FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction

Hao Zhu*, *Member, IEEE*, Haotian Yang*, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, *Senior Member, IEEE* and Xun Cao, *Member, IEEE*

Abstract—In this paper, we present a large-scale detailed 3D face dataset, *FaceScape*, and the corresponding benchmark to evaluate single-view facial 3D reconstruction. By training on FaceScape data, a novel algorithm is proposed 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. We also use FaceScape data to generate the in-the-wild and in-the-lab benchmark to evaluate recent methods of single-view face reconstruction. The accuracy is reported and analyzed on the dimensions of camera pose and focal length, which provides a faithful and comprehensive evaluation and reveals new challenges. The unprecedented dataset, benchmark, and code have been released to the public for research purpose[1].

Index Terms—3D Morphable Model, Dataset, Benchmark, 3D Face Reconstruction

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 [1], [2], [3] or active depth sensor such as Kinect [4] and coded light [5]. 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 in the age between 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.

- *Hao Zhu and Haotian Yang contributed equally to this work.*
- *Haotian Yang, Hao Zhu, Yidi Zhang, Longwei Guo, Yanru Wang, Mingkai Huang, Qiu Shen, and Xun Cao are with Nanjing University, Nanjing, China.*
- *Ruigang Yang is with University of Kentucky, Lexington, KY, USA.*

[1] *https://github.com/zhuhao-nju/facescape.git*

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 uniformed 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 dimensions. Experiments show that our generated bilinear model exceeds previous methods in representative ability. Besides, a new benchmark containing in-the-wild and in-the-lab data is presented to evaluate single-view face reconstruction. 14 recent methods are evaluated on the dimensions of camera pose and focal length, which provides a comprehensive evaluation and reveals new challenges.

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. The predicted model can be rigged to various expressions with plausible detailed geometry.

Our contributions are shown in Figure 1, and are summarized as following:

- We present a large-scale 3D face dataset, FaceScape, consisting of 18,760 extremely detailed 3D face mod-

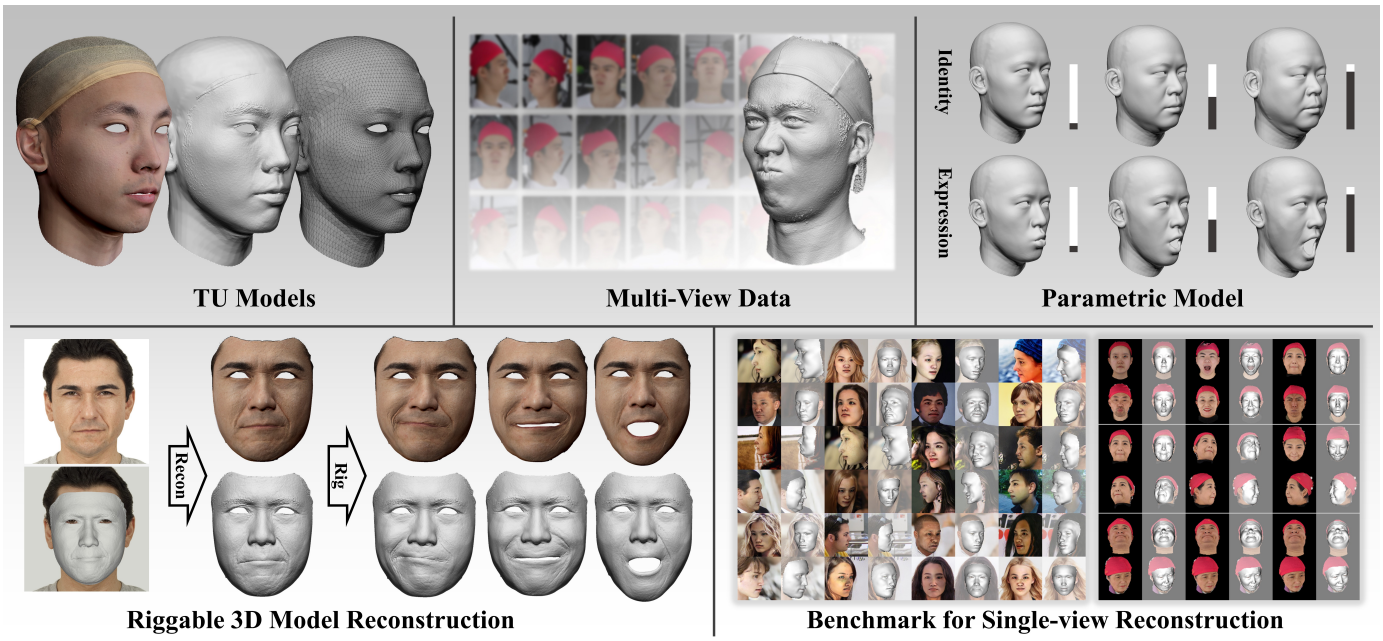


Fig. 1: FaceScape contains a large-scale detailed 3D face dataset and the benchmarks to evaluate single-view facial 3D reconstruction. By training on FaceScape data, a novel algorithm is proposed to predict elaborate riggable 3D face models from a single image input.

els. All the models are processed to topologically uniformed 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 acrossing 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.
- A new benchmark containing in-the-wild and in-the-lab data is presented to evaluate single-view face reconstruction, which reviews the SOTA methods on multiple dimensions and reveals new challenges.

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 [5], [6], [7], [8], [9] 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 [4], [10], [11], [12], [13], [14], [15], while the other works built sparse multi-view camera system [16], [17]. 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 [18], [19], [20]. 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. Our dataset outperforms previous works on both model quality and data amount.

3D Morphable Model. 3DMM is a statistical model which transforms the shape and texture of the faces into a vector space representation [21]. 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 [4], [22], [23], [24]. The models in expression dimension could be further transformed to a set of blendshapes [25], 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 [26], [27], [28], [29], [30], [31]. We recommend referring to the recent survey [32] for a comprehensive review.

Single-View Face Reconstruction. Generally the methods of single-view face reconstruction can be categorized into parametric methods and non-parametric methods. Parametric methods treat the reconstruction of facial shape as a 3DMM fitting problem, which optimizes or predicts the parameters of 3DMM from a single image [6], [27], [28], [30], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45]. These methods are also named 3D face alignment. Considering that parametric models cannot recover detailed shape due to the limited representation power, latter methods propose to firstly reconstruct a 3DMM, then refine it for

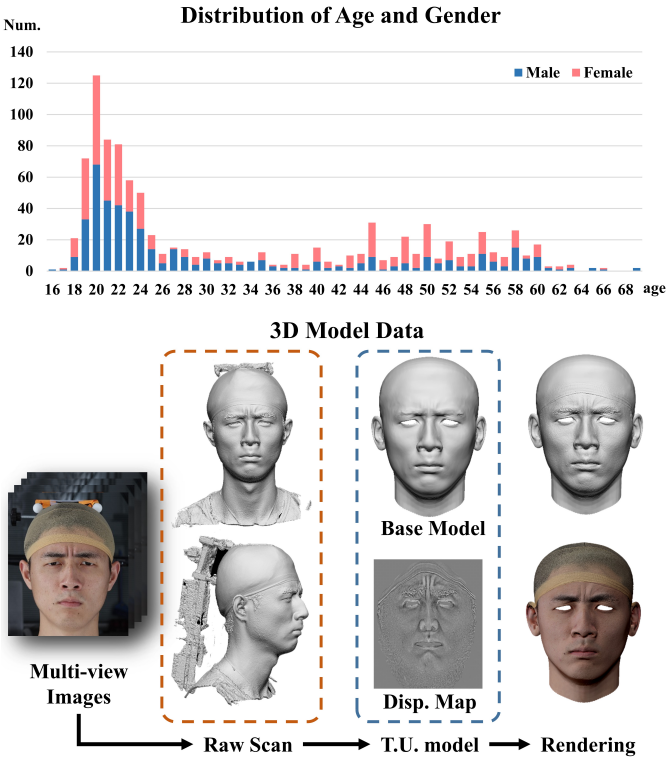


Fig. 2: Description of FaceScape dataset. In the upper side we show the histogram of subjects' age and gender. In the lower side we show the pipeline from the captured multi-view images to topologically uniformed models (T.U. models).

detailed shape [46], [47], [48], [49], [50], [51], [52], [53]. Non-parametric methods directly predict the facial shape in the form of mesh vertices [47], [54], [55], [56], [57], [58], depth [59], or volume [60]. Comparing to parametric model, non-parametric representation has higher degrees of freedom and therefore can express finer shape details. However, the free-form representation brings less constraint and makes the network difficult to predict the accurate shape.

Our work advances the state of the art in multiple aspects. In dataset, our FaceScape is by far the largest with the highest quality. In 3D face prediction, previous works focus on enhancing the static detailed facial shape, while we study the problem of recovering an **animatable** 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-depended 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 [4] 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.

3.2 3D shape Registration

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 [35] 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 Appearance-based Refinement

The 3D registration mentioned above only utilizes the geometry of the TU models to align mesh models. Though the registered shape is aligned with the raw scan, it fails to recover stretching deformations such as eyebrows raising and mouth stretching. Besides, the registered meshes may deviate from the semantic position of the texture in the flexible area like the mouth. The reason is that NICP only tries to minimize the nearest distance, disregarding the motions that are parallel to the surface. To solve these problems, we use the optical flow in UV texture space to refine the registration results.

We observe that the textures of a certain subject should be consistent in different expressions. However, due to the misalignment in the registration process, the texture maps of different expressions will shift, as shown in Figure 3. We project and blend the pixels of the multi-view images based on the preliminary registration to obtain the texture. Then the optical flow is estimated with DeepFlow [61] from the texture of other expressions to the neutral expression, the

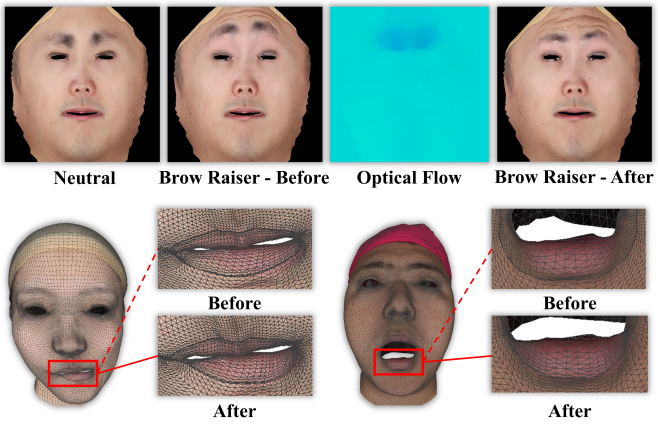


Fig. 3: In the first row, the texture shift between 'Neutral' and 'Brow Raiser' can be estimated as the optical flow, then the mesh grid can be refined with the updated texture shown on the right. In the second row, we compare the result before appearance-refinement and after appearance-refinement. The grid of lips are obviously misaligned in 'Before', and are fixed in 'After'.

vertex position of the 3D mesh is refined according to the optical flow.

Let M_n be the preliminary registration in neutral position and M_e in another expression, and the corresponding textures are I_n and I_e. The optical flow F from I_n to I_e is computed by DeepFlow if of the same size of the textures, which represent the dense correspondence between I_0 and I_1. The pixel with the coordinate (x, y) in $I_n$correspond to the pixel in I_1 with coordinate $(x-F(x,y,0), y-F(x,y,1))$. Based on the correspondence, we can obtain the updated vertices $\hat{V}'_e$ in M_e. That is, the vertices with texture coordinate (x, y) should have the same position of the vertices on M_e with texture coordinate $(x - F(x,y,0), y - F(x,y,1))$, so that the texture of M_e is consist to neutral expression. The position corresponding to $(x-F(x,y,0), y-F(x,y,1))$ is obtained by barycentric interpolation of triangle vertices. Due to the defect of optical flow calculation, directly updating the vertices' position will lead to a twisted mesh. So we add a smooth regularization item to ensure adjacent vertices do not change dramatically. The energy function to optimize the new vertices $\hat{V}$ is formulated as:

$$\hat{V} = \arg\min_{\hat{V}} ||\hat{V} - \hat{V}'||^2 + \sum_{(i,j)\in\Omega} ||(\hat{V}_i - \hat{V}'_i) - (\hat{V}_j - \hat{V}'_j)||^2 \tag{1}$$

where Ω are the edges in the mesh.

We show the registration before and after appearance-based refinement in Figure 3. Though the shape of the preliminary registration is close to the raw scan, the semantic meaning of the triangles and texture is misaligned, which is significantly alleviated after the refinement. More results showing the effectiveness of the appearance-refinement are shown in the appendix.

3.4 Bilinear Model

Bilinear model [22] 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 [25] to generate 52 blendshapes based on FACS [62] for each person. Then we follow the previous methods [4], [22] 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 $\mathbf{w}_{id}$ and expression parameter $\mathbf{w}_{exp}$ as:

$$V = C_r \times \mathbf{w}_{exp} \times \mathbf{w}_{id} \tag{2}$$

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) [4] and FLAME [23] 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 paremeters. Figure 4 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 4 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 group of works try to fit a parametric model to the source image, which will obtain an expression-riggable model, but the recovered geometry stays in the rough stage.

The emerge of FaceScape dataset makes it possible to estimate detailed and riggable 3D face model from a single image, as we can learn the dynamic details from the large amount of detailed facial models. We show our pipeline in Figure 6 to predict a detailed and riggable 3D face model from a single image. The pipeline consists of three stages: base model fitting, displacement map prediction and dynamic details synthesis. We will explain each stage in detail in the following sections.

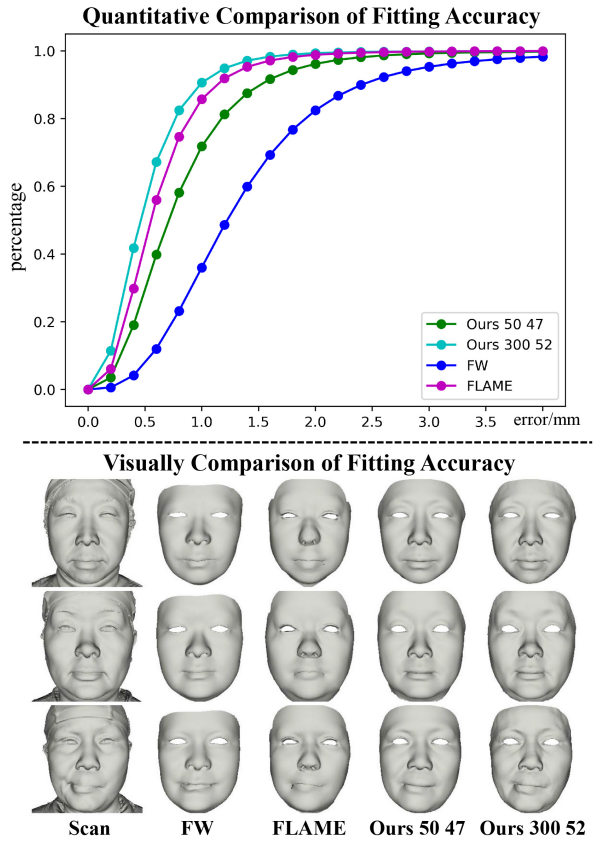


Fig. 4: Comparison of Reconstruction Error for parametric model generated by FaceScape and previous datasets.

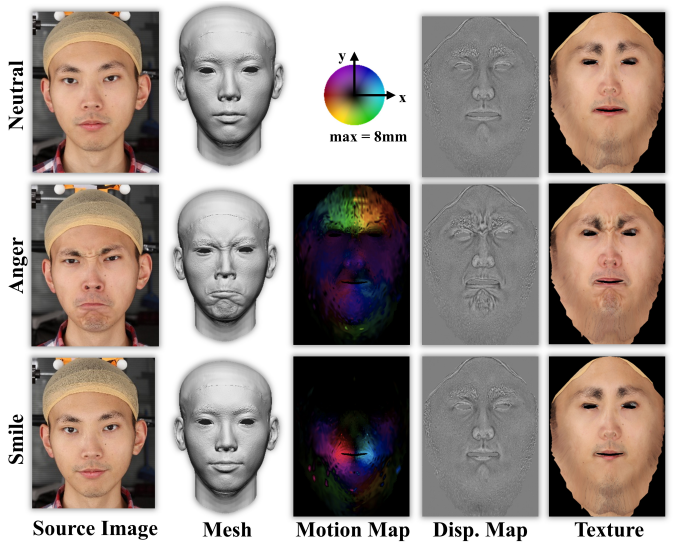


Fig. 5: Riggable details can be decoupled as static details and dynamic details. The static details can be estimated from the facial textures, while the dynamic details are strongly related to the facial deforming map.

4.1 Base Model Fitting

The bilinear model for base shape is inherently riggable as the parametric space is separated into identity dimension and expression dimension, so the rough riggable model can be generated by regressing the parameters of identity for the bilinear model. Following [34], we estimate parameters corresponding to a given image by optimizing an 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:

$$E = E_{lan} + \lambda_1 E_{pixel} + \lambda_2 E_{id} + \lambda_3 E_{exp} + \lambda_4 E_{alb} \quad (3)$$

where E_{id}, E_{exp} and E_{alb} are the regularization terms of expression, identity and albedo, respectively. λ_1, λ_2, λ_3 and λ_4 are the weights of different terms.

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

$$B_i = C_r \times \hat{w_{exp}}^{(i)} \times w_{id}, 0 \leq i \leq 51 \quad (4)$$

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

Learning-based fit. As an alternative to the optimization-based fit as explained above, we can train a neural network to regress the parameters of the bilinear model and camera pose. We use ShuffleNet_v2 [63] as our regressor, which takes a 256×256 image as input and predicts a 107-dimension vector. The predicted vector consists of 50 identity parameters, 51 expression parameters, rotation quaternion with scale (4 parameters), and translation in image coordinate (2 parameters). We use 17810 images from CelebA [64] as a training set, and generate the parameters label using the optimization-based fitting as explained above. Learning-based fit is much faster than the optimization-based method and achieves comparable accuracy. However, the learning-based fit is unstable for some specific poses such as head raising and bowing, because these poses are rare in the training set generated from CelebA [64]. The performance of learning based fit and optimization based fit are evaluated in Section 6.4, denoted as FaceScape(Learn) and FaceScape(Opti.) respectively.

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 deforming map to model the surface motion, which is defined as the difference of vertices' 3D

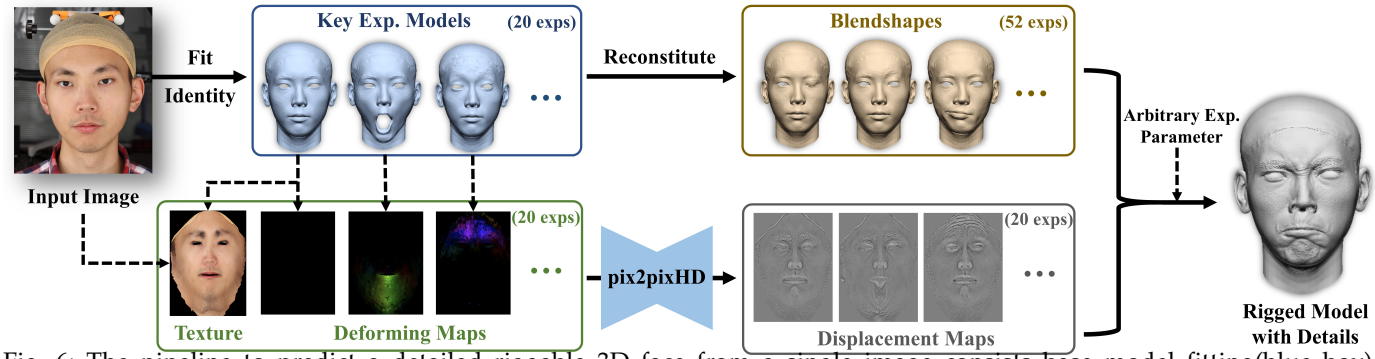


Fig. 6: The pipeline to predict a detailed riggable 3D face from a single image consists base model fitting(blue box), displacement map prediction (green and grey box), and dynamic details synthesis (yellow box and the followed arrow on the right).

position from source expression to target expression in the UV space. As shown in Figure 5, 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 [65] 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 texture in UV space, which can be computed from the recovered base model. Similar to [65], 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:

$$\min_{G}((\max_{D_1,D_2,D_3}\sum_{k=1,2,3}L_{adv}(G,D_k)) + \lambda\sum_{k=1,2,3}L_{FM}(G,D_k)) \tag{5}$$

where G is the generator, D_1, D_2 and D_3 are discriminators that have the same LSGAN [66] architecture but operate at different scales, λ is the weight of feature matching loss.

4.3 Dynamic Detail Synthesis

Inspired by [67], we synthesize displacement map F for an arbitrary expression corresponding to specific blendshape weight $\boldsymbol{\alpha}$, using a weighted linear combination of generated displacement maps $\hat{F}_0$ in neutral expression and $\hat{F}_i$ in other 19 key expressions:

$$F = M_0 \odot \hat{F}_0 + \sum_{i=1}^{19} M_i \odot \hat{F}_i \tag{6}$$

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:

$$A_j(p) = ||e_j(p) - e_0(p)||_2 \tag{7}$$

where $A_j(p)$ is the pixel value at position p of the jth 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 ith weight mask M_i is formulated as a linear combination of the activation masks weighted by the current blendshape weight $\boldsymbol{\alpha}$ and fixed blendshape weight $\hat{\boldsymbol{\alpha}}_i$ corresponding to the ith key expression:

$$M_i = \sum_{j=1}^{51} \boldsymbol{\alpha}^j \hat{\boldsymbol{\alpha}}_i^j A_j \tag{8}$$

where $\boldsymbol{\alpha}^j$ is the jth element of $\boldsymbol{\alpha}$. 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 [68], [69], [70] or single RGB camera [71], [72], [73]. 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 1024×1024. 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 7. 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 1. 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

Fig. 7: We show our predicted faces in source expression and rigged expressions. It is worth noting that the wrinkles in rigged expressions are predicted from the source image.

TABLE 1: 3D face Prediction Error

method	mean error	variance
our method (all exp.)	1.39	2.33
our method (source exp.)	**1.22**	**1.17**
DFDN [51] (source exp.)	2.19	3.20
Extreme3D [52] (source exp.)	2.06	2.55
3DDFA [75] (source exp.)	2.17	3.23

other parts remain the same. As shown in Figure 10, 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.

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 10, 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 8. The comparison of detail prediction is shown in Figure 9. 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

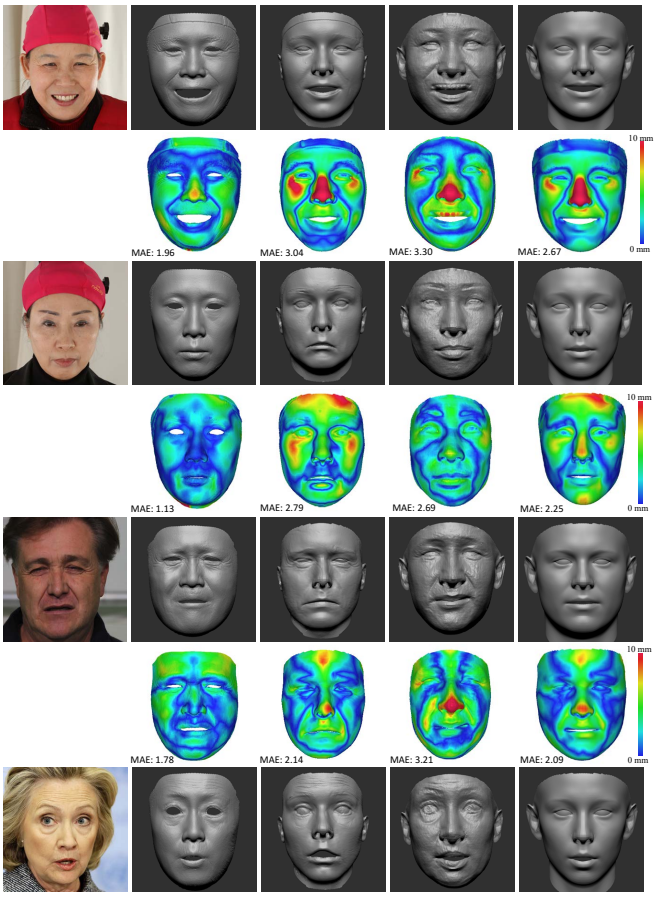


Fig. 8: Comparison of static 3D face prediction with previous methods. The images in top two rows are from FaceScape, the image in third row is from volker sequence [74], and the image in bottom row is from Internet. The top three images are with ground truth shapes, so we evaluate the reconstruction error and show the heat map below each row.

model, and the predicted details contribute to the visually plausible detailed geometry.

6 Benchmark

In recent years, a large number of methods are presented to recover 3D facial shapes from a single image. Surprisingly, there is very few benchmark and data for single-view face reconstruction that can take all aspects into account. The common evaluating way is to run the method on tuples of image-model pairs, where the image is the input, then the error between the output and the ground-truth model is reported. These image-model pairs for evaluation should meet the following requirements: 1) Accurate - the 3D model is accurately captured and is well aligned to the image; 2) Photo-realistic - the image should be like the in-the-wild photos; 3) Amount - The amount of data should be large enough to cover various appearance, poses, expressions, environments, and lightings. In this section, we propose FaceScape benchmark to evaluate the accuracy of 3D facial shape reconstruction, consisting of FaceScape-Wild (FS-Wild) and FaceScape-Lab (FS-Lab) data. FS-Wild data takes

Fig. 9: Comparison of detail prediction. We adopt NICP [35] to register the base meshes of different methods to ground truth scans, and visualize the predicted details on common base meshes.

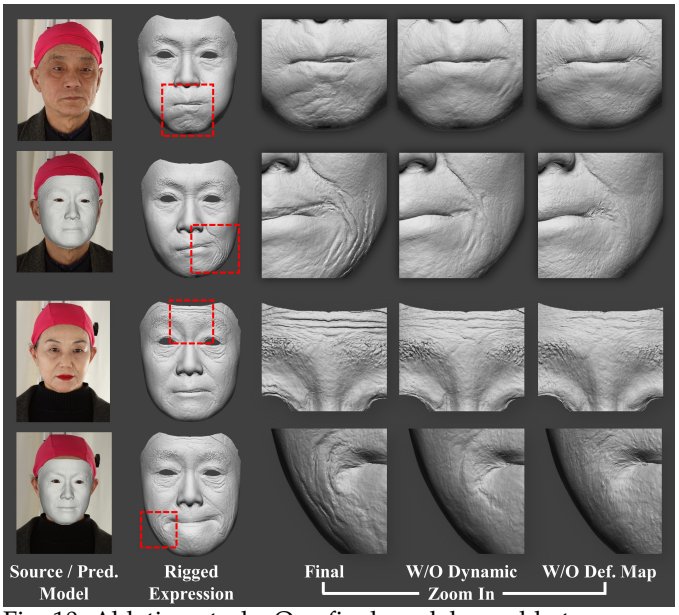


Fig. 10: Ablation study. Our final model are able to recover wrinkles in rigged expressions, while the method W/O demforming map and W/O dynamic details cannot.

advantage of face-swapping methods to synthesize photo-realistic in-the-wild images with exact ground-truth 3D shapes. FS-Lab dataset uses the unpublished in-the-lab images captured by our multi-view system, and corresponding accurate 3D model as ground-truth. We will explain the data generation, metrics, and evaluation results in the following subsections.

6.1 FS-Wild data generation

We create FS-Wild data by swapping the in-the-wild faces with rendered faces from ground-truth models. In this way, we obtain the photo-realistic in-the-wild facial images and the corresponding exact 3D ground-truth shape for evaluation. The pipeline to synthesize the final image is shown in Figure 11, which consists of three stages.

In the first stage (1/2/CompNet in Figure 11), we fit the bilinear model to the source in-the-wild image using the algorithm described in Section 4.1. A texture map is then randomly selected from the pool of TU models, according to the gender and age of the subject in the source image. The fitted model and texture map are then rendered to the image with spherical harmonic illumination [76]. Considering that the eyes and mouth are missing, we train a CompNet to complement the eyes and mouth. CompNet takes pix2pixHD [77] as the backbone and is trained using the captured images with eyes and mouth. We can see that CompNet generates photo-realistic eyes and mouth for the rendered image, and we also obtain the fitted model as the ground-truth shape.

In the second stage (3/4/SynNet in Figure 11), we first extract the semantic mask, then adjust the facial mask to fit the completed image. For the images from CelebAMask-HQ [78], the provided ground-truth masks are used. For the images from AFLW [79] that do not contain a ground-truth mask, BiSeNet [80] is used to extract the semantic mask from the source image. The aim of the adjustment is to make the facial mask of the source image and rendered image completely aligned. As model fitting cannot make the rendered mask exactly align with the facial mask, mask adjustment is essential for the following face-swap stage. Otherwise, the face-swapping result will contain artifacts around the cheek. Then, SynNet is used to generate photo-realistic faces from the adjusted semantic masks. We refer to SEAN [81] as our SynNet, which is a generative adversarial network conditioned on a semantic mask. In this way, the photo-realistic image that is exactly aligned with the fitted model is generated.

In the last stage, we use the SwapNet to replace the facial region of the adjusted image with that of the completed image. There are a large number of works that study the face-swapping problem, but most of them cannot maintain the 3D structure of the completed images. By experiments, we find the inpainting and blending module in FSGAN [82] meets our requirement, which keeps faithful 3D structure of the completed image while synthesizing photo-realistic in-the-wild images. Finally, we obtained the final image as the input for evaluation, and the fitted model as the corresponding ground-truth shape.

Implement details. In the rendering stage, we approximate the scene illumination with Spherical Harmonics [76], and set the spherical harmonic coefficient as 9. The backbones of our CompNet is the pix2pixHD [77] with a U-Net as the generator. The CompNet is trained with reconstruction loss $\mathcal{L}_{rec}$ and adversarial loss $\mathcal{L}_{adv}$:

$$\min_G((\max_D \mathcal{L}_{adv}(G,D)) + \lambda \mathcal{L}_{rec}(G)), \tag{9}$$

where G is the generator, D is the discriminator and λ is the weight of reconstruction loss. The reconstruction loss $\mathcal{L}_{rec}$ is given by:

$$\mathcal{L}_{rec} = \|G(x) - x\|_1 . \tag{10}$$

The adversarial loss is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))]. \tag{11}$$

Our SynNet is trained on CelebAMask-HQ following the training methods in SEAN. To adjust the semantic mask, we first remove the headgear region from the facial region of the rendered image, then take the intersection of the facial part of the semantic mask and the rendered image. For the region that is contained by the semantic mask but is not contained by the rendered image, the semantic label is replaced by the nearest non-facial label. In this way, the adjusted mask has the same face contour as the rendered image. The input image size of SynNet and CompNet is 512×512 and the output image size of SwapNet is 256×256.

Results of In-the-wild Face Rendering. The benchmark of FS-Wild consists of 400 face images of 400 synthesized subjects. The data are uniformly divided into 4 sets according to the angle between camera orientation and face orientation ($0-5°, 5°-30°, 30°-60°, 60°-90°$), with a reference 3D face model per subject. The images consist of indoor and outdoor images, neutral expression and expressive face images, and varying viewing angles ranging from frontal view to side view.

6.2 FS-Lab data generation

We render 330 images using the 20 detailed 3D models, which are randomly selected from the unpublished testing set of FaceScape. These subjects' age ranges from 17 to 63, with an average age of 38.7. Centering on the head and starting from the front, we select 11 different camera location with (yaw, pitch) coordinates:

- 1 **camera at exact front**: $(0°, 0°)$.
- 8 **cameras deflecting** 30°: $(0°, \pm 30°)$, $(\pm 21.47°, \pm 21.47°)$, $(\pm 30°, 0°)$.
- 2 **cameras deflecting** 60°: $(\pm 60°, 0°)$.

The samples images for 11 camera locations are shown in Figure 12.

The images are rendered in the resolution of 256×256 using 3 different focal length: 1200 (long focal), 600 (middle focal), 300 (short focal). The samples for different focal length are shown in Figure 12.

6.3 Metric

We propose to quantitatively evaluate the accuracy of the single-view face reconstruction methods using Chamfer Distance (CD), Mean Normal Error (MNE), and Complete Rate (CR). Considering that some methods may fail on certain images, these metrics are only computed from the valid outputs, and we additionally report the success rate for the whole datasets.

Pre-process. Firstly, we transform the predicted mesh and the ground-truth mesh into the camera coordinate. Considering different methods may use perspective projection camera or orthogonal projection camera, these projection models need to be adjusted to match the projection model used for ground-truth rendering. For FS-Wild data, all predicted mesh are transformed into orthogonal camera coordinate, and for FS-Lab data, all predicted meshes are transformed into the perspective camera coordinate with the ground-truth focal length. Then the depth of the predicted mesh is optimized to fit the ground-truth mesh by minimizing the depth difference. It is worth noting that we do not use ICP [84] or 7-point alignment [85] to register

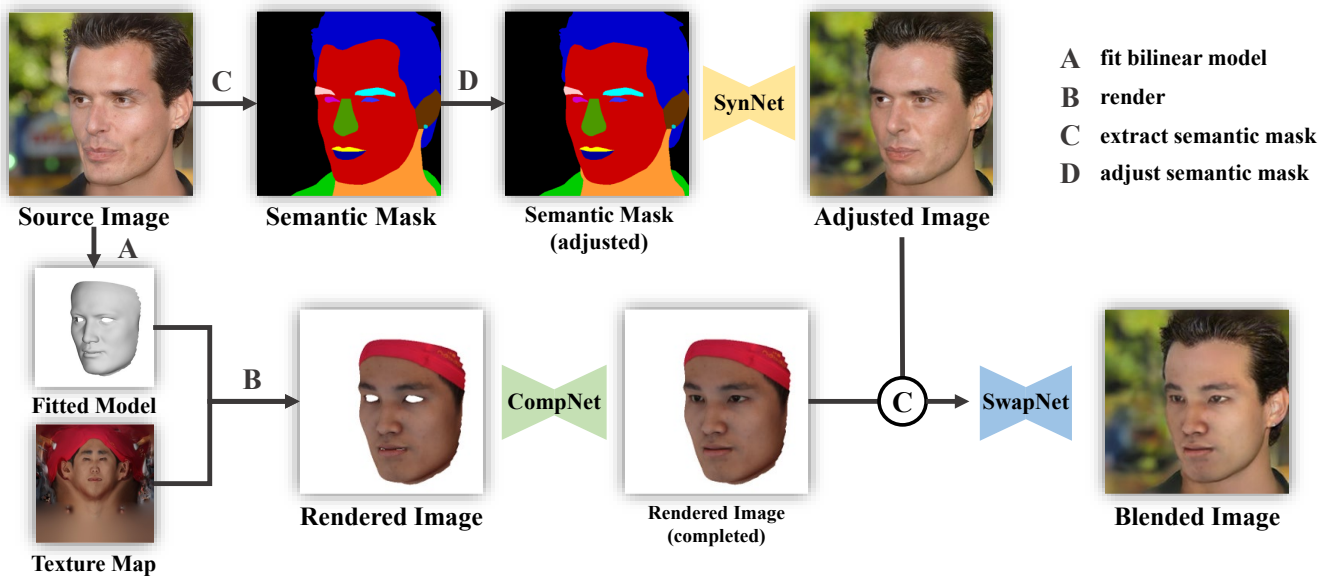


Fig. 11: Overview of the proposed in-the-wild face rendering approach.

TABLE 2: Quantitative evaluation on FS-Wild dataset categorised by pose angle.

Pose Angle → / Method ↓	$0° - 5°$ CD	$0° - 5°$ MNE	$0° - 5°$ CR	$5° - 30°$ CD	$5° - 30°$ MNE	$5° - 30°$ CR	$30° - 60°$ CD	$30° - 60°$ MNE	$30° - 60°$ CR	$60° - 90°$ CD	$60° - 90°$ MNE	$60° - 90°$ CR	Success Rate
Ext3dFace [52]	5.03	0.158	61.5	5.52	0.176	55.7	7.92	0.208	40.4	25.39	0.266	27.1	85.5
PRNet [54]	2.61	0.119	83.0	3.11	0.114	82.7	4.26	0.119	78.2	3.88	0.140	75.3	100.0
Deep3DFaceRec [37]	2.30	0.070	83.1	2.50	0.072	83.0	3.57	0.082	77.8	6.81	0.143	62.4	100.0
RingNet [40]	2.40	0.085	99.8	2.99	0.085	99.7	4.78	0.100	98.4	10.71	0.190	97.1	100.0
DFDN [56]	3.66	0.090	86.6	3.27	0.091	86.5	7.29	0.128	84.3	27.48	0.302	57.2	88.2
DF2Net [56]	2.92	0.121	57.1	4.21	0.128	55.3	6.55	0.159	46.3	19.76	0.309	30.8	68.8
UDL [50]	2.27	0.091	69.0	2.59	0.092	68.3	3.46	0.106	65.0	6.32	0.176	49.0	86.2
FaceScape(Opti.) [83]	2.81	0.086	83.7	3.17	0.092	82.0	4.09	0.108	79.0	6.57	0.162	67.9	96.0
FaceScape(Learn) [83]	2.70	0.085	86.9	3.69	0.092	86.5	4.23	0.099	85.2	9.10	0.151	70.6	100.0
MGCNet [38]	2.97	0.073	84.4	2.94	0.073	84.5	2.78	0.070	81.6	4.21	0.091	74.3	100.0
3DDFA_v2 [39]	2.49	0.074	86.5	2.66	0.074	86.0	3.18	0.078	83.1	3.67	0.093	79.9	100.0
SADRNet [55]	6.60	0.113	90.2	6.87	0.113	89.4	6.40	0.103	84.4	8.63	0.163	82.7	100.0
LAP [59]	4.19	0.111	93.5	4.47	0.116	92.8	6.16	0.148	87.3	13.71	0.205	68.1	100.0
DECA [53]	2.88	0.080	99.9	2.64	0.079	99.9	2.88	0.082	99.8	4.83	0.116	99.7	100.0

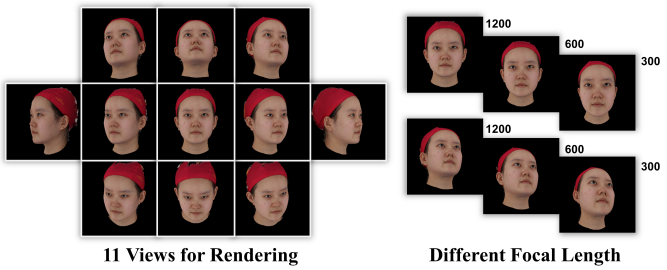


Fig. 12: Samples of in-the-lab data in different views and different focal lengths.

the predicted mesh to the ground-truth mesh, as used in Section 5.2, but only align them by moving the predicted mesh along the depth direction. We believe our alignment tells more because ICP and 7-point alignment changes the pose of the predicted mesh. Because of this, the chamfer distance reported in the benchmark is slightly larger than the error distance reported in Section 5.2.

We observed that the grid density of the mesh is not uniformed, which may cause bias in the quantitative evaluation. Therefore, all the predicted mesh and the ground-truth mesh are re-sampled by projecting them to the cylindrical coordinate. The central axis of the cylindrical coordinate crosses the barycenter of the ground-truth head mesh, pointing to the overhead direction. We found that projecting to cylindrical coordinate is better than sphere coordinate and Euclidean coordinate, because it causes the fewest occlusions and makes the sampling of points most uniform. When re-sampling meshes from the position map in the sphere coordinate, the faces with edges larger than $15mm$ are ignored.

Chamfer Distance. CD measures the overall error distance. Given the processed predicted mesh $\mathcal{M}_p$ and the ground-truth mesh $\mathcal{M}_g$, chamfer distance is formulated as:

$$CD(\mathcal{M}_p, \mathcal{M}_g) = \frac{1}{N_p} \sum_{x \in \mathcal{M}_p}^{N_p} \sum_{y \in \mathcal{M}_g}^{N_g} \min \|x - y\|_2 + \frac{1}{N_g} \sum_{y \in \mathcal{M}_g}^{N_g} \sum_{x \in \mathcal{M}_p}^{N_p} \min \|x - y\|_2 \tag{12}$$

where N_p, N_g are the numbers of the vertices of the predicted mesh and the ground-truth mesh respectively.

Mean Normal Error (MNE). MNE measures the accuracy of the detailed geometry in the middle-scale and small-

TABLE 3: Quantitative evaluation on FS-Lab benchmark categorised by pose angle.

Pose Angle → / Method ↓	0° CD	0° MNE	0° CR	30° CD	30° MNE	30° CR	60° CD	60° MNE	60° CR	Success Rate
Ext3dFace [52]	4.59	0.131	86.2	7.42	0.170	69.1	8.51	0.175	55.2	85.9
PRNet [54]	2.94	0.133	92.5	3.40	0.125	90.1	3.74	0.122	85.2	100.0
Deep3DFaceRec [37]	3.99	0.106	87.6	5.90	0.120	81.3	5.55	0.137	75.3	98.9
RingNet [40]	3.62	0.102	99.9	5.03	0.111	99.7	6.82	0.151	94.5	100.0
DFDN [56]	4.28	0.111	98.4	6.71	0.132	95.2	23.63	0.280	81.0	94.7
DF2Net [56]	4.48	0.152	64.1	7.64	0.200	52.2	-1.00	-1.000	-100.0	73.6
UDL [50]	2.21	0.092	79.5	5.34	0.123	71.3	5.63	0.167	61.9	87.0
FaceScape(Opti.) [83]	3.21	0.090	94.2	4.87	0.119	86.2	4.68	0.146	81.7	92.0
FaceScape(Learn) [83]	2.40	0.086	96.7	7.28	0.124	87.7	3.87	0.108	90.5	100.0
MGCNet [38]	3.45	0.085	92.7	3.91	0.093	90.1	3.65	0.090	83.2	100.0
3DDFA_v2 [39]	3.05	0.093	95.2	3.41	0.096	93.8	3.82	0.097	88.2	100.0
SADRNet [55]	4.25	0.109	95.8	7.07	0.137	94.9	7.09	0.148	87.6	100.0
LAP [59]	4.27	0.112	96.4	7.33	0.149	93.2	8.70	0.195	85.6	99.2
DECA [53]	3.30	0.093	99.8	4.14	0.100	99.4	4.20	0.107	97.1	100.0

TABLE 4: Quantitative evaluation on FS-Lab benchmark categorised by focal length.

Focal Length → / Method ↓	Long(1200) CD	Long(1200) MNE	Long(1200) CR	Mid(600) CD	Mid(600) MNE	Mid(600) CR	Short(300) CD	Short(300) MNE	Short(300) CR	Success Rate
Ext3dFace [52]	7.25	0.167	69.4	6.72	0.162	64.9	6.03	0.160	61.4	85.9
PRNet [54]	3.42	0.125	89.4	3.48	0.124	89.0	3.79	0.128	90.2	100.0
Deep3DFaceRec [37]	5.67	0.122	80.8	5.28	0.117	79.2	4.90	0.114	81.1	98.9
RingNet [40]	5.23	0.117	98.8	5.25	0.117	99.4	5.37	0.119	99.8	100.0
DFDN [56]	9.05	0.153	93.3	9.40	0.149	92.8	9.30	0.146	94.6	94.7
DF2Net [56]	7.26	0.194	53.7	6.97	0.191	51.2	6.39	0.183	49.5	73.6
UDL [50]	5.06	0.126	70.9	4.91	0.124	69.2	4.95	0.125	69.7	87.0
FaceScape(Opti.) [83]	4.69	0.120	86.4	4.77	0.121	85.2	5.47	0.126	83.6	92.0
FaceScape(Learn) [83]	6.21	0.118	89.0	6.19	0.118	88.8	6.43	0.125	86.4	100.0
MGCNet [38]	3.82	0.091	89.1	4.01	0.091	89.4	4.18	0.098	91.3	100.0
3DDFA_v2 [39]	3.45	0.096	92.9	3.51	0.094	92.7	3.85	0.097	93.5	100.0
SADRNet [55]	6.81	0.137	93.6	6.82	0.132	95.0	6.60	0.131	97.1	100.0
LAP [59]	7.30	0.154	92.1	7.06	0.151	91.4	6.75	0.150	91.7	99.2
DECA [53]	4.07	0.100	99.0	4.19	0.101	99.6	5.81	0.122	99.8	100.0

scale. To compute MNE, we first render the predicted mesh and the ground-truth mesh in the cylindrical coordinate, generating predicted normal map $\mathcal{N}_p$ and ground-truth normal map $\mathcal{N}_g$. Only the intersection of the valid region for $\mathcal{N}_p$ and $\mathcal{N}_g$ are reserved, so the pixels in $\mathcal{N}_p$ and $\mathcal{N}_g$ are one-to-one matched. Then, MNE is formulated as:

$$MNE(\mathcal{N}_p, \mathcal{N}_g) = \frac{1}{N_n} \sum_{x,y \in \mathcal{N}_p, \mathcal{N}_g}^{N_n} \left(\frac{x \cdot y}{\|x\|_2 \cdot \|y\|_2} \right) \tag{13}$$

where N_n is the number of valid pixels in the two normal maps.

Complete Rate (CR). CR measures the integrity of the reconstruction results. To compute CR, we first render the predicted mesh and the ground-truth mesh in the cylindrical coordinate as the position map P_p and P_g respectively. then CR is formulated as:

$$\eta = \frac{S(P_p \cap P_g)}{S(P_g)} \tag{14}$$

where $S(P)$ is the function that returns the area of the position map P.

6.4 Evaluation

We evaluate 14 recent single-view face reconstruction methods, and the quantitative results are reported in Table 2 3 4. Table 2 show the scores on FS-Wild data categorised by pose angle. Table 3 and Table 4 show the scores on FS-Lab data categorised by pose angle and focal length respectively.

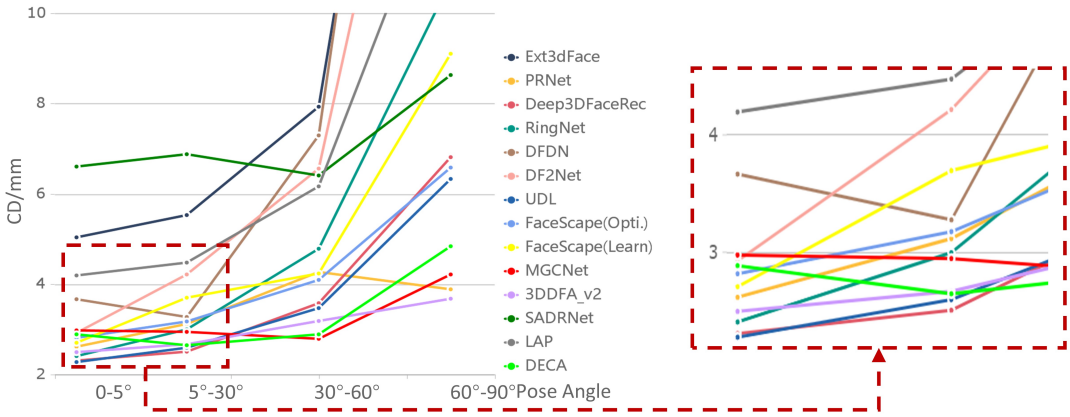
It is worth noting that some methods are special in certain aspects: FaceScape(Opti./Learn) [83] and DECA [53] predicts riggable mesh model; LAP [59] is trained with the in-the-wild photo collections, and MGCNet [38] is trained with multi-view images; Ext3dFace [52], DFDN [51], and DF2Net [56] explicitly reconstructs a rough base model and a refined model, here we use the refined model for evaluation.

Seeing from pose angle. In Table 2 3 and Figure 13-(a/b), we can see that most of the methods performs well in frontal views, but degraded severely for the side views. PRNet [54], 3DDFA_v2 [39], and MGCNet [38] are few methods that are relatively stable for side views.

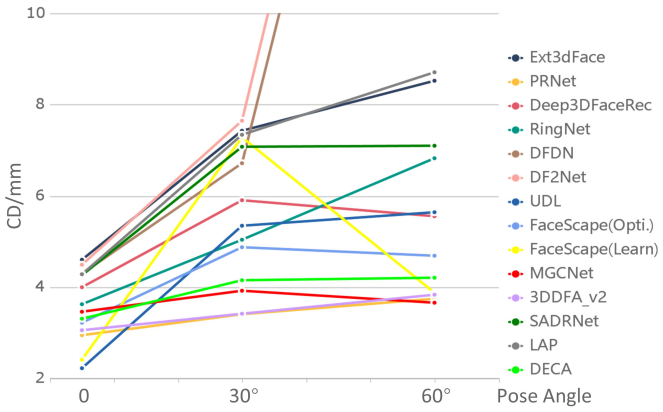
Seeing from focal length. In Table 4 and Figure 13-(c), we can see that some methods performs better for long focal while others do the opposite. The reason is that some methods assume the orthogonal projection camera that is more close to the long focal camera, while the others assume a perspective projection camera with a pre-defined focal length. Reconstructing accurate faces for both long and short focal images is still a challenge.

6.5 Comparison with Other Benchmarks

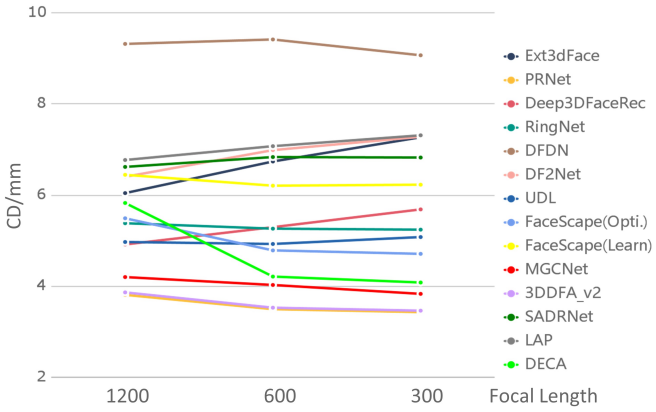
Early benchmarks for 3D face reconstruction use fitted 3D landmarks [86] or fitted 3DMM [8], [9] for evaluation. Considering the accuracy of fitted 3DMM is low, latter benchmarks choose to capture RGBD data for evaluation [85], [87].



(1) CD evaluated on FS-Wild dataset (pose angle)



(2) CD evaluated on FS-Lab dataset (pose angle)



(3) CD evaluated on FS-Lab dataset (focal length)

Fig. 13: Charts to visualize the quantitative evaluation.

NoW benchmark [87] and Feng *et al*. 's benchmark [85] are the only large-scale reconstructed benchmarks for 3D face reconstruction. NoW benchmark contains 2054 images of 100 subjects captured with an iPhone X, and a separate 3D head scan for each subject. Feng *et al*. 's benchmark contains 2000 images of 125 subjects captured by an RGB camera and a head scanner. Both benchmarks take various head poses, expressions, and lighting conditions into account, while the accuracy of ground-truth shape is limited to the hand-hold 3D scanner and the aligning algorithm. Comparing to these methods, our benchmark provides extremely well-aligned and fidelity 3D shape captured by multi-view camera array (error ¡ 0.2mm) as ground-truth, and provides in-the-wild data and in-the-lab data for a comprehensive evaluation. Our benchmark considers comparing methods according to poses, focal length, and expressions, which reveals more challenges. Besides, our benchmark and the previous two are complementary in terms of race: the subjects of our benchmark are mostly oriental faces, while the subjects of NoW and Feng *et al*. 's benchmark are mostly western faces. The comparison between facescape-benchmark and NoW benchmark are summarized in Table 5.

7 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, and also a comprehensive benchmark for evaluating single-view face reconstruction. We explore to predict 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 future researches on 3D facial modeling and parsing.

References

[1] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3d morphable model learnt from 10,000 faces," in *CVPR*, 2016, pp. 5543–5552. 1

[2] H. Dai, N. Pears, W. A. Smith, and C. Duncan, "A 3d morphable model of craniofacial shape and texture variation," in *ICCV*, 2017, pp. 3085–3093. 1

[3] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *CVPR*, 2005, pp. 947–954. 1

[4] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *TVCG*, vol. 20, no. 3, pp. 413–425, 2013. 1, 2, 3, 4

[5] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301. 1, 2

[6] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3d solution," in *CVPR*, 2016, pp. 146–155. 2

[7] Y. Guo, J. Cai, B. Jiang, J. Zheng *et al*., "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *PAMI*, vol. 41, no. 6, pp. 1294–1307, 2018. 2

TABLE 5: Comparison with previous benchmarks.

benchmark	NoW [87]	Feng *et al.* [85]	FS-Lab	FS-Wild
Subject Amount	100	135	20	400
Accuracy of GT	high	high	very high	very high
Variety of Identity	western faces	western faces	oriental faces	oriental faces
Variety of Expression	daily expressions	daily expressions	20 specified expressions	daily expressions
Variety of Enviroment	street scenes	indoor scenes	laboratory studio	in-the-wild as CelebA [78])
Multiple Focal Length	No	No	No	three types
Sense of Reality	real-captured	real-captured	real-captured	somewhat photo-realistic

[8] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, "3d face morphable models "in-the-wild"," in *CVPR*, 2017, pp. 5464–5473. 2, 11

[9] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou, "3d reconstruction of in-the-wild faces in images and videos," *PAMI*, vol. 40, no. 11, pp. 2638–2652, 2018. 2, 11

[10] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *FG*, 2006, pp. 211–216. 2

[11] L. Yin, X. Chen, Y. Sun, W. Tony, and J. R. Michael, "A high-resolution 3d dynamic facial expression database, 2008," in *FG*, vol. 126, 2008. 2

[12] Y. Baocai, S. Yanfeng, W. Chengzhang, and G. Yun, "Bjut-3d large scale 3d face database and information processing," *Journal of Computer Research and Development*, vol. 6, p. 020, 2009. 2

[13] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3d face analysis," in *European Workshop on Biometrics and Identity Management*, 2008, pp. 47–56. 2

[14] W. Sankowski, P. S. Nowak, and P. Krotewicz, "Multimodal biometric database dmcsv1 of 3d face and hand scans," in *MIXDES*, 2015, pp. 93–97. 2

[15] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications," in *CVPR*, 2018, pp. 5117–5126. 2

[16] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014. 2

[17] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling," in *ICCV*, 2011, pp. 2296–2303. 2

[18] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, vol. 1, 2006, pp. 519–528. 2

[19] H. Zhu, Y. Nie, T. Yue, and X. Cao, "The role of prior in image based 3d modeling: a survey," *Frontiers of Computer Science*, vol. 11, no. 2, pp. 175–191, 2017. 2

[20] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, "Video-based outdoor human reconstruction," *TCSVT*, vol. 27, no. 4, pp. 760–770, 2016. 2

[21] V. Blanz, T. Vetter *et al.*, "A morphable model for the synthesis of 3d faces," in *Siggraph*, vol. 99, no. 1999, 1999, pp. 187–194. 2

[22] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," in *ToG*, vol. 24, no. 3, 2005, pp. 426–433. 2, 4

[23] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans," *ToG*, vol. 36, no. 6, p. 194, 2017. 2, 4

[24] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang, "Disentangled representation learning for 3d face shape," in *CVPR*, 2019, pp. 11 957–11 966. 2

[25] H. Li, T. Weise, and M. Pauly, "Example-based facial rigging," in *ToG*, vol. 29, no. 4, 2010, p. 32. 2, 4

[26] T. Bagautdinov, C. Wu, J. Saragih, P. Fua, and Y. Sheikh, "Modeling facial geometry using compositional vaes," in *CVPR*, 2018, pp. 3877–3886. 2

[27] 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 *CVPR*, 2018, pp. 2549–2559. 2

[28] L. Tran, F. Liu, and X. Liu, "Towards high-fidelity nonlinear 3d face morphable model," in *CVPR*, 2019, pp. 1126–1135. 2

[29] L. Tran and X. Liu, "On learning 3d face morphable model from in-the-wild images," *PAMI*, 2019. 2

[30] ——, "Nonlinear 3d face morphable model," in *CVPR*, 2018, pp. 7346–7355. 2

[31] S. Cheng, M. Bronstein, Y. Zhou, I. Kotsia, M. Pantic, and S. Zafeiriou, "Meshgan: Non-linear 3d morphable models of faces," *arXiv preprint arXiv:1903.10384*, 2019. 2

[32] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani *et al.*, "3d morphable face models—past, present, and future," *ToG*, vol. 39, no. 5, pp. 1–38, 2020. 2

[33] S. Romdhani and T. Vetter, "Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in *CVPR*, vol. 2, 2005, pp. 986–993. 2

[34] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR*, 2016, pp. 2387–2395. 2, 5

[35] B. Amberg, S. Romdhani, and T. Vetter, "Optimal step nonrigid icp algorithms for surface registration," in *CVPR*, 2007, pp. 1–8. 2, 3, 8

[36] P. Dou, S. K. Shah, and I. A. Kakadiaris, "End-to-end 3d face reconstruction with deep neural networks," in *CVPR*, 2017, pp. 5908–5917. 2

[37] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *CVPR Workshops*, 2019, pp. 0–0. 2, 10, 11

[38] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, and L. Quan, "Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency," in *ECCV*, 2020, pp. 53–70. 2, 10, 11

[39] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *ECCV*, 2020, pp. 152–168. 2, 10, 11

[40] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *CVPR*, 2019, pp. 7763–7772. 2, 10, 11

[41] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," in *CVPR*, 2018, pp. 8377–8386. 2

[42] T. Koizumi and W. A. Smith, ""look ma, no landmarks!"–unsupervised, model-based dense face alignment," in *ECCV*. Springer, 2020, pp. 690–706. 2

[43] F. Liu, R. Zhu, D. Zeng, Q. Zhao, and X. Liu, "Disentangling features in 3d face shapes for joint face reconstruction and recognition," in *CVPR*, 2018, pp. 5216–5225. 2

[44] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *CVPR*, 2019, pp. 1155–1164. 2

[45] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, and J. Feng, "3d face reconstruction from a single image assisted by 2d face images in the wild," *IEEE Transactions on Multimedia*, vol. 23, pp. 1160–1172, 2020. 2

[46] E. Richardson, M. Sela, R. Orel, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *CVPR*, 2017, pp. 5553–5562. 3

[47] M. Sela, E. Richardson, and R. Kimmel, "Unrestricted facial geometry reconstruction using image-to-image translation," in *ICCV*, 2017, pp. 1576–1585. 3

[48] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "Sfsnet: Learning shape, reflectance and illuminance of faces in the wild'," in *CVPR*, 2018, pp. 6296–6305. 3

[49] 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 *CVPR*, 2018, pp. 8407–8416. 3

[50] Y. Chen, F. Wu, Z. Wang, Y. Song, Y. Ling, and L. Bao, "Self-supervised learning of detailed 3d face reconstruction," *IEEE Transactions on Image Processing*, vol. 29, pp. 8696–8705, 2020. 3, 10, 11

[51] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu, "Photo-realistic facial details synthesis from single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9429–9439. 3, 7, 11

[52] A. T. Tran, T. Hassner, I. Masi, E. Paz, Y. Nirkin, and G. G. Medioni, "Extreme 3d face reconstruction: Seeing through occlusions," in *CVPR*, 2018, pp. 3935–3944. 3, 7, 10, 11

[53] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021. 3, 10, 11

[54] 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, pp. 534–551. 3, 10, 11

[55] Z. Ruan, C. Zou, L. Wu, G. Wu, and L. Wang, "Sadrnet: Self-aligned dual face regression networks for robust 3d dense face alignment and reconstruction," *IEEE Transactions on Image Processing*, 2021. 3, 10, 11

[56] X. Zeng, X. Peng, and Y. Qiao, "Df2net: A dense-fine-finer network for detailed 3d face reconstruction," in *ICCV*, 2019, pp. 2315–2324. 3, 10, 11

[57] Y. Zhou, J. Deng, I. Kotsia, and S. Zafeiriou, "Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders," in *CVPR*, 2019, pp. 1097–1106. 3

[58] X. Zhu, F. Yang, D. Huang, C. Yu, H. Wang, J. Guo, Z. Lei, and S. Z. Li, "Beyond 3dmm space: Towards fine-grained 3d face reconstruction," in *ECCV*, 2020, pp. 343–358. 3

[59] Z. Zhang, Y. Ge, R. Chen, Y. Tai, Y. Yan, J. Yang, C. Wang, J. Li, and F. Huang, "Learning to aggregate and personalize 3d face from in-the-wild photo collection," in *CVPR*, 2021, pp. 14 214–14 224. 3, 10, 11

[60] 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 *ICCV*, 2017, pp. 1031–1039. 3

[61] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," in *ICCV*, 2013, pp. 1385–1392. 3

[62] E. Friesen and P. Ekman, "Facial action coding system: a technique for the measurement of facial movement," *Palo Alto*, vol. 3, no. 2, p. 5, 1978. 4

[63] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, 2018, pp. 116–131. 5

[64] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[65] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018, pp. 8798–8807. 6

[66] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017, pp. 2794–2802. 6

[67] K. Nagano, J. Seo, J. Xing, L. Wei, Z. Li, S. Saito, A. Agarwal, J. Fursund, H. Li, R. Roberts *et al.*, "pagan: real-time avatars using dynamic textures," *ToG*, vol. 37, no. 6, pp. 258–1, 2018. 6

[68] T. Weise, S. Bouaziz, H. Li, and M. Pauly, "Realtime performance-based facial animation," in *ToG*, vol. 30, no. 4, 2011, p. 77. 6

[69] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives," *ToG*, vol. 32, no. 4, pp. 42–1, 2013. 6

[70] S. Bouaziz, Y. Wang, and M. Pauly, "Online modeling for realtime facial animation," *ToG*, vol. 32, no. 4, p. 40, 2013. 6

[71] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ToG*, vol. 32, no. 4, p. 41, 2013. 6

[72] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ToG*, vol. 33, no. 4, p. 43, 2014. 6

[73] B. Chaudhuri, N. Vesdapunt, and B. Wang, "Joint face detection and facial motion retargeting for multiple faces," in *CVPR*, 2019, pp. 9719–9728. 6

[74] L. Valgaerts, C. Wu, A. Bruhn, H.-P. Seidel, and C. Theobalt, "Lightweight binocular facial performance capture under uncontrolled lighting," *ToG*, vol. 31, no. 6, pp. 187–1, 2012. 8

[75] X. Zhu, Z. Lei, S. Z. Li *et al.*, "Face alignment in full pose range: A 3d total solution," *PAMI*, 2017. 7

[76] R. Ramamoorthi and P. Hanrahan, "A signal-processing framework for inverse rendering," in *Proceedings of Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 117–128. 9

[77] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *CVPR*, 2018. 9

[78] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *CVPR*, 2020, pp. 5549–5558. 9, 13

[79] P. M. R. Martin Koestinger, Paul Wohlhart and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 9

[80] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *ECCV*, 2018. 9

[81] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *CVPR*, 2020, pp. 5104–5113. 9

[82] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *CVPR*, 2019, pp. 7184–7193. 9

[83] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *CVPR*, 2020, pp. 601–610. 10, 11

[84] K. S. Arun, T. S. Huang, and S. D. Blostein, "Least-squares fitting of two 3-d point sets," *IEEE Transactions on pattern analysis and machine intelligence*, no. 5, pp. 698–700, 1987. 9

[85] Z.-H. Feng, P. Huber, J. Kittler, P. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Rätsch, "Evaluation of dense 3d reconstruction from 2d face images in the wild," in *FG*, 2018, pp. 780–786. 9, 11, 12, 13

[86] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017, pp. 1021–1030. 11

[87] S. Sanyal, T. Bolkart, H. Feng, and M. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *CVPR*, 2019, pp. 7763–7772. 11, 12, 13