FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction

Haotian Yang*1, Hao Zhu*1,2,5, Yanru Wang1, Mingkai Huang1, Qiu Shen1, Ruigang Yang2,3,4,5, Xun Cao1
1Nanjing University, 2Baidu Research, 3University of Kentucky, 4Inceptio Inc., 5National Engineering Laboratory for Deep Learning Technology and Applications, China

Abstract

In this paper, we present a large-scale detailed 3D face dataset, FaceScape, and propose a novel algorithm that is able to predict elaborate riggable 3D face models from a single image input. FaceScape dataset provides 18,760 textured 3D faces, captured from 938 subjects and each with 20 specific expressions. The 3D models contain the pore-level facial geometry that is also processed to be topologically uniformed. These fine 3D facial models can be represented as a 3D morphable model for rough shapes and displacement maps for detailed geometry. Taking advantage of the large-scale and high-accuracy dataset, a novel algorithm is further proposed to learn the expression-specific dynamic details using a deep neural network. The learned relationship serves as the foundation of our 3D face prediction system from a single image input. Different than the previous methods, our predicted 3D models are riggable with highly detailed geometry under different expressions. The unprecedented dataset and code will be released to public for research purpose†.

1. Introduction

Parsing and recovering 3D face models from images have been a hot research topic in both computer vision and computer graphics due to its many applications. As learning based methods have become the mainstream in face tracking, recognition, reconstruction and synthesis, 3D face datasets becomes increasingly important. While there are numerous 2D face datasets, the few 3D datasets lack in 3D details and scale. As such, learning-based methods that rely on the 3D information suffer.

Existing 3D face datasets capture the face geometry using sparse camera array[8, 19, 34] or active depth sensor such as Kinect[12] and coded light[33]. These setups limit the quality of the recovered faces. We captured the 3D face model using a dense 68-camera array under controlled illumination, which recovers the 3D face model with wrinkle and pore level detailed shapes, as shown in Figure 1. In addition to shape quality, our dataset provides considerable amount of scans for study. We invited 938 people between the ages of 16 and 70 as subjects, and each subject is guided to perform 20 specified expressions, generating 18,760 high quality 3D face models. The corresponding color images and subjects’ basic information (such as age and gender)
are also recorded.

Based on the high-fidelity raw data, we build a powerful parametric model to represent the detailed face shape. All the raw scans are firstly transformed to a topologically unified base model representing the rough shape and a displacement map representing detailed shape. The transformed models are further used to build bilinear models in identity and expression dimension. Experiments show that our generated bilinear model exceeds previous methods in representative ability.

Using FaceScape dataset, we study how to predict a detailed riggable face model from a single image. Prior methods are able to estimate rough blendshapes where no wrinkle and subtle features are recovered. The main problem is how to predict the variation of small-scale geometry caused by expression changing, such as wrinkles. We propose the dynamic details which can be predicted from a single image by training a deep neural network on FaceScape dataset. Cooperated with bilinear model fitting method, a full system to predict detailed riggable model is presented. Our system consists of three stages: base model fitting, displacement map prediction and dynamic details synthesis. As shown in Figure 1, our method predicts detailed 3D face model which contains subtle geometry, and achieves high accuracy due to the powerful bilinear model generated from FaceScape dataset. The predicted model can be rigged to various expressions with plausible detailed geometry.

Our contributions are summarized as following:

- We present a large-scale 3D face dataset, FaceScape, consisting of 18,760 extremely detailed 3D face models. All the models are processed to topologically unified base models for rough shape and displacement maps for detailed shape. The data are released free for non-commercial research.

- We model the variation of detailed geometry across expressions as dynamic details, and propose to learn the dynamic detail from FaceScape using a deep neural network.

- A full pipeline is presented to predict detailed riggable face model from a single image. Our result model can be rigged to various expressions with plausible geometric details.

2. Related Work

3D Face Dataset. 3D face datasets are of great value in face-related research areas. Existing 3D face datasets could be categorized according to the acquisition of 3D face model. Model fitting datasets[33, 60, 23, 5, 7] fit the 3D morphable model to the collected images, which makes it convenient to build a large-scale dataset on the base of wild faces. The major problem of the fitted 3D model is the uncertainty of accuracy and the lack of detailed shape. To obtain the accurate 3D face shape, a number of works reconstructed the 3D face using active method including depth sensor or scanner[53, 52, 3, 38, 37, 12, 17], while the other works built sparse multi-view camera system[54, 18]. Traditional depth sensors and 3D scanners suffer from the limited spatial resolution, so they can’t recover detailed facial geometry. The sparse multi-view camera system suffers from the unstable and inaccurate reconstruction[39, 56, 55]. The drawbacks of these methods limit the quality of 3D face model in previous datasets. Different from the datasets above, FaceScape obtained the 3D face model from a dense multi-view system with 68 DSLR cameras, which provides extremely high quality face models. The parameters measuring 3D model quality are listed in Table 1. Our dataset outperforms previous works on both model quality and data amount. Note that Table 1 doesn’t list the datasets which provide only parametric model but no source 3D models[8, 6, 33, 30].

3D Morphable Model. 3DMM is a statistical model which transforms the shape and texture of the faces into a vector space representation[4]. As 3DMM inherently contains the explicit correspondences from model to model, it is widely used in model fitting, face synthesis, image manipulations, etc. The recent research on 3DMM can be generally divided into two directions. The first direction is to separate the parametric space to multiple dimensions like identity, expression and visemes, so that the model could be controlled by these attributes separately[49, 12, 29, 26]. The models in expression dimension could be further transformed to a set of blendshapes[27], which can be rigged to generate individual-specific animation. Another direction is to enhance the representation power of 3DMM by using deep neural network to present 3DMM bases[2, 42, 45, 47, 46, 16].

Single-view shape Prediction. Predicting 3D shape from a single image is a key problem to many applications like view synthesis[22, 57, 58] and stereoscopic video generation[13, 24]. The emergence of 3DMM has simplified the single-view reconstruction of face to a model fitting problem, which could be well solved by fitting facial landmarks and other features[36, 43] or regressing the parameter of 3DMM with a deep neural network[20, 60]. However, fitting 3DMM is difficult in recovering small details from the input image due to the limited representation power. To solve this problem, several recent works adopt the multi-layer refinement structures. Richardson et al.[35] and Sela et al.[40] both proposed to firstly predict a rough facial shape and render it to the depth map, then refine the depth map to enhance the details from the registered source image. Sengupta et al.[41] proposed to train the SfSNet on combination of labeled synthetic data and unlabeled in-the-wild data to estimate plausible detailed shape in un-
Table 1: Comparison of 3D Face Datasets

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BU-3DFE[53]</td>
<td>100</td>
<td>25</td>
<td>10k-20k</td>
<td>1300 × 900 / -</td>
<td>structure light</td>
</tr>
<tr>
<td>BU-4DFE[52]</td>
<td>101</td>
<td>6(video)</td>
<td>10k-20k</td>
<td>1040 × 1329 / -</td>
<td>structure light</td>
</tr>
<tr>
<td>BJUT-3D[3]</td>
<td>500</td>
<td>1-3</td>
<td>≈200k</td>
<td>478 × 489 / -</td>
<td>laser scanner</td>
</tr>
<tr>
<td>Bosphorus[38]</td>
<td>105</td>
<td>35</td>
<td>≈35k</td>
<td>1600 × 1200 / -</td>
<td>structure light</td>
</tr>
<tr>
<td>FaceWarehouse[12]</td>
<td>150</td>
<td>20</td>
<td>≈11k</td>
<td>640 × 480 / -</td>
<td>kinect</td>
</tr>
<tr>
<td>4DFAB[17]</td>
<td>180</td>
<td>6(video)</td>
<td>≈100k</td>
<td>1200 × 1600 / -</td>
<td>kinect+cameras(7)</td>
</tr>
<tr>
<td>D3DFACS[18]</td>
<td>10</td>
<td>38AU(video)</td>
<td>≈30k</td>
<td>- / 1024 × 1280</td>
<td>multi-view system(6)</td>
</tr>
<tr>
<td>BP4D-Spontaneous[54]</td>
<td>41</td>
<td>27AU(video)</td>
<td>≈37k</td>
<td>1040 × 1392 / -</td>
<td>multi-view system(3)</td>
</tr>
<tr>
<td>FaceScape (Ours)</td>
<td>938</td>
<td>20</td>
<td>≈2m</td>
<td>4k-8k / 4096×4096</td>
<td>multi-view system(68)</td>
</tr>
</tbody>
</table>

constrained images. Tran et al. [44] proposed to predict a bump map to represent the wrinkle-level geometry base on a rough base model. Huynh et al. [25] utilized image-to-image network and super-resolution network to recover the mesoscopic facial geometry in the form of displacement map. Chen et al. [15] also tried to predict the displacement map with a conditional GAN based on the 3DMM model, which enables to recover detailed shape from an in-the-wild image.

Our work advances the state of the art in multiple aspects. In dataset, our FaceScape is by far the largest with the highest quality. A detailed quantitative comparison with previous datasets are made in Table 1. In 3D face prediction, previous works focus on enhancing the static detailed facial shape, while we study the problem of recovering an animable model from a single image. We demonstrate for the first time that a detailed and rigged 3D face model can be recovered from a single image. The rigged model exhibits expression-dependent geometric details such as wrinkles.

3. Dataset

3.1. 3D Face Capture

We use a multi-view 3D reconstruction system to capture the raw mesh model for the datasets. The multi-view system consists of 68 DSLR cameras, 30 of which capture 8K images focusing on front side, and the other cameras capture 4K level images for the side part. The camera shutters are synced to be triggered within 5ms. We spend six months to invite 938 people to be our capturing subjects. The subjects are between 16 and 70 years old, and are mostly from Asia. We follow FaceWarehouse[12] which asks each subject to perform 20 specific expressions including neutral expression for capturing. The total reconstructed number reach to roughly 18,760, which is the largest amount comparing to previous expression controlled 3D face datasets. The reconstructed model is triangle mesh with roughly 2 million vertices and 4 million triangle faces. The meta information for each subject is recorded, including age, gender, and job (by voluntary). We show the statistical information about the subjects in our dataset in Figure 2, and a comparison with prior 3D face datasets in Table 1.

3.2. Topologically Uniformed Model

We down-sample the raw recovered mesh into rough mesh with less triangle faces, namely base shape, and then build 3DMM for these simplified meshes. Firstly, we roughly register all the meshes to the template face model...
by aligning 3D facial landmarks, then the NICP\cite{1} is used to deform the templates to fit the scanned meshes. The deformed meshes can be used to represent the original scanned face with minor accuracy loss, and more importantly, all of the deformed models share the uniform topology. The detailed steps to register all the raw meshes are described in the supplementary material.

After obtaining the topology-uniformed base shape, we use displacement maps in UV space to represent middle and fine scale details that are not captured by the base model due to the small number of vertices and faces. We find the surface points of base mesh corresponding to the pixels in the displacement map, then inverse-project the points to the raw mesh along normal direction to find its corresponding points. The pixel values of the displacement map is set to the signed distance from the point on base mesh to its corresponding point.

We use base shapes to represent rough geometry and displacement maps to represent detailed geometry, which is a two-layer representation for our extremely detailed face shape. The new representation takes roughly 2\% of the original mesh data size, while maintaining the mean absolute error to be less than 0.3mm.

3.3. Bilinear Model

Bilinear model is firstly proposed by Vlasic et al.\cite{49}, which is a special form of 3D morphable model to parameterize face models in both identity and expression dimensions. The bilinear model can be linked to a face-fitting algorithm to extract identity, and the fitted individual-specific model can be further transformed to riggable blendshapes. Here we describe how to generate bilinear model from our topologically uniformed models. Given 20 registered meshes in different expressions, we use the example based facial rigging algorithm\cite{27} to generate 52 blendshapes based on FACS\cite{21} for each person. Then we follow the previous methods\cite{49, 12} to build the bilinear model from generated blendshapes in the space of 26317 vertices \times 52 expressions \times 938 identities. Specifically, we use Tucker decomposition to decompose the large rank-3 tensor to a small core tensor \(C_r\) and two low dimensional components for identity and expression. New face shape can be generated given the the identity parameter \(w_{id}\) and expression parameter \(w_{exp}\) as:

\[
V = C_r \times w_{exp} \times w_{id} \tag{1}
\]

where \(V\) is the vertex position of the generated mesh.

The superiority in quality and quantity of FaceScape makes the generated bilinear model own higher representation power. We evaluate the representation power of our model by fitting it to scanned 3D meshes not part of the training data. We compare our model to FaceWarehouse(FW)\cite{12} and FLAME\cite{29} by fitting them to our self-captured test set, which consists of 1000 high quality meshes from 50 subjects performing 20 different expressions each. FW has 50 identity parameters and 47 expression parameters, so we use the same number of parameters for fair comparison. To compare with FLAME which has 300 identity parameters and 100 expression parameters, we use 300 identity parameters and all 52 expression parameters. Figure 3 shows the cumulative reconstruction error. Our bilinear face model achieves much lower fitting error than FW using the same number of parameters and also outperform FLAME using even less expression parameters. The visually comparison in Figure 3 shows ours model could produce more mid-scale details than FW and FLAME, leading to more realistic fitting results.

4. Detailed Riggable Model Prediction

As reviewed in the related works in Section 2, existing methods have succeed in recovering extremely detailed 3D facial model from a single image. However, these recovered models are not riggable in expression space, since the recovered detail is static to the specific expression. Another
The objective function consisting of three parts. The first part is landmark alignment term. Assuming the camera is weak perspective, the landmark alignment term $E_{lan}$ is defined as the distance between the detected 2D landmark and its corresponding vertex projected on the image space. The second part is pixel-level consistency term $E_{pixel}$ measuring how well the input image is explained by a synthesized image. The last part is regularization term which formulates identity, expression, and albedo parameters as multivariate Gaussians. The final objective function is given by:

$$E = E_{lan} + \lambda_1 E_{pixel} + \lambda_2 E_{id} + \lambda_3 E_{exp} + \lambda_4 E_{alb}$$  \hspace{1cm} (2)

where $E_{id}$, $E_{exp}$ and $E_{alb}$ are the regularization terms of expression, identity and albedo, respectively. $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are the weights of different terms.

After obtaining the identity parameter $w_{id}$, individual-specific blendshapes $B_i$ can be generated as:

$$B_i = C_e \times \hat{w}_{exp}^{(i)} \times w_{id}, 0 \leq i \leq 51$$  \hspace{1cm} (3)

where $\hat{w}_{exp}^{(i)}$ is the expression parameter corresponding to blendshape $B_i$ from Tucker decomposition.

### 4.2. Displacement Map Prediction

Detailed geometry is expressed by displacement maps for our predicted model. In contrast to the static detail which is only related to the specific expression in a certain moment, dynamic detail expresses the geometry details in varying expressions. Since the single displacement map cannot represent the dynamic details, we try to predict multiple displacement maps for 20 basic expressions in FaceScape using a deep neural network.

We observed that the displacement map in a certain expression could be decoupled into static part and dynamic part. The static part tends to keep static in different expressions, and is mostly related to the intrinsic feature like pores, nevus, and organs. The dynamic part varies in different expressions, and is related to the surface shrinking and stretching. We use a deformable map to model the surface motion, which is defined as the difference of vertices’ 3D position from source expression to target expression in the UV space. As shown in Figure 4, we can see the variance between displacement maps is strongly related to the deforming map, and the static features in displacement maps are related to the texture. So we feed motion maps and textures to a CNN to predict the displacement map for multiple expressions.

We use pix2pixHD[50] as the backbone of our neural network to synthesize high resolution displacement maps. The input of the network is the stack of deforming map and static features in displacement maps. The combination of adversarial loss $L_{adv}$ and feature matching loss $L_{FM}$ is used to train our net with the loss function formulated as:

$$\min_G \left( \max_{D_1, D_2, D_3} \sum_{k=1,2,3} L_{adv}(G, D_k) \right) + \lambda \sum_{k=1,2,3} L_{FM}(G, D_k)$$  \hspace{1cm} (4)

Figure 4: Riggable details can be decoupled as static details and dynamic details. The static details can be estimated from the facial textures, while the dynamic details are strongly related to the facial deforming map.
where $G$ is the generator, $D_1$, $D_2$ and $D_3$ are discriminators that have the same LSGAN[31] architecture but operate at different scales, $\lambda$ is the weight of feature matching loss.

### 4.3. Dynamic Detail Synthesis

Inspired by[32], we synthesize displacement map $F$ for an arbitrary expression corresponding to specific blendshape weight $\alpha$, using a weighted linear combination of generated displacement maps $\hat{F}_0$ in neutral expression and $\hat{F}_i$ in other 19 key expressions:

$$F = M_0 \odot \hat{F}_0 + \sum_{i=1}^{19} M_i \odot \hat{F}_i$$

(5)

where $M$ is the weight mask with the pixel value between 0 and 1 , $\odot$ is element-wise multiplication operation. To calculate the weight mask, considering the blendshape expressions change locally, we first compute an activation mask $A_j$ in UV space for each blendshape mesh $e_j$ as:

$$A_j(p) = ||e_j(p) - e_0(p)||_2$$

(6)

where $A_j(p)$ is the pixel value at position $p$ of the $j$th activation mask, $e_j(p)$ and $e_0(p)$ is the corresponding vertices position on blendshape mesh $e_j$ and neutral blendshape mesh $e_0$, respectively. The activation masks are further normalized between 0 and 1. Given the activation mask $A_j$ for each of the 51 blendshape meshes, the $i$th weight mask $M_i$ is formulated as a linear combination of the activation masks weighted by the current blendshape weight $\alpha$ and fixed blendshape weight $\alpha_i$ corresponding to the $i$th key expression:

$$M_i = \sum_{j=1}^{51} \alpha^j \hat{\alpha}_i^j A_j$$

(7)

where $\alpha^j$ is the $j$th element of $\alpha$. $M_0$ is given by $M_0 = \max(0, 1 - \sum_{i=1}^{19} M_i)$.

There are many existing performance driven facial animation methods generating blendshape weights with depth camera[51, 28, 9] or single RGB camera[11, 10, 14]. As blendshape weights have semantic meaning, it’s easy for artists to manually adjust the rigging parameters.

### 5. Experiments

#### 5.1. Implement Detail

We use 888 people in our dataset as training data with a total of 17760 displacement maps, leaving 50 people for testing. We use the Adam optimizer to train the network with learning rate as $2e^{-4}$. The input textures and output displacement maps’ resolution of our network is both $1024 \times 1024$. We use 50 identity parameters, 52 expression parameters and 100 albedo parameters for our parametric model in all experiments.

#### 5.2. Evaluation of 3D Model Prediction

The predicted riggable 3D faces are shown in Figure 6. To show riggable feature of the recovered facial model, we rig the model to 5 specific expressions. We can see the results of rigged models contain the photo-realistic detailed wrinkles, which cannot be recovered by previous methods. The point-to-plane reconstruction error is computed between our model and the ground-truth shape. The mean error is reported in Table 2. More results and the generated animations are shown in the supplementary material.

#### 5.3. Ablation Study

**W/O dynamic detail.** We try to use only one displacement map from source image for rigged expressions, and the other parts remain the same. As shown in Figure 9, we find that the rigged model with dynamic detail shows the wrinkles caused by various expressions, which are not found in W/O dynamic method.
Table 2: 3D face Prediction Error

<table>
<thead>
<tr>
<th>method</th>
<th>mean error</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>our method (all exp.)</td>
<td>1.39</td>
<td>2.33</td>
</tr>
<tr>
<td>our method (source exp.)</td>
<td>1.22</td>
<td>1.17</td>
</tr>
<tr>
<td>DFDN[15] (source exp.)</td>
<td>2.19</td>
<td>3.20</td>
</tr>
<tr>
<td>Extreme3D[44] (source exp.)</td>
<td>2.06</td>
<td>2.55</td>
</tr>
<tr>
<td>3DDFA[59] (source exp.)</td>
<td>2.17</td>
<td>3.23</td>
</tr>
</tbody>
</table>

**W/O deforming Map.** We change the input of our displacement map prediction network by replacing the deforming map with one-hot encoding for each of 20 target expressions. As shown in Figure 9, we find the results without deforming map (W/O Def. Map) contain few details caused by expressions.

5.4. Comparisons to Prior Works

We show the predicted results of our result and other works in Figure 7. The comparison of detail prediction is
Figure 7: Comparison of static 3D face prediction with previous methods. The images in top two rows are from FaceScape, the image in third row is from volker sequence[48], and the image in bottom row is from Internet. The top three images are with ground truth shapes, so we evaluate the reconstruction error and show the heat map below each row. Our method predicts the lowest error comparing to three previous methods.

shown in Figure 8. As most of the detailed face predicted by other works cannot be directly rigged to other expressions, we only show the face shape in the source expression. Our results are visually better than previous methods, and also quantitatively better in the heat map of error. We consider the major reason for our method to perform the best in accuracy is the strong representation power of our bilinear model, and the predicted details contribute to the visually plausible detailed geometry.

6. Conclusion

We present a large-scale detailed 3D facial dataset, FaceScape. Comparing to previous public large-scale 3D face datasets, FaceScape provides the highest geometry quality and the largest model amount. We explore to pre-

dict a detailed riggable 3D face model from a single image, and achieve high fidelity in dynamic detail synthesis. We believe the release of FaceScape will spur the future researches including 3D facial modeling and parsing.

Acknowledgements

This work was supported by the grants – NSFC 61627804 / U1936202, USDA 2018-67021-27416, JSNSF BK20192003, and a grant from Baidu Research.
References

[21] Paul Ekman and Wallace V. Friesen. Facial action coding system: a technique for the measurement of facial movement. 1978. 4
[34] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik


Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. PAMI, 2019.
