Photo-Realistic Facial Details Synthesis from Single Image

Anpei chen Zhang Chen Guli Zhang Ziheng Zhang
Kenny Mitchell Jingyi Yu
shanghaiTech University

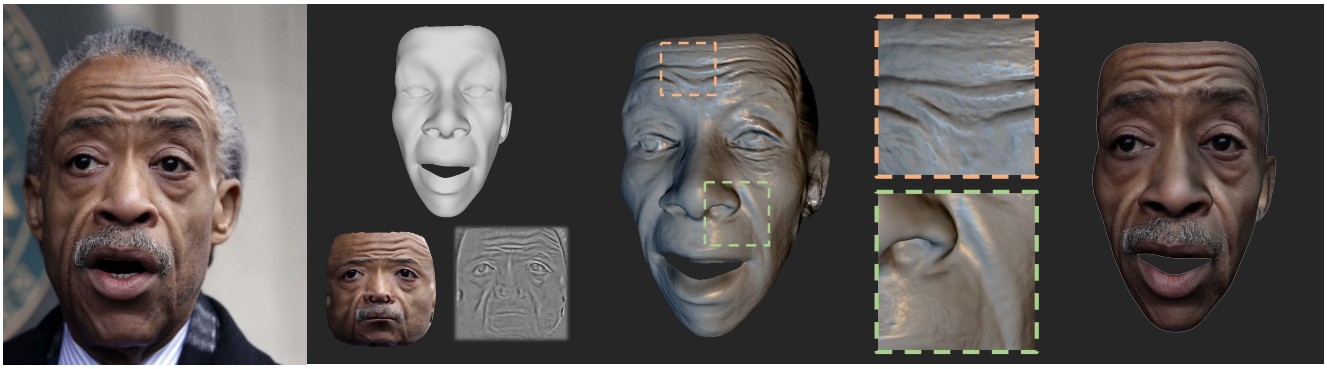

Figure 1. From left to right: An input face image, proxy 3D face, texture and displacement map produced by our Emotion-to-Face (E2F) framework, detailed face geometry using our learning based synthesis technique, and the re-rendered result.

Abstract

We present a single-image 3D face synthesis technique that can handle challenging facial expressions while recovering fine geometric details. Our technique employs expression analysis for proxy face geometry generation and combines supervised and unsupervised learning for facial detail synthesis. On proxy generation, we conduct emotion prediction to determine a new expression-informed proxy. On detail synthesis, we present a Deep Facial Detail Net (DFDN) based on Conditional Generative Adversarial Net (CGAN) that employs both geometry and appearance loss functions. For geometry, we capture 366 high-quality 3D scans from 122 different subjects under 3 facial expressions. For appearance, we use additional 20K in-the-wild face images and apply image-based rendering to accommodate lighting variations. Comprehensive experiments demonstrate that our framework can produce high-quality 3D faces with realistic details under challenging facial expressions.

1. Introduction

Producing high quality human faces with fine geometric details has been a core research area in computer vision and graphics. Geometric structure details such as wrinkles are important indicators of age and facial expression and are essential for producing realistic virtual human [2]. Successful solutions by far rely on complex and often expensive capture systems such as stereo-based camera domes [20] or photometric-based LightStage [32, 12]. Although such solutions have become increasingly popular and affordable with the availability of low-cost cameras and lights, they are still bulky and hence do not support portable scanning. In addition, they are vulnerable to low texture regions such as bare skins.

We aim to produce high-quality 3D faces with fine geometric details from a single image, with quality comparable to those produced from the dome systems and LightStage. Existing single-image solutions first construct a 3D proxy face from templates and then refine the proxy by deforming geometry and adding details. Such proxies can be derived from 3D Morphable Model (3DMM) [10, 9, 42, 51, 16] by blending base face geometry. More complex techniques employ sparse coding on 3D face dictionaries to further improve robustness and quality [41, 10, 9, 42, 51, 16, 23, 43, 7]. However, artifacts arise from these approaches such as over-smoothing and incorrect expression, where a relatively small number of parameters are used to approximate the high dimensional space for real face. Shape-from-shading [28], photometric stereo [12], and deep learning [52, 39, 14]

have been used to generate the missing details. However, existing methods have limits in attaining correct shape under unseen emotional expressions and lighting, thus delivering insufficient or inaccurate geometric details, as shown in Fig. 7

In this paper, we present a novel learning-based technique to produce accurate geometric details from a single face image. Our approach takes into account emotion, expression and appearance. For proxy generation, we employ the Basel Face Model (BFM) [37] composed of shape, expression and surface reflectance (albedo). 3D expressions, however, exhibit strong ambiguity after being projected onto 2D images: a pair of 3D meshes that represent very different emotional expressions can have similar 2D landmarks on images. Therefore, we first devise a learning-based approach to conduct emotion prediction and then use the result to determine an expression informed proxy.

For geometric detail synthesis, we devise a Deep Facial Detail Net (DFDN) based on Conditional Generative Adversarial Net (*CGAN*) to map an image patch to a detailed displacement map. Our DFDN has two components: a medium scale geometry module that learns the PCA coefficients (in our case 64) of each patch and a fine scale geometry module that refines the PCA based result with additional details. For training, we captured a total of 366 high quality 3D scans from 122 different subjects under three facial expressions (one neutral and two extreme expressions). We augment the training data with 340 high resolution meshes from ICT-3DRFE [47]. The loss function is defined in terms of geometric differences between the estimation and the ground truth. However, we observe that these training data are still insufficient to cover a wide range of lighting conditions. Hence, we introduce an additional unsupervised learning procedure (with an additional 20K images captured in the wild) where for each image we obtain its proxy geometry using our emotion-driven shape estimator and then approximate the corresponding environment lighting using spherical harmonics (SH). We use DFDN to obtain an estimate of the geometry, but since we do not have the ground truth geometry, we re-render these results using the estimated albedo and environment lighting, and compute the loss function in terms of the image differences. Finally, we alternate the supervised and the unsupervised learning processes, on geometry and image, respectively.

2. Related Work

Existing approaches for producing high quality 3D face geometry either rely on reconstruction or synthesis.

Reconstruction-based Techniques. Multi-View Stereo (MVS) 3D face reconstruction systems employ stereo [33] or structure-from-motion [57]. A sparse set of cameras produce large scale geometry [20] whereas denser and hence more expensive settings [2] provide more accurate measures. In either case, the reconstruction quality depends heavily on the feature matching results as they act as anchor points dominating the final shape. For regions with few textures such as bare skin, the reconstruction tends to be overly smooth due to lack of features. For example, wrinkles caused by facial expressions are particularly difficult to reconstruct: even though they cause shading variations, their geometry is too slight to capture using stereo, especially when the camera baseline is small. Recently, Graham et al. [20] use 24 entry-level DSLR photogrammetry cameras and 6 ring flashes to capture facial specular response independently and then combine shape-from-chroma and shape-from-specularity for high quality reconstruction.

Another class of multi-shot techniques employed in face reconstruction is Photometric Stereo (PS). PS is based on analyzing image intensity variations under different illuminations from a fixed viewpoint. Instead of directly reconstructing 3D geometry, PS intends to first recover the normal map and then the 3D mesh, e.g., via normal integration. A common artifact in PS is low-frequency distortions in the final reconstruction [35, 48] caused by perspective projection violating the orthographic assumption. Accurate calibrations on both the light sources and camera, though able to mitigate the problem, are cumbersome. Most recent techniques [36, 58, 15] combine PS with MVS by using the MVS results as a proxy for calibration and then refine the results. Aliaga et al. [3] simulates a MVS setup by employing multiple digital projectors as both light sources and virtual cameras. We refer the readers to [1] for a comprehensive review of PS variants.

Synthesis-based approaches. The availability of high quality mobile cameras and the demand on portable 3D scanning have promoted significant advances on producing high quality 3D faces from a single image. The seminal work of Blanz and Vetter [5] pre-captures a database of face models and extracts a 3D morphable model (3DMM) composed of base shapes and albedos. Given an input image, it finds the optimal combination of the base models to fit the input. Their technique can also handle geometric deformations under expression [37, 18] if the database includes expressions, e.g., captured by RGBD cameras [11]. More extensive facial databases have been recently made publicly available [59, 24, 55, 30, 6], with an emphasis on handling complex expressions [30, 6]. Most recently, Li *et al.* [30] capture pose and articulations of jaw, neck, and eyeballs with over 33,000 3D scans that have helped boost the performance of single-image/video face reconstruction/tracking [41, 10, 9, 46, 61, 42, 51, 16, 23, 43, 7]. The current databases, however, still lack mid- and high-frequency geometric details such as wrinkles and pores that are epitomes to realistic 3D faces. Shading based compensations can improve the visual appearance [16, 27] but remain far behind quality reconstruction of photos.

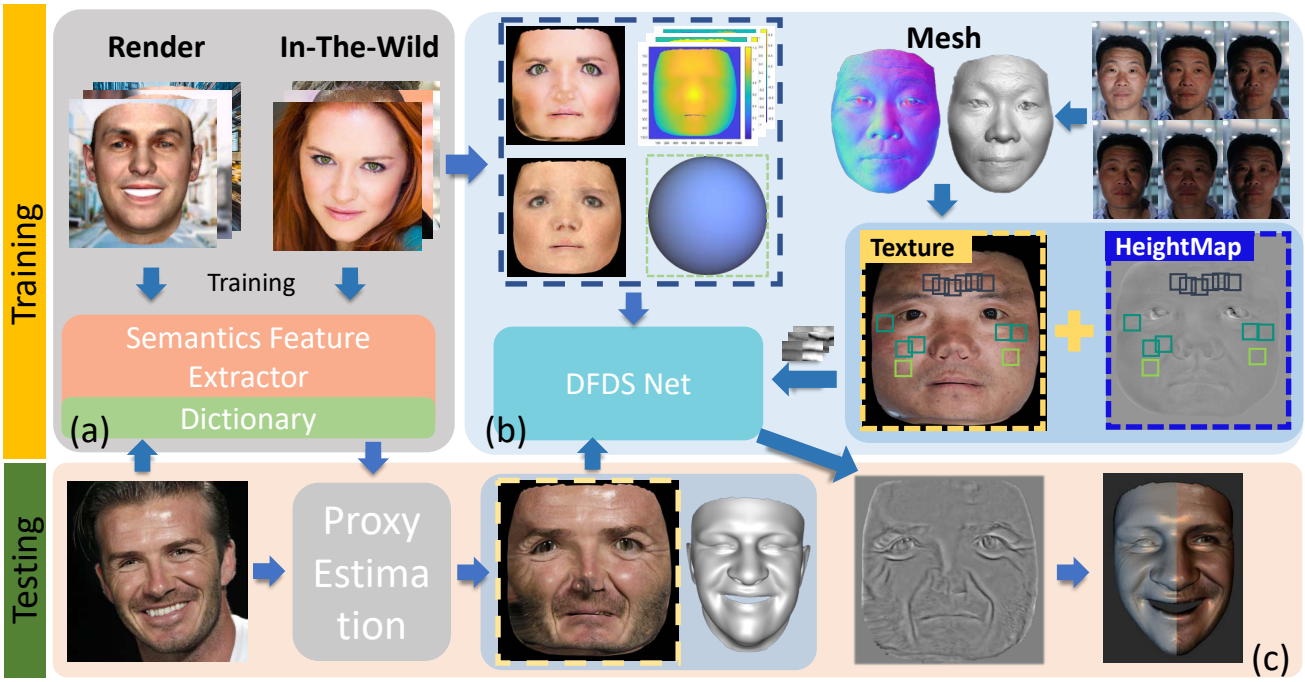

Figure 2. Our processing pipeline. Top: training stage for emotion-driven proxy generation (a) and facial detail synthesis (b). Bottom: applying the trained network on a real image.

Our approach is part of the latest endeavor that uses learning to synthesize fine geometric details from a single image. Real [53] or synthetically rendered [39, 14] face images were used as training datasets, and convolutional neural networks (CNNs) are then used to estimate 3D model parameters. Sela *et al.* [44] use synthetic images for training but directly recover depth and correspondence maps instead of model parameters. Richardson *et al.* [40] apply supervised learning to first recover model parameters and then employ shape-from-shading (SfS) to recover fine details. Guo *et al.* [21] adopts a two-stage network to reconstruct facial geometry at different scales. Tewari *et al.* [50] adopts a self-supervised approach based on an autoencoder where a novel decoder depicts the image formation process. Kim *et al.* [29] combine the advantages of both synthetic and real data to [49] jointly learn a parametric face model and a regressor for its corresponding parameters. Learning based techniques can also produce volumetric representations [26] or normal fields [45]. Yet, few approaches can generate very fine geometric details. Cao *et.al* [8] capture 18 high quality scans and employ a principal component analysis (*PCA*) model to emulate wrinkles as displacement maps. Huynh *et al.* [25] use high precision 3D scans from the LightStage [19]. Although effective, their technique assumes similar environment lighting as the LightStage.

3. Expression-Aware Proxy Generation

Our first step is to obtain a proxy 3D face with surface albedo map and accurate facial expressions. We employ the *Basel Face Model (BFM)* [37], which consists of three components: shape M_{sha}, expression M_{exp} and albedo M_{alb}. Shape M_{sha} and expression M_{exp} determine vertex locations while albedo M_{alb} encodes per-vertex albedo:

$$M_{sha}(\alpha) = a_s + E_s \cdot \alpha \quad (1)$$

$$M_{exp}(\beta) = a_e + E_e \cdot \beta \quad (2)$$

$$M_{alb}(\gamma) = a_{alb} + E_{alb} \cdot \gamma \quad (3)$$

where $a_s, a_e, a_{alb} \in \mathbb{R}^{3n}$ represent the mean of corresponding *PCA* space. $E_s \in \mathbb{R}^{3n\times199}$, $E_e \in \mathbb{R}^{3n\times100}$ contain basis vectors for shape and expression while $E_{alb} \in \mathbb{R}^{3n\times199}$ contain basis vectors for albedo. α, β, γ correspond to the parameters of the PCA model.

3.1. Proxy Estimation

Given a 2D image, we first extract 2D facial landmarks $\mathbf{L} \in \mathbb{R}^{2m}$ and use the results to compute *PCA* parameters α, β for estimation of proxy shape. Specifically, we set out to find the parameters that minimize the reprojection error on landmarks:

$$E = \sum_k w_k \|\mathbf{L}_k - P(\mathbf{l}_k(\alpha,\beta))\|_2 + \lambda_s\|\alpha\|_2 + \lambda_e\|\beta - \beta_{prior}\|_2 \quad (4)$$

where $\mathbf{l}_k(\alpha, \beta)$ corresponds to the kth facial vertex landmark and $P(\cdot)$ is the camera projection operator that maps 3D vertices to 2D image coordinates. w_k controls the weight for each facial landmark whereas λ_s, λ_e imposes regularization on the parameters. β_{prior} denotes the emotion-based expression prior that we will describe in Section 3.2.

To solve for Eq. 4, we use the iterative linear method [24]. Specifically, the camera projection operator $P(\cdot)$ is parameterized as an affine camera matrix. For each round of iterations, we first fix α, β and solve for $P(\cdot)$ using the *Gold Standard Algorithm* [22]. We then fix $P(\cdot)$ and solve for α, β. To bootstrap this iterative scheme, we initialize α, β as $\mathbf{0}$.

3.2. Imposing Expression as Priors

The most challenging component in proxy estimation is expression. 3D expressions exhibit a significant ambiguity after being projected onto 2D images, e.g., different expressions may have similar 2D facial landmarks after projection. Fig. 3 shows an example of this ambiguity: the landmarks of the two faces are extremely close to each other while their expression parameters and shapes are vastly different, especially near the facial decree areas. So it is hard to define or train a mapping directly from image to 3D expression. In our experiments, we also observe that the reprojection-based loss function can easily fall into local minimum that reflects such ambiguity.

We propose to use facial semantic information to narrow the 3D expression parameter solving space via converting the problem into a conditional distribution. Our facial expression features include physical based appearance features(e.g. gender,age and FACS) and underlay high level emotion features.

To obtain the emotion features, we reuse a 11 discrete expression emotion dataset [34] to train an emotion feature predictor *Emotion-Net*. Next, we randomly generate expression parameters β from normal distribution in interval $[-3, 3]$ and render 90K images with different facial expressions. We feed the images into the trained *Emotion-Net* and obtain a total of 90K emotion feature vectors. Using these emotion feature vectors along with their corresponding appearance feature parameters, we formulate a dictionary $\Psi: \Psi_{emo} \rightarrow \Psi_{exp}$ that record semantics features Ψ_{emo} to expression parameters Ψ_{exp}. We utilize the 18-layer Residual Net (*ResNet-18*) to train our emotion feature predictor and use the output of the second last layer $f \in \mathbb{R}^{512}$ as the feature vectors to represent human emotions. Once we obtain the trained model and the expression dictionary, we can predict expression parameters β_{prior} as a prior for proxy estimation.

Given a new image I, we first feed it to *Emotion-Net* and physical appearance features predictor[4] to obtain its semantics feature vector. We then find its closest semantics vector in the dictionary and use the corresponding expression parameters for β_{prior}:

$$\beta_{prior} = \Psi(\arg\min_{\psi_{emo}} \|\textit{Emotion-Net}(I) - \psi_{emo}\|_2) \quad (5)$$

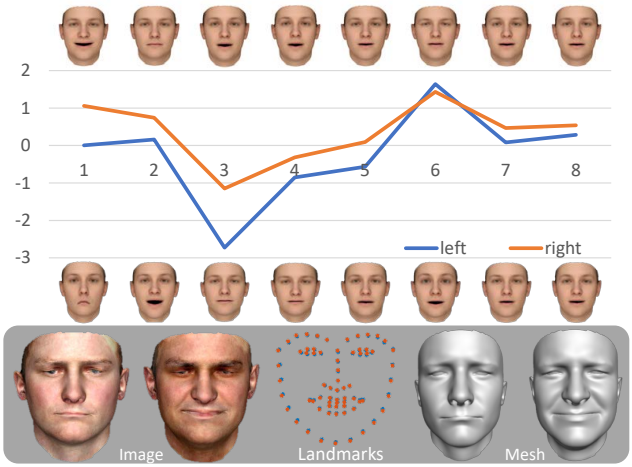

Figure 3. Expression projection ambiguity. Top: Visualization of the two models' first eight dimensions of 3D expression parameters. Bottom: 2D facial images, their landmarks layered onto one image, and the corresponding geometric. 'left, right' refer to both rendered images and meshes.

4. Deep Facial Detail Synthesis

With the 3D proxy face, we synthesize geometric details by estimating displacement map and applying to the proxy mesh. The key observation here is that for facial details such as wrinkles, there is strong correlation between geometry and appearance.

4.1. Network Architecture

Fig. 4 shows our *Deep Facial Detail Net* (*DFDN*) with two main cascaded modules. The *Partial Detail Inference Module* (*PDIM*) takes 2D image patches as inputs and generates 3D facial geometric details using a *PCA*-based technique (Section 4.2). Such a scheme dramatically reduces the parameter space and is stable for both training and inference process. However, *PCA*-based approximations lose high frequency features that are critical to fine detail synthesis. We therefore introduce the *Partial Detail Refinement Module* (*PDRM*) to further refine high-frequency details. The reason we explicitly break down facial inference procedure into linear approximation and non-linear refinement is that facial details consist of both regular patterns like wrinkles and characteristic features such as pores and spots. By using a two-step scheme, we encode such priors into our network.

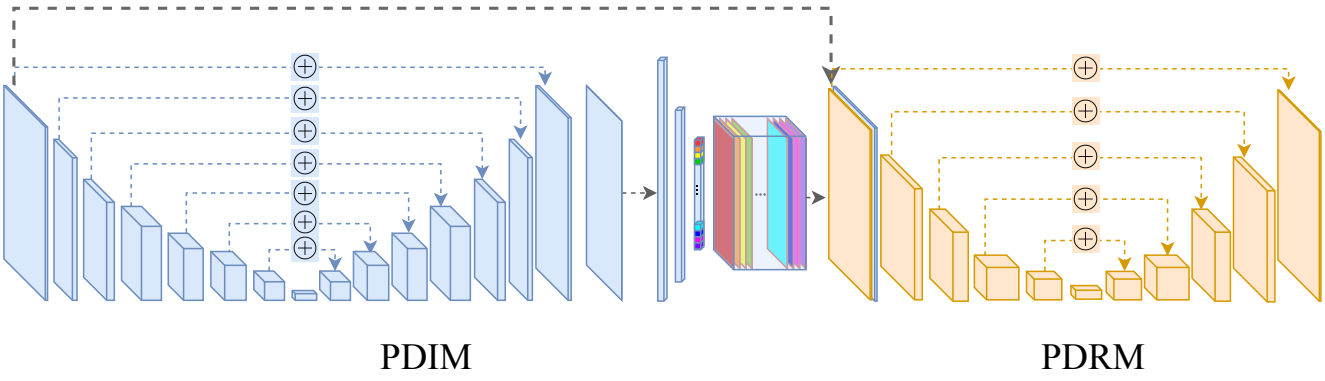

Figure 4. Network architecture for facial detail synthesis. PDIM for median frequency detail (wrinkles) synthesis and PDRM for high frequency detail (pores) synthesis.

In *PDIM* module, we use *UNet-8* structure concatenated with 4 fully connected layers to learn the mapping from texture map to *PCA* representation of displacement map. The sizes of the 4 fully connected layers are $2048, 1024, 512, 64$. Except for the last fully connected layer, each linear layer is followed by an *ReLU* activation layer. In the subsequent *PDRM* module, we use *UNet-6*, i.e. 6 layers of convolution and deconvolution, each of which uses 4×4 kernel, 2 for stride size, and 1 for padding size. Apart from this, we adopt *LeakReLU* activation layer except for the last convolution layer and then employ *tanh* activation.

To train *PDIM* and *PDRM* modules, we combine supervised and unsupervised training techniques based on Conditional Generative Adversarial Nets (*CGAN*), aiming to handle variations in facial texture, illumination, pose and expression. Specifically, we collect 706 high precision 3D human faces and over 20K unlabeled facial images captured in-the-wild to learn a mapping from the observed image x and the random noise vector z to the target displacement map y by minimizing the generator objective G and maximizing log-probability of 'fooling' discriminator D as:

$$\mathcal{L} = \arg\min_G \max_D (\mathcal{L}_{cGAN(G,D)} + \lambda \mathcal{L}_{L1(G)}), \tag{6}$$

where we set $\lambda = 100$ in all our experiments and

$$\begin{aligned}\mathcal{L}_{cGAN(G,D)} =& \mathbb{E}_{x,y}[log\ D(x,y)] + \\ & \mathbb{E}_{x,z}[log(1 - D(x, G(x,z)))].\end{aligned} \tag{7}$$

A major drawback of the supervised learning scheme mentioned above is that the training data, captured under fixed setting (controlled lighting, expression, etc.), are insufficient to emulate real face images that exhibit strong variations caused by environment lighting and expressions. We hence devise a semi-supervised generator G, exploiting labeled 3D face scans for supervised loss L_{scans} as well as image-based modeling and rendering for unsupervised reconstruction loss L_{recon} as:

$$\mathcal{L}_{L1(G)} = L_{scans}(x,z,y) + \eta L_{recon}(x,z,y) \tag{8}$$

Where η controls the contribution of reconstruction loss and we fix it as 0.5 in our case. In the following subsections, we discuss how to construct the supervised loss L_{scans} for geometry and unsupervised loss L_{recon} for appearance.

4.2. Geometry Loss

The geometry loss compares the estimated displacement map with ground truth. To do so, we need to capture ground truth facial geometry with fine details.

Face Scan Capture. To acquire training datasets, we implement a small-scale facial capture system similar to [12] and further enhance photometric stereo with multi-view stereo: the former can produce high quality local details but is subject to global deformation whereas the latter shows good performance on low frequency geometry and can effectively correct deformation.

Our capture system contains 5 Cannon 760D DSLRs and 9 polarized flash lights. We capture a total of 23 images for each scan, with uniform illumination from 5 different viewpoints and 9 pairs of vertically polarized lighting images (only from the central viewpoint). The complete acquisition process only lasts about two seconds. For mesh reconstruction, we first apply multi-view reconstruction on the 5 images with uniform illumination. We then extract the specular/diffuse components from the remaining image pairs and calculate diffuse/specular normal maps respectively using photometric stereo. The multi-view stereo results serve as a depth prior z_0 for normal integration [38] in photometric stereo as:

$$\begin{aligned}\min \iint_{(u,v)\in I} & [(\nabla z(u,v) - [p(u,v), q(u,v)]^\top)^2 \\ & + \mu(z(u,v) - z_0(u,v))^2] du dv,\end{aligned} \tag{9}$$

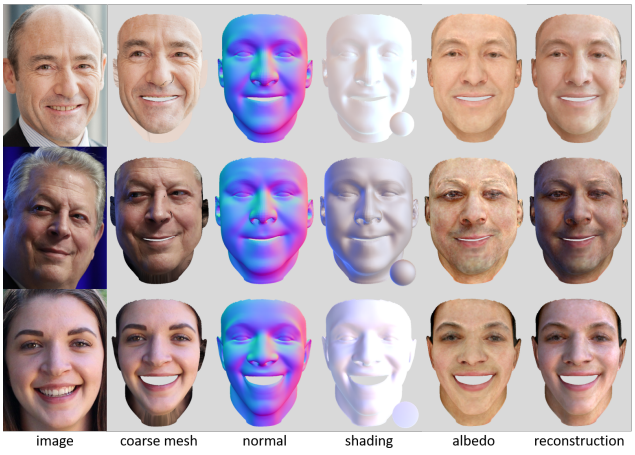

Figure 5. From left to right: from the image, we estimate its proxy mesh, normal, lighting/shading, and albedo to re-render an image.

where u, v represents image coordinates, p, q represents approximations to $\partial_u z$ and $\partial_v z$ respectively, z_0 is the depth prior. μ controls the contribution of prior depth z_0. In order to generate geometry pairs with and without details, we set the weight parameter μ to $1e^{-5}$ and $1e^{-3}$ respectively. Then we obtain a ground truth displacement map for each geometry pair.

PCA-Based Displacement Map. In our training scheme, we choose not to directly feed complete face images as inputs to the network: such training can easily cause overfitting since we do not have sufficient 3D face models with fine details to start with. Instead, we observe that despite large scale variations on different faces, local texture details present strong similarities even if the faces appear vastly different. Hence we adopt the idea from [8, 12] to enhance our network generalization by training the network with texture/displacement patches of 256×256 resolution. We model the displacement using *PCA*, where each patch is a linear combination of 64 basis patches.

Our geometric loss is then defined as:

$$L_{scans}(x,z,y) = \sum ||\mathcal{PDIM}(x,z) - \mathcal{PCA}(y)||_1 + ||\mathcal{PDRM}[\mathcal{PDIM}(x,z)] - y||_1 \quad (10)$$

where we combine the loss in PCA space with the per-pixel loss to recover finer details.

For patch sampling, we unfold each facial image into a 4096×4096 resolution texture map and regionally sample training patches based on semantic facial segmentation. For the sampling scheme, we iteratively reduce the displacement map gradient with a weighted Gaussian kernel for the training set, and uniformly sample patches with 50% overlap in our validations.

4.3. Appearance Loss

Recall that the small amount of labeled facial geometry is insufficient to cover a broad range of illumination conditions and surface reflectance. Thus, we further adopt a rendering-based, unsupervised learning approach: we obtain 20K in-the-wild images, estimate its proxy (using E2F) and geometric details (using *DFDN*), and then use this information to calculate lighting and albedo. Finally, we re-render an image with all these estimations and compute reconstruction loss against the input image. Genova et al. [17] adopted a similar approach by creating a vertex buffer and conducting rasterization and interpolation.

To obtain per-pixel normals with geometric details added, we propose an texture space manipulation using the proxy mesh's position map $\mathcal{P}_{proxy}$ (shown in Fig. 2, the middle of first row) and the output displacement map $G(x,z)$ from *DFDN*:

$$\mathcal{P}_{fine} = \mathcal{P}_{proxy} + G(x,z) * \mathcal{N}_{proxy} \quad (11)$$

$$\mathcal{N}_{fine} = \mathcal{F}(\mathcal{P}_{fine}) \quad (12)$$

where $\mathcal{N}_{proxy}, \mathcal{N}_{fine}$ represent proxy and fine scale geometric normal map and $\mathcal{P}_{fine}$ is position map for fine detailed mesh. $\mathcal{F}(\cdot)$ is normalized cross product operator on position difference:

$$\mathcal{F}(\mathcal{P}_{fine}) = \frac{conv_h(\mathcal{P}_{fine}) \times conv_v(\mathcal{P}_{fine})}{||conv_h(\mathcal{P}_{fine})|| \cdot ||conv_v(\mathcal{P}_{fine}||} \quad (13)$$

We compute position difference via nearby horizontal and vertical 3 pixels in texture space, giving rise to convolution kernels of $[-0.5, 0, 0.5]$ and $[-0.5, 0, 0.5]^\top$ for $conv_h$, $conv_v$ respectively.

With known normal of fine detailed mesh, we assume a Lambertian skin reflectance model and represent the global illumination using Spherical Harmonics (SH) to estimate environment lighting and surface albedo. Under this model, we can compute the radiance L emitting from point v at any viewing direction as:

$$L(v) = \rho(v)S(n_v) = \rho(v)\sum_{i=1}^{n} l_i Y_i(n_v) \quad (14)$$

where $\rho(v)$ denotes the surface albedo, S irradiance, Y_i the basis of spherical harmonics (see details in supplementary material), l_i the corresponding weight, and n_v the normal of vertex v. We also represent albedo using *BFM*:

$$L(v) = (a^v_{alb} + E^v_{alb} \cdot \gamma)\sum_{i=1}^{n} l_i Y_i(n_v) \quad (15)$$

with a^v_{alb} and E^v_{alb} being the mean and principle component albedo at vertex v. We use the first nine harmonic basis and have:

$$L(v) = (a^v_{alb} + E^v_{alb} \cdot \gamma)H_v \cdot l \quad (16)$$

where $H_v = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} Y_1(n_{v_i}) & \cdots & Y_9(n_{v_i}) \end{bmatrix}$ and $l = \begin{bmatrix} l_1^1, & \cdots & l_9^1, & l_1^2, & \cdots & l_9^2, & l_1^3, & \cdots & l_9^3 \end{bmatrix}^T$. Accordingly, a reconstructed face image I_{recon} can be represented by

$$I_{recon} = (a_{alb} + E_{alb} \cdot \gamma) \circ (H \cdot l) \tag{17}$$

where $H = \begin{bmatrix} H_{v_1}^T, & \cdots, & H_{v_n}^T \end{bmatrix}^T$, and $H \in \mathbb{R}^{3n \times 27}$.

We estimate lighting and albedo by minimizing the energy function Eq. 18 on the illumination coefficients l and the albedo parameters γ:

$$E(l, \gamma) = \|I_{input} - I_{recon}\|_2^2 \tag{18}$$

where I_{input} is the intensity value of vertices. In order to achieve a reliable estimation, in our implementation, we first use a self-adaptive mask to select vertices that have reliable normals with which to apply the optimization.

We adopt an iterative optimization scheme similar to [56]. The complete algorithm is shown in **Algorithm** 1 where M, ξ_1, ξ_2 are termination threshold. They are set as 50, 0.05 and 50 in our experiments.

Algorithm 1 lighting and albedo estimation

Require: $I_{input}, H, a_{alb}, E_{alb}, M, \xi_1, \xi_2, i = 0$
Ensure: $l, \gamma = \arg\min_{l,\gamma} E(l, \gamma)$
$i \leftarrow 0$
$\gamma \leftarrow \mathbf{0}$
while $i \leq M$ **do**
  $l \leftarrow \arg\min_l \|I_{input} - (a_{alb} + E_{alb} \cdot \gamma) \circ (H \cdot l)\|_2^2$
  $\delta I \leftarrow I_{input} - (a_{alb} \circ (H \cdot l))$
  $\delta\gamma \leftarrow \arg\min_{\delta\gamma} \|\delta I - (E_{alb}\delta\gamma) \circ (H \cdot l)\|_2^2$
  $\gamma \leftarrow \gamma + \delta\gamma$
  $i \leftarrow i + 1$
  if $\|\delta\gamma\|_2^2 < \xi_1$ **or** $\|\delta I\|_2^2 < \xi_2$ **then return** l, γ
return l, γ

For training, we synthesize high resolution facial images from the emotion dataset *AffectNet* [34] which contains more than 1M facial images collected from the Internet and about half of the retrieved images (440K) were manually annotated. Importantly, *AffectNet* is the largest database of facial expressions, valence, and arousal in the wild.

In our experiments, We also use HSV color space instead of RGB to accommodate environment lighting variations and employ a two-step training approach, i.e. only back propagate PCA parameters loss for the first 10 epochs, which we found the loss reduces much faster than that in directly training. Moreover, we train 250 epochs for each facial area based on our observation that the loss is smaller when there are more epochs, but the mesh is noisier in the experiment. To sum up, our expression estimation and detail synthesis networks borrow the idea of residual learning, breaking down the final target into a few small tasks, which facilitates training and improves performance in our tasks.

5. Experimental Results

In order to verify the robustness of our algorithm, we have tested our emotion-driven proxy generation and facial detail synthesis approach on over 20,000 images (see supplementary material for many of these results).

Expression Generation. We downsample all images from *AffectNet* dataset into 256×256 (the downsampling is only for proxy generation, not for detail synthesis) and randomly sample 10% of the images for the validation set. We use the Adam optimization framework with a momentum of 0.9. We train a total of 150 epochs and set learning rate to be 0.0001 for the first 50 epochs, and gradually reduce it to 0 in the rest epochs. Our trained *Emotion-Net* achieves a test accuracy of 47.1%. Recall that facial emotion classification is a challenging task and even human annotators achieve only 60.7% accuracy. Since our goal focuses on producing more realistic 3D facial models, we find this accuracy is sufficient for producing reasonable expression prior.

Fig. 6 shows some samples of our proxy generation results. Compared with the state-of-the-art solutions of 3D expression prediction [60, 13], we find that all methods are able to produce reasonable results in terms of eyes and mouth shape. However, the results from 3DDFA [60] and ExpNet [13] exhibit less similarity with input images on regions such as cheeks, Nasolabial folds and under eye bags while ours show significantly better similarity and depict person-specific characteristics. This is because such regions are not covered by facial landmarks. Using landmarks alone falls into the ambiguity mentioned in Section 3.2 and cannot faithfully reconstruct expressions on these regions. Our emotion-based expression predictor exploits global information from images and is able to more accurately capture expressions, especially for jowls and eye bags.

Facial Detail Synthesis. We sampled a total of 10K patches for supervised training and 12K for unsupervised training. We train 250 epochs in total, and uniformly reduce learning rate from 0.0001 to 0 starting at 100th epoch. Note, we use supervised geometry loss for the first 15 epochs, and then alternate between supervised geometry loss and unsupervised appearance loss for the rest epochs.

Our facial detail synthesis aims to reproduce details from images as realistically as possible. Most existing detail synthesis approaches only rely on illumination and reflectance model [31, 54]. A major drawback of these methods lies in that their synthesized details resemble general object surface without considering skin's spatial correlation, as shown

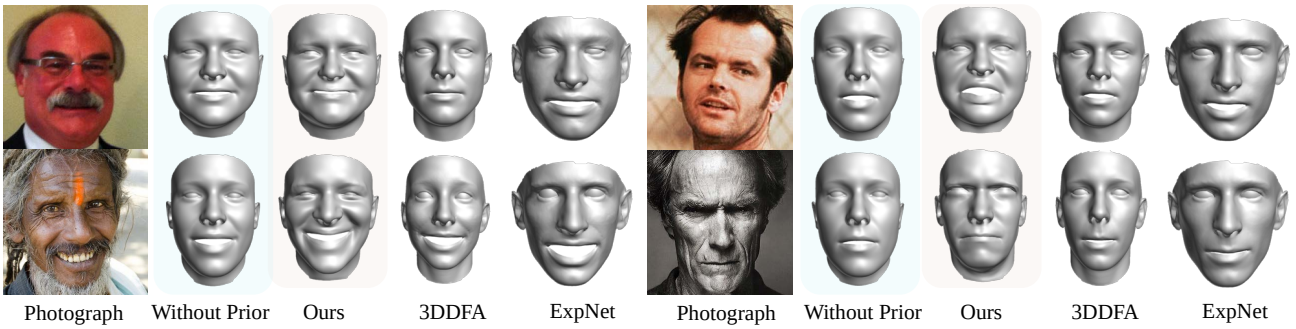

Figure 6. Comparisons of our emotion-driven proxy estimation vs. the state-of-the-art (3DDFA [60] and ExpNet [13])

.

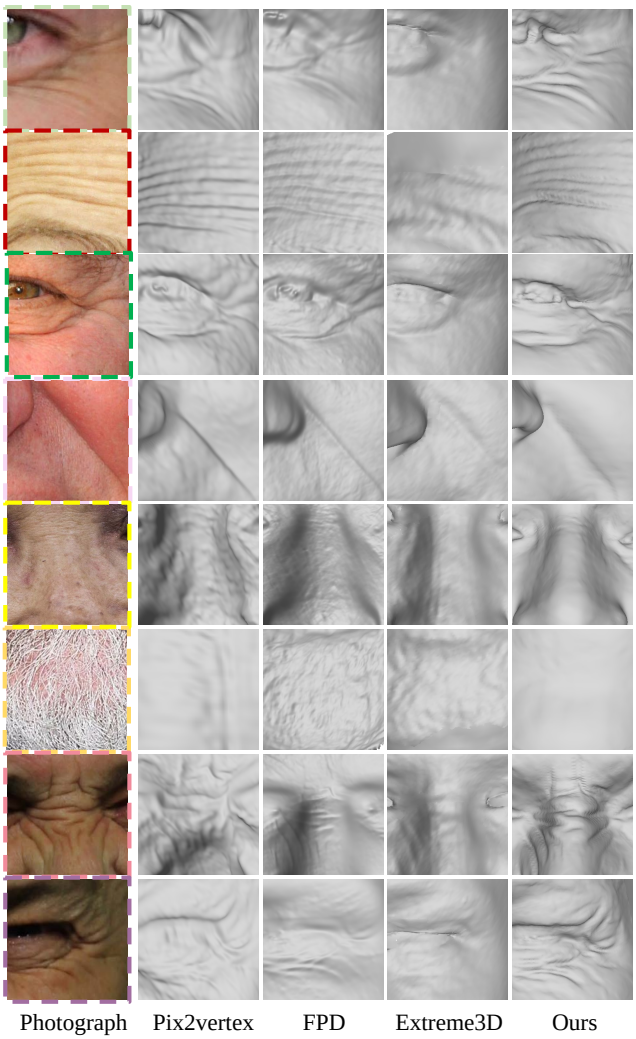

Figure 7. Close-up views of the synthesized meshes using Pix2vertex [44], FPD [31], Extreme3D [54] and ours.

in close-up views in Fig. 7 (full mesh in supplementary material). Our wrinkles are more similar to real skin surface while the other three approaches are more like cutting with a knife on the surface. We attribute this improvement to combining illumination model with human face statistics from real facial dataset and wrinkle PCA templates.

Our approach also has better performance on handling the surface noise from eyebrows, braids and beards while preserving skin details (2, 5 and 6th row of Fig. 7). Finally, our output displacement map is easy to integrate with existing rendering pipelines and can produce high-fidelity results, as shown in Fig. 1.

6. Conclusion and Future Work

We have presented a single-image 3D face synthesis technique that can handle challenging facial expressions while preserving fine geometric structures. Our technique combines cues provided by emotion, expression, appearance, and lighting for producing high fidelity proxy geometry and fine geometric details. Specifically, we have conducted emotion prediction to obtain an expression informed proxy and we have demonstrated that our approach can handle a wide range of expressions. For detail synthesis, our Deep Facial Detail Net (DFDN) employs both geometry and appearance loss functions and was trained on both real captured and synthesized data from in-the-wild images. Comprehensive experiments have shown that our technique can produce, from a single image, ultra high quality 3D faces with fine geometric details under various expressions and lighting conditions.

Although our solution is capable of handling a variety of lighting conditions, it has not yet considered the effects caused by occlusions (e.g., shadows), skin scattering (e.g., specularity), or non-facial objects (hair or glasses) that may cause incorrect displacement estimations. For shadows, it may be possible to directly use the proxy to first obtain an ambient occlusion map and then correct the image. Shadow detection itself can be directly integrated into our learning-based framework with new sets of training data. Another

limitation of our technique is that it cannot tackle low resolution images: our geometric detail prediction scheme relies heavily on reliable pixel appearance distribution. Two specific types of solutions we plan to investigate are to conduct (learning-based) facial image super-resolution that already accounts for lighting and geometric details as our input and to design a new type of proxy face model that includes deformable geometric details.

References

[1] J. Ackermann, M. Goesele, et al. A survey of photometric stereo techniques. *Foundations and Trends® in Computer Graphics and Vision*, 9(3-4):149–254, 2015. 2

[2] O. Alexander, M. Rogers, W. Lambeth, J.-Y. Chiang, W.-C. Ma, C.-C. Wang, and P. Debevec. The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications*, 30(4):20–31, 2010. 1, 2

[3] D. G. Aliaga and Y. Xu. A self-calibrating method for photogeometric acquisition of 3d objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):747–754, 2010. 2

[4] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE, 2015. 4

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2

[6] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018. 2

[7] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Poumpis, Y. Panagakis, and S. P. Zafeiriou. 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2

[8] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):46, 2015. 3, 6

[9] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):43, 2014. 1, 2

[10] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 1, 2

[11] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 2

[12] X. Cao, Z. Chen, A. Chen, X. Chen, S. Li, and J. Yu. Sparse photometric 3d face reconstruction guided by morphable models. In *CVPR*, 2018. 1, 5, 6

[13] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni. Expnet: Landmark-free, deep, 3d facial expressions. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 122–129. IEEE, 2018. 7, 8

[14] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–26, 2017. 1, 3

[15] C. H. Esteban, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008. 2

[16] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):28, 2016. 1, 2

[17] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 6

[18] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 75–82. IEEE, 2018. 2

[19] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)*, volume 30, page 129. ACM, 2011. 3

[20] P. Graham, G. Fyffe, B. Tonwattanapong, A. Ghosh, and P. Debevec. Near-instant capture of high-resolution facial geometry and reflectance. In *ACM SIGGRAPH 2015 Talks*, page 32. ACM, 2015. 1, 2

[21] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng. Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 3

[22] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 4

[23] G. Hu, F. Yan, J. Kittler, W. Christmas, C. H. Chan, Z. Feng, and P. Huber. Efficient 3d morphable face model fitting. *Pattern Recognition*, 67:366–379, 2017. 1, 2

[24] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2, 4

[25] L. Huynh, W. Chen, S. Saito, J. Xing, K. Nagano, A. Jones, P. Debevec, and H. Li. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 3

[26] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1031–1039. IEEE, 2017. 3

[27] L. Jiang, J. Zhang, B. Deng, H. Li, and L. Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, 2018. 2

[28] I. Kemelmacher-Shlizerman and R. Basri. 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405, 2011. 1

[29] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4625–4634, 2018. 3

[30] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017. 2

[31] Y. Li, L. Ma, H. Fan, and K. Mitchell. Feature-preserving detailed 3d face reconstruction from a single image. In *Proc. of the 15th ACM SIGGRAPH European Conference on Visual Media Production*. ACM, 2018. 7, 8

[32] W.-C. Ma, T. Hawkins, P. Peers, C.-F. Chabert, M. Weiss, and P. Debevec. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 183–194. Eurographics Association, 2007. 1

[33] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004. 2

[34] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985*, 2017. 4, 7

[35] T. Papadhimitri and P. Favaro. A new perspective on uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1474–1481, 2013. 2

[36] J. Park, S. N. Sinha, Y. Matsushita, Y.-W. Tai, and I. S. Kweon. Robust multiview photometric stereo using planar mesh parameterization. *IEEE transactions on pattern analysis and machine intelligence*, 39(8):1591–1604, 2017. 2

[37] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. Ieee, 2009. 2, 3

[38] Y. Quéau. *Reconstruction tridimensionnelle par stéréophotométrie*. PhD thesis, 2015. 5

[39] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 460–469. IEEE, 2016. 1, 3

[40] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5553–5562. IEEE, 2017. 3

[41] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 986–993. IEEE, 2005. 1, 2

[42] S. Saito, T. Li, and H. Li. Real-time facial segmentation and performance capture from rgb input. In *European Conference on Computer Vision*, pages 244–261. Springer, 2016. 1, 2

[43] S. Schönborn, B. Egger, A. Morel-Forster, and T. Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017. 1, 2

[44] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1585–1594. IEEE, 2017. 3, 8

[45] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6296–6305, 2018. 3

[46] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Transactions on Graphics (TOG)*, 33(6):222, 2014. 2

[47] G. Stratou, A. Ghosh, P. Debevec, and L.-P. Morency. Effect of illumination on automatic expression recognition: a novel 3d relightable facial database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 611–618. IEEE, 2011. 2

[48] A. Tankus and N. Kiryati. Photometric stereo under perspective projection. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 611–616. IEEE, 2005. 2

[49] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018. 3

[50] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 5, 2017. 3

[51] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 1, 2

[52] A. T. Tran, T. Hassner, I. Masi, and G. Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1493–1502. IEEE, 2017. 1

[53] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[54] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proc. CVPR*, 2018. 7, 8

[55] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM transactions on graphics (TOG)*, 24(3):426–433, 2005. 2

[56] Y. Wang, L. Zhang, Z. Liu, G. Hua, Z. Wen, Z. Zhang, and D. Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2009. 7

[57] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. Reynolds. structure-from-motionphotogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012. 2

[58] C. Wu, Y. Liu, Q. Dai, and B. Wilburn. Fusing multiview and photometric stereo for 3d reconstruction under uncalibrated illumination. *IEEE transactions on visualization and computer graphics*, 17(8):1082–1095, 2011. 2

[59] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006. 2

[60] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. 7, 8

[61] X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Z. Li. Discriminative 3d morphable model fitting. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. Citeseer, 2015. 2