Learning 3D Faces from Photo-Realistic Facial Synthesis

Ruizhe Wang * †[1], Chih-Fan Chen *†[1], Hao Peng†[1], Xudong Liu †[1, 2], and Xin Li ‡[2]

[1]Oben, Inc
[2]West Virginia University

2020 International Conference on 3D Vision (3DV) | 978-1-7281-8128-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/3DV50981.2020.00096

Abstract

We present an approach to efficiently learn an accurate and complete 3D face model from a single image. Previous methods heavily rely on 3D Morphable Models to populate the facial shape space as well as an over-simplified shading model for image formulation. By contrast, our method directly augments a large set of 3D faces from a compact collection of facial scans and employs a high-quality rendering engine to synthesize the corresponding photo-realistic facial images. We first use a deep neural network to regress vertex coordinates from the given image and then refine them by a non-rigid deformation process to more accurately capture local shape similarity. We have conducted extensive experiments to demonstrate the superiority of the proposed approach on 2D-to-3D facial shape inference, especially its excellent generalization property on real-world selfie images.

1. Introduction

Acquiring high quality 3D face models is an essential task in many vision applications including virtual reality/augmented reality, teleconferencing, virtual try-on, computer games, special effect, and so on. A common practice, adopted by most professional production studios, is to *manually* create avatars from 3D scans or photo references by skillful artists. This process is often time consuming and labor intensive because each model requires days of manual processing and touching-up. It is desirable to automate the process of 3D model generation by leveraging rapid advances in computer vision/graphics and image/geometry processing.

Inferring a 3D face from a single image is challenging due to the lack of publicly available 3D face data, as well as the intrinsic ambiguity of image formulation process. Conventional wisdom attempts to address this issue by employing 3D Morphable Model (3DMM) [3] to explain the space of face shape variations, and by using a simplified shading model [1, 25] to simulate the process of image formulation. More recently, several deep learning based approaches have been proposed - either in a supervised setup to directly regress the 3DMM parameters [42, 9, 36] or in an unsupervised fashion [32, 14, 26] with the help of a differentiable rendering process. These methods are mostly limited by the expressiveness of the underlying 3DMM representation as well as the realism of the rendering process.

Figure 1: Sample outputs of our proposed method. From left to right: input image, inferred shape model that is accurate, efficient and complete, model with transferred texture

To meet those challenges, we propose two novel alternatives. First, instead of relying on the 3DMM representation to populate the face space, we directly augment a large collection of 3D faces from a small collection of facial scans by using the deformation representation feature. This process better interpolates the space of shape variations and leads to more accurate 2D-3D shape inference. Second, instead of adopting an over-simplified rendering process, we use an off-the-shelf high quality rendering engine to generate photo-realistic facial images. Our approach is capable of more accurately characterizing real-world complexities (e.g., sub-surface scattering, shadows caused by self-occlusion and skin-related reflectance [7]).

We train a deep neural network to directly regress vertex coordinates of a generic head model from the given image. To improve the generalization ability and robustness of the

*equal contribution
†{ruizhe, chihfan, hpeng, xudong}@oben.com
‡xin.li@mail.wvu.edu

2475-7888/20/$31.00 ©2020 IEEE
DOI 10.1109/3DV50981.2020.00096

trained model, we propose to first extract deep facial identity features [30, 21] which encodes each face into a unique latent representation (similar to [14]) and then decode the latent representation to a 3D face. Given the regressed 3D face model with neutral facial expression in a canonical coordinate system, we further optimize for camera intrinsic, pose, facial expression, as well as a per-vertex displacement field. Our approach is capable of better capturing local shape similarity and enabling faithful texture transfer via camera projections. Extensive experiments demonstrate the superiority of the proposed approach for 2D-to-3D facial shape inference, especially its excellent generalization property on real-world selfie images. When trained on a small number of 512 subjects, our approach can outperform the current state-of-the-art [12] trained on 10,000 real facial scans.

Our key contributions can be summarized as follows:

• A novel scheme of photo-realistic facial synthesis, by using high quality rendering on augmented shapes from a small collection of facial scans, for training facial shape inference.

• An efficient method for generating an accurate and complete 3D face model with texture from a single image by using a combination of deep neural network for shape regression and optimization for shape refinement and texture transfer.

• Extensive experimental evaluation against other benchmarks and ablation study to demonstrate the superiority of the propose method on 2D-to-3D shape inference.

2. Related Works

3D Face Representation: 3D Morphable Model (3DMM) [3] uses Principal Component Analysis (PCA) on aligned 3D neutral faces to reduce the dimension of 3D face representation making the face fitting problem more tractable. The FaceWareHouse technique [6] enhances the original PCA-based neutral face model with expressions by applying multi-linear analysis [37] to a large collection of 4D facial scans captured with RGB-D sensors. The quality of multi-linear model was further improved in [4] by jointly optimizing the model and the group-wise registration of 3D scans. In [5], a Large Scale Facial Model with 10,000 faces was generated to maximize the coverage of gender and ethnics. The training data was further enlarged in [20], which created a linear shape space trained from 4D scans of 3800 human heads. More recently, a non-linear model was proposed in [34] from a large set of unconstrained face images without the necessity of collecting 3D face scans.

Fitting via Inverse Rendering: Inverse rendering [1, 3] formulates 3D face modeling as an optimization problem over the entire parameter space seeking the best fitting for the observed image. In addition to pixel intensity values, other constraints such as facial landmarks and edge contours, are exploited for more accurate fitting [25]. More recently, GanFit [12] used a generative neural network for facial texture modeling and utilized an additional facial identity loss function in the optimization formulation. The inverse rendering based modeling approach has been widely used in many applications [40, 15, 33, 13].

Supervised Shape Regression: Convolutional Neural Network (CNN) based approaches have been proposed to directly map an input image to the parameters of a 3D face model such as 3DMM [8, 42, 18, 35, 41]. In [16], a volumetric representation was learned from an input image. In [28], an input color image was mapped to a depth image using an image translation network. In [9], a network was proposed to jointly reconstruct the 3D facial structure and provide dense alignment in the UV space. The work of [36] took a layered approach toward decoupling low-frequency geometry from its mid-level details estimated by a shape-from-shading approach. It is worth mentioning that many CNN-based approaches use facial shape estimated by inverse rendering as the ground truth during training. In [24], a 3D face model is learnt from synthetically rendered images. Unlike their facial synthesis approach, we employ photo-realistic rendering on augmented facial shapes.

Unsupervised Learning: Most recently, face modeling from images via unsupervised learning becomes popular because it affords almost unlimited amount of data for training. An image formation layer was introduced in [32] as the decoder jointly working with an auto-encoder architecture for end-to-end unsupervised training. SfSNet [29] explicitly decomposes an input image into albedo, normal and lighting components, which are then composed back to approximate the original input image. 3DMM parameters were first directly learned in [14] from facial identity encoding and then the problem of parameter optimization was formulated in an unsupervised fashion by introducing a differentiable renderer and a facial identity loss on the rendered facial image. A multi-level face model, (i.e., 3DMM with corrective field) was developed in [31] following an inverse rendering setup that explicitly models geometry, reflectance and illumination per vertex.

Deep Facial Identity Feature: Recent advances in face recognition [30, 21, 27] attempt to encode all facial images of the same subject under different conditions into identical feature representations, namely deep facial identity features. Several attempts have been made to utilize this robust feature representation for face modeling. GanFit [12] used an additional deep facial identity loss to the commonly used landmark and pixel intensity losses. In [14], 3DMM parameters were directly learned from deep facial features. Although our shape regression network is similar to theirs, the choice of training data is different. Unlike their unsupervised setting, we opt to work with supervision by synthetically rendered facial images.

3. Proposed Method

3.1. Overview

An overview of the proposed method is shown in Figure 2. To facilitate facial image synthesis (Section 3.2) for training a shape regression neural network (Section 3.3), we have collected and processed a prioritized 3D face dataset, from which we can sample augmented 3D face shape with UV-texture to render a large collection of photo-realistic facial images. During testing, the input image is first used to directly regress the 3D vertex coordinates of a 3D face model with the given topology, which are furthered refined to fit the input image with a per-vertex non-rigid deformation approach (Section 3.4.1). Upon accurate fitting, selfie texture is projected to the UV space to infer a complete texture map (Section 3.4.2).

3.2. Photo-Realistic Facial Synthesis

3.2.1 3D Scan Database

. The most widely used Basel Face Model (BFM) [22] has two major drawbacks. First, it consists of 200 subjects but mainly Caucasian, which might lead to biased face shape estimation. Second, each face is represented by a dense model with high polygon count, per-vertex texture appearance and frontal face only, which limits its use for production-level real-time rendering. To overcome these limitations, we have collected a total of 512 subjects using a professional-grade multi-camera stereo scanner (3dMD LLC, Atlanta [1]) across different gender and ethnicity as shown in Table 1.

Sex/Ethnicity	White	Asian	Black	Total
Male	82 / 5	178 / 5	8 / 5	268 / 15
Female	45 / 5	164 / 5	5 / 5	214 / 15
Total	127 / 10	342 / 10	13 / 10	482 / 30

Table 1: The distribution of gender and ethnicity in our database. Note that we randomly select 5 subjects for each group for testing and the rest subjects are used for training and validation.

As shown in Figure 3, we process a raw textured 3D facial scan data to generate our 3D face representation that consists of a shape model with low polygon count and a high-resolution diffuse map for preserving details. A face representation containing a head model of 2925 vertices and a diffuse map sized by 2048×2048 is used. We take a non-rigid alignment approach [6] of deforming a generic head model to match the captured facial scan. Then we transfer the texture onto the generic model's UV space. With further manual artistic touch up, we obtain the final high-fidelity diffuse map.

[1] http://www.3dmd.com/

3.2.2 Data Augmentation

482 subjects are far from enough to cover all possible facial shape variations. While it is expensive to collect thousands of high-quality facial scans, we adopt an alternative shape augmentation approach to improve the generalization ability of the trained neural network. First, we adopt a recent deformation representation (DR) [38, 11] to model a 3D facial mesh $\mathbf{P}$. DR feature encodes the $i\text{-}th$ vertex $\mathbf{P}^i = [P^i_x, P^i_y, P^i_z]$ as a $\mathbb{R}^9$ vector. Hence the DR feature of the entire mesh is represented as a vector $\mathbf{D} \in \mathbb{R}^{|\mathbf{P}|\times 9}$. $\mathbf{D}$ encodes local deformation around each vertex of $\mathbf{P}$ with respect to a reference mesh $\mathbf{P}^R$ into a $\mathbb{R}^9$ vector. We use the mean face of all 482 processed facial models as the reference mesh.

Encode $\mathbf{D}$ *from* $\mathbf{P}$ We denote the $i\text{-}th$ vertex as $\mathbf{p}_i$ and $\mathbf{p}^R_i$ respectively. The deformation gradient in the closest neighborhood $\mathcal{N}_i$ of the $i\text{-}th$ vertex from the reference model to the deformed model is defined by the affine transformation matrix $\mathbf{T}_i$ that minimizes the following energy

$$E(\mathbf{T}_i) = \sum_{i\in\mathcal{N}_i} c_{ij} \|(\mathbf{p}_i - \mathbf{p}_j) - \mathbf{T}_i(\mathbf{p}^R_i - \mathbf{p}^R_j)\|^2, \quad (1)$$

where c_{ij} is the cotangent weight depending on the reference model to handle irregular tessellation. With polar decomposition, $\mathbf{T}_i$ is decomposed into a rotation component $\mathbf{R}_i$ and a scaling/shear component $\mathbf{S}_i$ such that $\mathbf{T}_i = \mathbf{R}_i\mathbf{S}_i$. The rotation matrix can be represented with a rotation axis ω_i and rotation angle θ_i pair, and we further convert them to the matrix logarithm representation:

$$log\mathbf{R}_i = \theta_i \begin{pmatrix} 0 & -\omega_{i,z} & \omega_{i,y} \\ \omega_{i,z} & 0 & -\omega_{i,x} \\ -\omega_{i,y} & \omega_{i,x} & 0 \end{pmatrix}. \quad (2)$$

Finally the DR feature for $\mathbf{p}_i$ is represented by $\mathbf{d}_i = \{log\mathbf{R}_i; \mathbf{S}_i - \mathbf{I}\}$ where $\mathbf{I}$ is the identity matrix. Since $\|\omega_i\| = 1$ and $\mathbf{S}_i$ is symmetric, $\mathbf{d}_i$ has 9 DoF.

Recover $\mathbf{P}$ *from* $\mathbf{D}$ Given the DR feature $\mathbf{D}$ and the reference mesh $\mathbf{P}^R$, we first recover the affine transformation $\mathbf{T}_i$ for each vertex. Then we try to recover the optimal $\mathbf{P}$ that minimizes:

$$E(\mathbf{P}) = \sum_{\mathbf{p}_i\in\mathbf{P}} \sum_{j\in\mathcal{N}_i} c_{ij} \|(\mathbf{p}_i - \mathbf{p}_j) - \mathbf{T}_i(\mathbf{p}^R_i - \mathbf{p}^R_j)\|^2. \quad (3)$$

For each $\mathbf{p}_i$, we obtain it by solving $\frac{\partial E(\mathbf{P})}{\partial \mathbf{p}_i} = 0$ which gives

$$2\sum_{j\in\mathcal{N}_i} c_{ij}(\mathbf{p}_i - \mathbf{p}_j) = \sum_{j\in\mathcal{N}_i} \mathbf{T}_i(\mathbf{p}^R_i - \mathbf{p}^R_j). \quad (4)$$

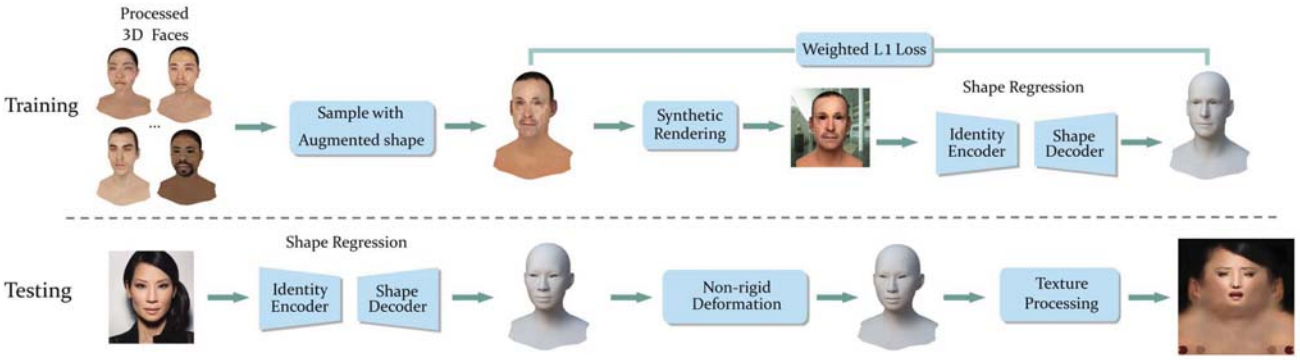


Figure 2: Overview of the proposed approach. During training, we learn a shape regression neural network on photo-realistic synthetic facial images. During testing, we infer a low polygon count shape model with a UV diffuse map generated from the projected texture.

Figure 3: (left-most) a raw facial scan with dense topology, (left) the model with UV texture; (right) the processed face model with sparse topology, and (right-most) the model with UV texture.

The resulting equations for all $\mathbf{p}_i \in \mathbf{P}$ lead to a linear system which can be written as $\mathbf{AP} = \mathbf{b}$. By specifying the position of one vertex, we can get the single solution to the equation to fully recover $\mathbf{P}$.

Upon obtaining a set of DR features as $(\mathbf{D}_1, ..., \mathbf{D}_N)$ where N is the total number of subjects, we follow [17] to sample new DR features. More specifically, we sample a vector $(r, \theta_1, ..., \theta_{m_1})$ in Polar coordinates, where r observes a uniform distribution $\mathbf{U}[0.6, 1.3]$ and θ_i follows uniform distribution $\mathbf{U}[0, \pi/2]$. We calculate its corresponding Cartesian coordinates $(a_1, a_2, ..., a_m)$ and interpolate the sampled DR features as $\sum_{i=1}^{m} a_i \mathbf{D}_i$, from which we further calculate the corresponding facial mesh. Note that the sampling in this paper is under a different scenario than [17], which inferring 3D faces from 2D images. In our experiments, we use $m = 5$ and only select samples from the same gender and ethnicity for blending. We generate 10,000 new 3D faces with a ratio of $0.65/0.30/0.05$ across Asian/Caucasian/Black and a ratio of $0.5/0.5$ across Male/Female.

Synthetic Rendering For each new sampled face, we assign its UV texture by choosing the closest 3D face in the same ethnicity and gender from existing 482 subjects. We use an off-the-shelf high quality rendering engine V-ray [2]. With artistic assistance, we set up a shader graph to render photo-realistic facial images given a custom diffuse map and a generic specular map. We manually set up 30 different lighting conditions and further randomize head rotation $[-15^\circ, +15^\circ]$ in roll, yaw and pitch. The background of rendered images are randomized with a large collection of indoor and outdoor images. We opt not to render eye models and mask out the eye areas when testing by using detected local eye landmarks. Please see the supplementary materials for more details.

3.3. Regressing Vertex Coordinates

Our shape regression network consists of a feature encoder and a shape decoder. Deep facial identity feature is known for its robustness under varying conditions such as lighting, head pose and facial expression, providing a naturally ideal option for the encoded feature. Although any off-the-shelf facial recognition network would be sufficient for our task, we propose to adopt Light CNN-29V2 [39] due to its good balance between network size and encoding efficiency. A pre-trained Light CNN-29V2 model is used to encode an input image into a 256-dimensional feature vector. We have used a weighted per-vertex L1 loss: weight of 5 for vertices on the facial area (within a radius of 95mm from the nose tip) and weight of 1 for other vertices.

For shape decoder, we have used three fully connected (FC) layers, with the output size of 128, 200 and 8,775 respectively. The last FC layer directly predicts concatenated vertex coordinates of a generic head model consisting of 2,925 points, and it is initialized with 200 pre-computed PCA components explaining more than 99% of the variance observed in the 10,000 augmented 3D facial shapes.

[2] https://vray.us/

3.4. Refinement and Texture Transfer

3.4.1 Non-rigid Deformation

3D vertex coordinates generated by the shape regression neural network is not directly applicable to texture projection because facial images usually contain unknown factors such as camera intrinsic, head pose and facial expression. Meanwhile, since shape regression predicts the overall facial shape, local parts such as eyes, nose and mouth are not accurately reconstructed; but they are equally important to quality perception when comparing against the original face image. We propose to utilize facial landmarks detected in a coarse-to-fine fashion and formulate non-rigid deformation as an optimization problem that jointly optimizes over camera intrinsic, camera extrinsic, facial expression and a per-vertex displacement field. Please see the supplementary material for more details.

3.4.2 Texture Processing

Upon non-rigid deformation, we project selfie texture to the UV space of the generic model using the estimated camera intrinsic, head pose, facial expression and per-vertex correction. While usually only the frontal area on a selfie is visible, we recover textures on other areas, e.g., back of head and neck, by using the UV texture of one of the 482 subjects that is *closest* to the query subject. We define closeness as L1 loss on the distance between LightCNN-29V2 embeddings, (i.e, through face recognition). Finally given a foreground projected texture and a background default texture, we blend them using the Poisson Image Editing [23].

4. Experimental Results

4.1. Implementation Details

For shape regression, we use Adam optimizer with a learning rate of 0.0001 and the momentum $\beta_1 = 0.5$, $\beta_2 = 0.999$ for 500 epochs. We train on a total of 10,000 synthetically rendered facial images with a batch size of 64.

4.2. Database and Evaluation Setup

Stirling/ESRC 3D Faces Database: The ESRC [10] is the latest public 3D faces database captured by a Di3D camera system. The database also provides several images captured from different viewpoints under various lighting condition. We select those subjects who have both 3D scan and a frontal neutral face for evaluation. There are total 129 subjects (62 male and 67 female) for testing. Note that in this dataset, around 95% of people are Caucasian.

JNU-Validation Database: The JNU-Validation Database is a part of the JNU 3D face Database collected by the Jiangnan University [19]. It has 161 2D images of 10 Asians and their 3D face scans captured by 3dMD. Since the validation database was not used during training, we consider it as a test database for Asians. The 2D images of each subject are in range of $[3, 26]$. To minimize the impact of imbalance data, we select three frontal images of each subject for quantitative comparison.

Our Test Data Since there is no public database available for testing, which shall cover all the gender and races, we randomly pick five subjects from the six group in Table 1 and form a total 30 subjects as the evaluation database. The other 482 scans are used as for data augmentation and training/validation stage for both geometry and texture. Each subject has two testing images: a selfie captured by a Samsung Galaxy S7 and an image captured from a Sony a7R DSLR camera by a photographer.

Evaluation Setup: We compared our method with several state-of-the-art-methods including 3DMM-CNN [35], Extreme 3D Face (E3D) [36], PRNet [9], RingNet [26], and GanFit [12]. The reconstructed model detail of each methods are shown in Table 2. Note that for our method and RingNet, both eyes, teeth and tongue and their model holders are removed before comparison. Because the evaluation metric is using the point-to-plane error, unrelated data will increase the over all error. Although removing those parts will also slightly increase the error (e.g., no data in the eyes area to compare), the introduced error is much smaller than the error of directly using the original models. For a fair comparison with all other methods, unless clearly stated, evaluation numbers of our method is *without the estimated per-vertex displacement field*.

4.3. Quantitative Comparison

Evaluation Metric: To align the reconstructed model with ground truth, we followed the step of [35, 14, 12] and the challenge [10]. Since the topology of each method is fixed, seven pre-selected vertex index is first used to roughly align the reconstructed model to the ground truth and then the model was further refined by iterative closest point (ICP) [2]. The position of vertex of the tip of the nose v_t is chosen to be the center of the ground truth and reconstructed models. Given a threshold d mm, we discard those vertex v_i, where $||v_i - v_t|| > d$. To evaluate the reconstructed model with ground truth, we used the Average Root Mean Square Error (ARMSE) [3] as suggested by the 2_{nd} 3DFAW Challenge [4] , where it computes the closest point-to-mesh distance between the ground truth and predicted model and vice versa.

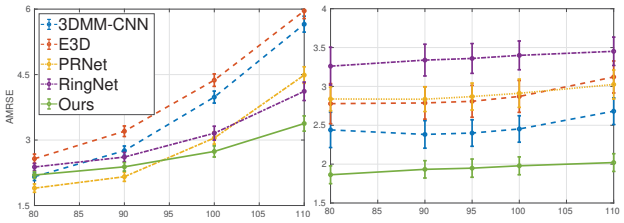
ESRC and JNU-validation Dataset: In Figure 4, we have chosen d = [80, 90, 100, 110] and computed the ARMSE for each reconstructed model and ground truth. Note that the annotation provided by ESRC database only

[3] https://codalab.lri.fr/competitions/572#learn_the_details-evaluation

[4] https://3dfaw.github.io/

	Ours	RingNet [26]	GanFit [12]	PRNet [9]	E3D [36]	3DMM-CNN [35]
Full Head	Yes	Yes	No	No	No	No
Vertex	2.9K (2.7K)	5.0K (3.8K)	53.2K	43.7K	∼155K	47.0K
Face	5.8K (5.3K)	10.0 K (7.4K)	105.8K	86.9K	∼150K	93.3K

Table 2: The geometric complexity of our method and other method. Note that except E3D, the other methods used the same topology for their reconstructed model. The number inside the parentheses in both our method and RingNet are the details of head models after unrelated mesh removal.



(a) ESRC (b) JNU-Validation

Figure 4: The quantitative results of our method compare to 3DMM-CNN [35], E3D [36], PRNet [9] and RingNet [26] on both ESRC and JNU-Validation database.

has the seven landmark for alignment, thus instead of using the tip of nose, we use the average of the 7 landmark as the center of face. In ESRC, our result is better than other methods when $d > 95$ and our performance is more resilient as d increases. This indicates that our method can better replicate the shape of the entire head than other methods. In JNU-validation database, since other methods are trained from a Caucasian-dominated 3DMM model, while the other races are also considered during our augmented stage, we can achieve much smaller reconstructed error at every d value.

Our Test Dataset: In Figure 5 (a), the centers of each error-bar are the average of the ARMSE from the 60 reconstructed meshes. The range of the errorbar is $\pm 1.96 \times SE$, where SE is the standard error. It is shown that our reconstructed models is slightly better than GanFit and significantly better than other methods. It is worth mentioning that our vertex number is only $\sim 70\%$ of RingNet and less than 6% of other methods. In Figure 5 (b), the cropped mesh of the ground truth and each methods are shown under different threshold of d. To utilize the reconstructed models for real-world application, we believe that $d = 110$ is the best value because it captured the entire head instead of the frontal face. We further investigate the performance under different races and the results are shown in Figure 5 (c). Our method can correctly replicate the model to under 2.5 mm of error in all ethnicity, while other methods such as RingNet and PRNet are very sensitive to the ethnicity differences. Although GanFit performed slightly better than our method on White and Black races, the overall performance is not as good as ours because they are not able to recover the Asian geometries well. It is worth noting that we used 10000 synthetic images augmented from less than 500 scan data, which is only 5% of the data used in GanFit. To fairly visualize the error between methods without the effect of different topology, we find the closest point-to-plane distance from ground truth to reconstructed model and generate the heat-map for each method in Figure 6.

4.4. Ablation Study

To demonstrate the effectiveness of the individual modules in the proposed approach, we modify one variable at a time and compare with the following alternatives:

• **No Augmentation (No-Aug):** Without any augmentation, we simply repetitively sample 10,000 faces from 482 subjects.

• **3DMM Augmentation with class (3DMM-C):** Instead of DR Feature based sampling, we propose a 3DMM based shape augmentation method considering race and gender. We train a 3DMM representation from 482 subjects, and for each group in Table 1, a Gaussian random vector $\mathbf{x} \sim \mathcal{N}(\mu_i, \Sigma_i^2)$ is used to create weights of the principal shape components, where μ_i and Σ_i^2 are the mean vector and co-variance matrix of those coefficients in the group. We sample 10,000 faces with this augmentation approach.

• **Game engine Rendering (Unity):** Instead of using high-quality photo-realistic renderer, we use Unity, a standard game rendering engine, to synthesize facial images. The quality of rendered images are comparatively lower than V-ray. We keep the DR feature based augmentation approach and rendered exactly the same 10000 synthetic faces mentioned in Section 3.2.

In Figure 7, our proposed approach outperforms all other alternatives. It is expected that without data augmentation (i.e., No-Aug), the reconstructed error is the worst among all methods. The difference between 3DMM-C and our method demonstrates that DR based augmentation scheme better interpolates facial shape space comparing with the traditional 3DMM representation. The results between Unity and our method shows that rendering quality plays an important role in bridging the gap between real and synthetic images. The proposed non-rigid deformation method (Ours+ND) not only plays a critical role in texture transfer, but also helps to improve local shape similarity.

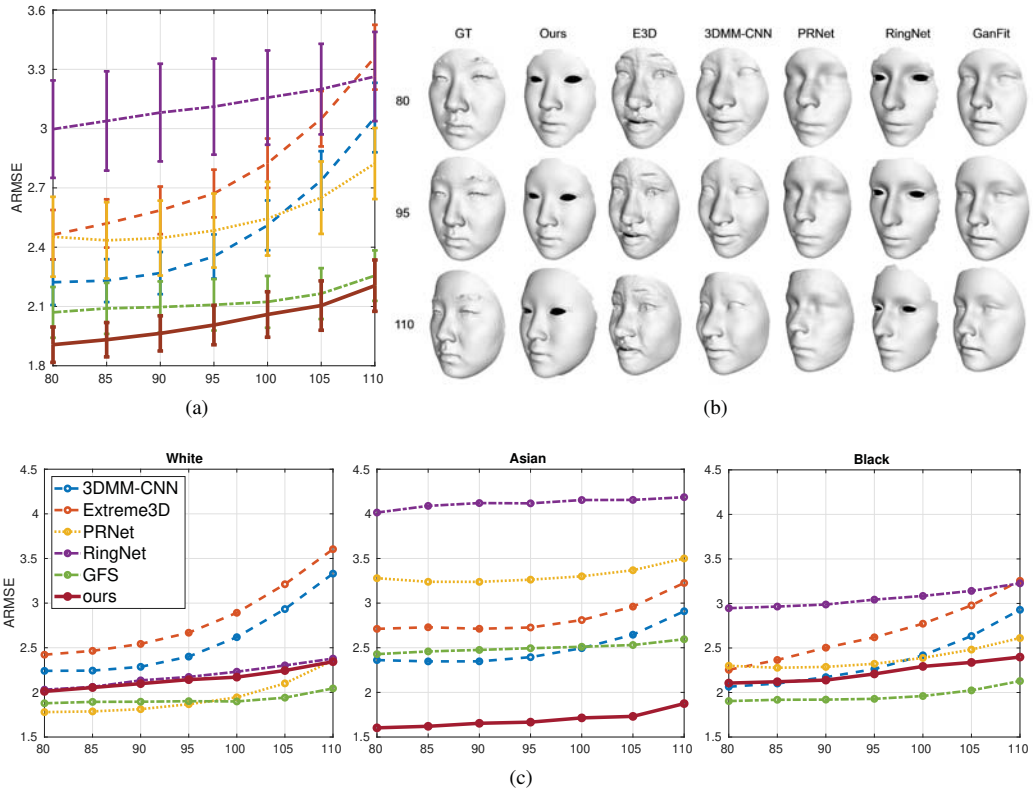


Figure 5: The quantitative results of our method compare to E3D [36], 3DMM-CNN [35], PRnet [9], RingNet [26] and GanFit [12]. (a) The overall performance of each method. (b) The qualitative comparison of cropped meshes with ground truth in $d = [80, 95, 110]$. (c) The evaluation results on different ethnicity.

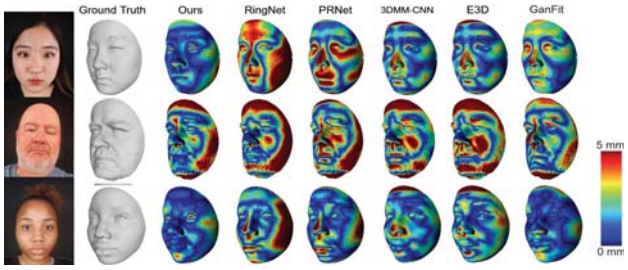


Figure 6: The heatmap visualization of reconstructed models in $d = 110$. Vertices colored red fall above the 5mm error tolerance, while blue vertices are those which lie within the tolerance.

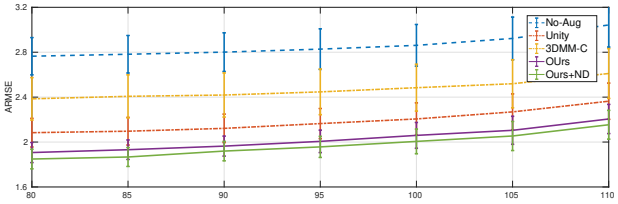


Figure 7: The quantitative results of No-Aug, 3DMM-C, Unity, our method and our method with non-rigid deformation. The proposed method achieve the best performance at all time.

4.4.1 Qualitative Comparison

Figure 8 shows our shape estimation method on frontal face images side-by-side with the state-of-the-arts in MoFA test database. We picked the same images shown in GanFit [12]. Our method creates accurate face geometry, while also capturing discriminate features which allow the identity of each face to be easily distinguishable from the others.

Meanwhile, as shown in Table 2, our result maintains a low geometric complexity. This allows our avatars to be production ready even in demanding cases such as on mobile platforms. In Figure 9, we choose a few celebrity to verify the geometry accuracy of our method comparing to others. In Figure 10, we present the results of several celebrities and compare our method not only for geometry but also in appearance. Note that by projecting the selfie to a high-resolution UV texture, our reconstructed models has photo-

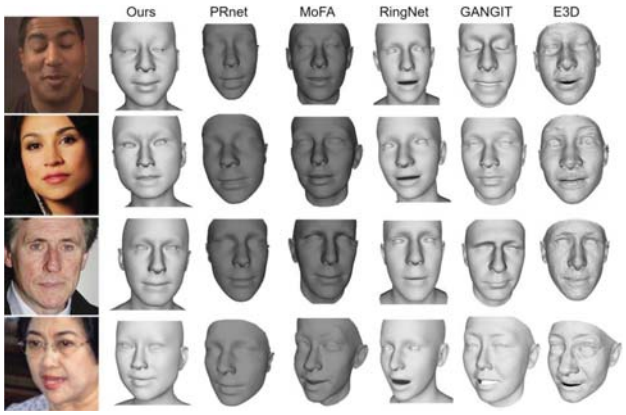


Figure 8: The qualitative comparison of our method with PRNet [9], MoFA [32], RingNet [26], GanFit [12] and E3D [36]. Our method accurately reconstructs the geometry, while maintaining a much lower vertices count, which is more suitable for production.

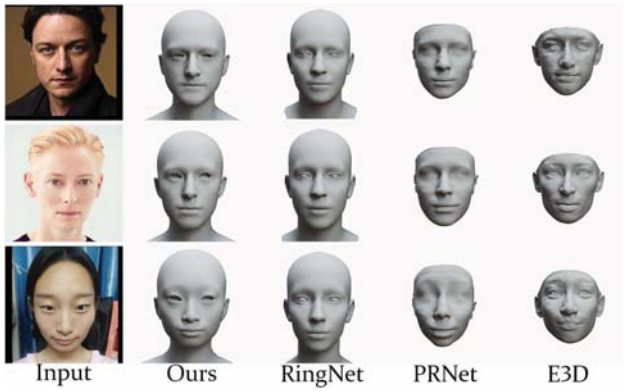


Figure 9: The showcase of our reconstruction results of several celebrities comparing to RingNet [26], PRNet [9] and E3D [36].

realistic appearance while 3DMM-CNN [35] and PRNet [9] used vertex color results in limited texture reapplication. In Figure 11, we demonstrate our final results with blended diffuse maps in Section 3.4.2.

4.5. Discussion of Data Distribution

The qualitative and quantitative comparison concludes that the geometry of human faces are *gender-* and *race-dependent*, thus representing human faces with a single 3DMM model is problematic. Furthermore, in our ablation study further shows that Gaussian distribution assumption might not be ideal for each category. In contrast to other method, our proposed framework has the ability to extend to more races such as South Asian or Latino, or any categories (e.g., ages) that can better describe a group of people with similar facial geometry.

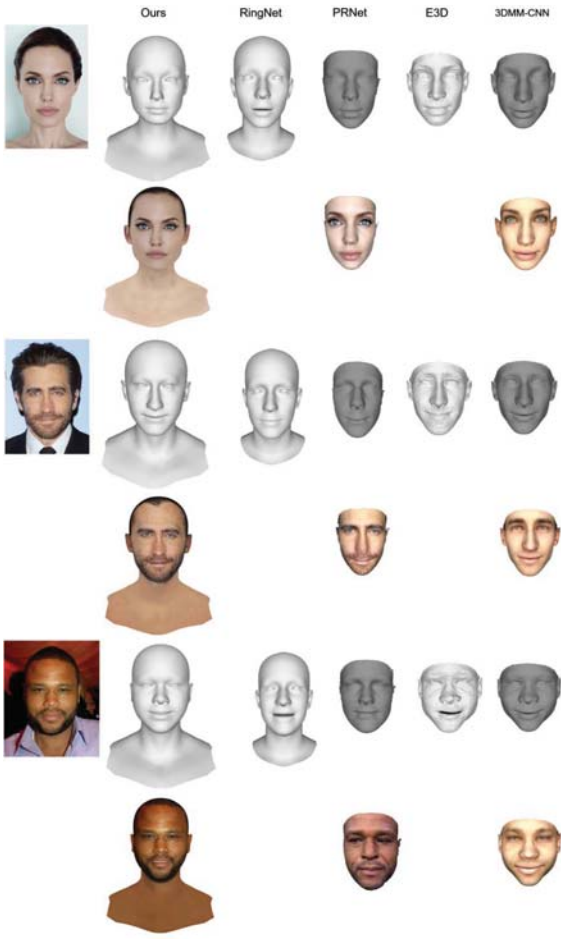


Figure 10: Qualitative results of our method compare to RingNet [26], PRnet [9], E3D [36], and 3DMM-CNN [35].

Figure 11: Our final results with blended diffuse maps.

5. Conclusions

In this paper, we propose a novel scheme of photo-realistic facial synthesis, by using high quality rendering on augmented shapes from a small collection of facial scans, for training facial shape inference from a single image. We further adopt an optimization based approach for capturing higher local similarity and enabling texture transfer. Extensive experimental evaluation against state-of-the-art methods demonstrates effectiveness of the propose method on facial shape inference.

References

[1] O. Aldrian and W. A. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012. 1, 2

[2] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 5

[3] 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*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1, 2

[4] T. Bolkart and S. Wuhrer. A groupwise multilinear correspondence optimization for 3d faces. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3604–3612, Dec 2015. 2

[5] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahy, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5543–5552, June 2016. 2

[6] 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, March 2014. 2, 3

[7] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156. ACM Press/Addison-Wesley Publishing Co., 2000. 1

[8] P. Dou, S. K. Shah, and I. A. Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5908–5917, 2017. 2

[9] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 1, 2, 5, 6, 7, 8

[10] Z. Feng, P. Huber, J. Kittler, P. Hancock, X. Wu, Q. Zhao, P. Koppen, and M. Raetsch. Evaluation of dense 3d reconstruction from 2d face images in the wild. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 780–786, May 2018. 5

[11] L. Gao, Y.-K. Lai, J. Yang, Z. Ling-Xiao, S. Xia, and L. Kobbelt. Sparse data driven mesh deformation. *IEEE transactions on visualization and computer graphics*, 2019. 3

[12] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 2, 5, 6, 7, 8

[13] Z. Geng, C. Cao, and S. Tulyakov. 3d guided fine-grained face manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9821–9830, 2019. 2

[14] 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. 1, 2, 5

[15] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering, nov 2017. 2

[16] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 1031–1039, 2017. 2

[17] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang. Disentangled representation learning for 3d face shape. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[18] A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016. 2

[19] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin. Gaussian mixture 3d morphable face model. *Pattern Recogn.*, 74(C):617–628, Feb 2018. 5

[20] 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, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2

[21] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In X. Xie, M. W. Jones, and G. K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015. 2

[22] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, Sep. 2009. 3

[23] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003. 5

[24] E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 fourth international conference on 3D vision (3DV)*, pages 460–469. IEEE, 2016. 2

[25] S. Romdhani and T. Vetter. 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)*, volume 2, pages 986–993. IEEE, 2005. 1, 2

[26] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision, 2019. 1, 5, 6, 7, 8

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[28] M. Sela, E. Richardson, and R. Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. *arxiv*, 2017. 2

[29] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 2

[30] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 2

[31] 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. 2

[32] 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. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3735–3744, 2017. 1, 2, 8

[33] 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. 2

[34] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 2018. 2

[35] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. 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*, pages 5163–5172, 2017. 2, 5, 6, 7, 8

[36] A. Tuan Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 1, 2, 5, 6, 7, 8

[37] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, jul 2005. 2

[38] Q. Wu, J. Zhang, Y.-K. Lai, J. Zheng, and J. Cai. Alive caricature from 2d to 3d. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7336–7345, 2018. 3

[39] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 4

[40] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):162, 2018. 2

[41] H. Yi, C. Li, Q. Cao, X. Shen, S. Li, G. Wang, and Y.-W. Tai. Mmface: A multi-metric regression network for unconstrained face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7663–7672, 2019. 2

[42] 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. 1, 2