</tr>
<tr>
<td>3DDFA_V2 [20]</td>
<td>3.59 ± 0.36</td>
<td>3.58 ± 0.36</td>
<td>3.87 ± 0.45</td>
</tr>
<tr>
<td>MGCNet [44]</td>
<td>3.64 ± 0.38</td>
<td>3.58 ± 0.36</td>
<td>3.87 ± 0.45</td>
</tr>
<tr>
<td>WSL-based [7]</td>
<td>3.64 ± 0.38</td>
<td>3.58 ± 0.36</td>
<td>3.87 ± 0.45</td>
</tr>
<tr>
<td>PRNet [13]</td>
<td>3.64 ± 0.38</td>
<td>3.58 ± 0.36</td>
<td>3.87 ± 0.45</td>
</tr>
</tbody>
</table>

5.3 Qualitative Evaluation

We also conducted qualitative evaluations to visually compare the face reconstruction results between our method and the above five state-of-the-art methods [1, 20, 44, 7, 13, 41] on 9 face images randomly selected from the CelebA test set. Compared to the methods in [44] and the WSL-based method [7], our method can better reconstruct the expressive faces, especially in the areas of the eyes and mouth. Comparisons are shown in Figure 7. Our method can recover accurate albedo colors and lighting conditions and more realistic face shapes. Figure 9 shows some comparison results among our method, the MoFA method [48], and the method in [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 [48] and the method in [17].

6 DISCUSSION AND CONCLUSION

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

Our current method faces several limitations. First, it cannot model multiple light sources in an input image. Therefore, certain specular...
Acknowledgments
This work was in part supported by NSF IIS-2005430, and a research grant from the SEED lab of Electronic Arts.


