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

1. Animation

We recommend watching the supplementary video, where the FaceScape dataset is briefly introduced and the generated animations are shown. In the animation part, the 3D face model is predicted from a single wild image, then is rigged to the expressions captured by FaceShift[7]. As shown in the video, the face model predicted by our method can be rigged to various expressions while recovers the dynamic details, such as the wrinkles caused by expressions. We also use the same rigging parameters to drive 3 different predicted models, and find that they appear different dynamic details. This is because these details are related to the source subjects, not the rigging parameters.

2. Model Processing Details

The generation of topologically uniformed model has been briefly introduced in Section 3.2 of the main paper. Here we supplement a detailed description of model registration and displacement map generation.

Registration of base shape. We down-sample the raw recovered mesh into rough mesh with fewer triangle faces, namely base shape, and then build 3DMM for these simplified meshes. Firstly, the 2D landmarks are extracted from the frontal image, then the corresponding 3D landmarks are obtained by inverse-projection 2D landmarks. The Procrustes transformation[3] is used to register all landmarks to a standard 3D facial template with landmark annotations. In this way, the pose and scale for all the scanned meshes are roughly aligned to the standard facial template. Then we use Non-rigid ICP[1] to register the standard template mesh to scanned mesh in neutral expression. For scanned meshes in other 19 expressions, similar to [2], the deformation transfer algorithm[5] is firstly used to deform the registered mesh in neutral expression to other expressions mimicking the deformation of a set of template meshes in corresponding expressions. Then the Non-rigid ICP[1] is used to register these deformed individual-specific templates to scanned meshes to fit the scans in non-neutral expressions more accurately.

Displacement map generation. 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. The most straightforward way to compute the displacement map is to calculate the distance from the surface of the registered model to the raw mesh. However, we find that there will be artifacts in the displacement map caused by the defects in the registration procedure. Thus the raw scan is firstly smoothed with Laplacian mesh smoothing. Then we trace the surface points of base mesh corresponding to pixels in the displacement map, and inverse-project the points to the raw mesh along normal direction to find its corresponding points. The pixel value of the displacement map is set to the signed distance from the point on raw mesh to its corresponding point on the smoothed mesh.

3. Base Model Fitting

The base model fitting method has been briefly introduced in Section 4.1 of the main paper. Here we provide a detailed description of three parts in the objective function.

Landmark Alignment. Firstly the 2D landmarks L are extracted from the image using an off-the-shelf facial landmark detector. Assuming the camera is weak perspective, the landmark alignment term is defined as the distance between the detected 2D landmark $L^{(k)}$ and its corresponding vertex projected on the image space:

$$E_{lan} = ||(sR(C_r \times \mathbf{w}_{exp} \times \mathbf{w}_{id})^{(k)} + \mathbf{t}) - L^{(k)}||_2^2 \quad (1)$$

where s is the scale factor of the weak perspective function, R is the rotation matrix and $\mathbf{t}$ is the translation.

Pixel Level Consistency. The pixel-level reconstruction term is used to match the geometry more accurately in the regions where no feature points such as cheeks exists. Under the assumption of Lambertian surfaces, we use the first three bands of Spherical Harmonics(SH)[4] for illumination representation. The per-vertex albedo is represented as a PCA model based on our dataset with albedo parameter

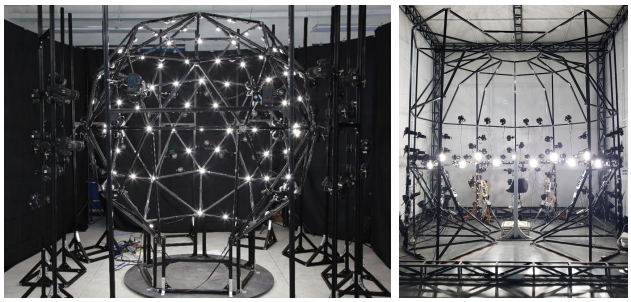

Figure 1: Our multi-view system to reconstruct the high quality detailed 3D face. We captured the data in two different places, so there are two frameworks shown as system A and system B.

$\mathbf{w}_{alb}$. The objective function is formulated as:

$$E_{pixel} = \frac{1}{|\mathcal{V}|}\sum_{q\in\mathcal{V}}||\hat{I}(q) - I(q)||_2 \quad (2)$$

where $\mathcal{V}$ is the set of pixels corresponding to frontal vertices of the fitted mesh, $\hat{I}$ is the synthetic face, I is the input image.

Regularization. We formulate the prior of identity, expression and albedo parameters as multivariate Gaussians around the average of our dataset for regularization. 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. We optimize the parameters alternatively. Following [8], the vertex indices corresponding to contour landmarks of the face are updated after each iteration.

4. Facial Capture System

The capturing system has been briefly introduced in Section 3.1 of the main paper. Here we supplement the pictures of our system in Figure 1. The system consists of the 68 DSLR camera array, controlled lighting and a centralized control sever.

5. More Results

More results are supplemented in Figure 5 as the extension of Figure 6 in our main paper. It shows that our results recover 3D faces with photo-realistic details. The faces can be further rigged to other expressions, and the details in the new expressions are synthesized to make the rigged model plausible.

Figure 2: We use our recovered model for synthesizing images in another expression with detailed shading.

Figure 3: Failure cases. In the left, our prediction cannot recover the aquiline nose well, as this feature is not common in our dataset. In the right, the wrong displacement map is predicted due to occlusion.

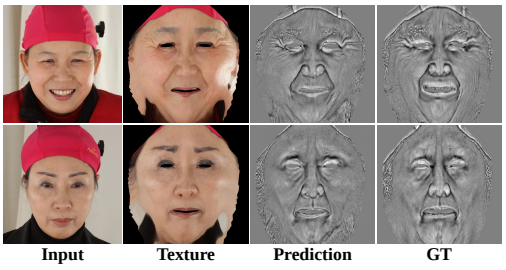

Figure 4: Predicted displacement maps using our method and ground truth.

We supplement the comparison of the predicted and ground-truth displacement maps in Figure 4 as the extension of Figure 7 in our main paper.

6. More Models

We show the 20 captured expressions for each subject in Figure 6, and show more subjects in neutral expression in Figure 7. The diversity of models in expression and identity dimensions ensures the quality of bilinear face model generated on FaceScape dataset.

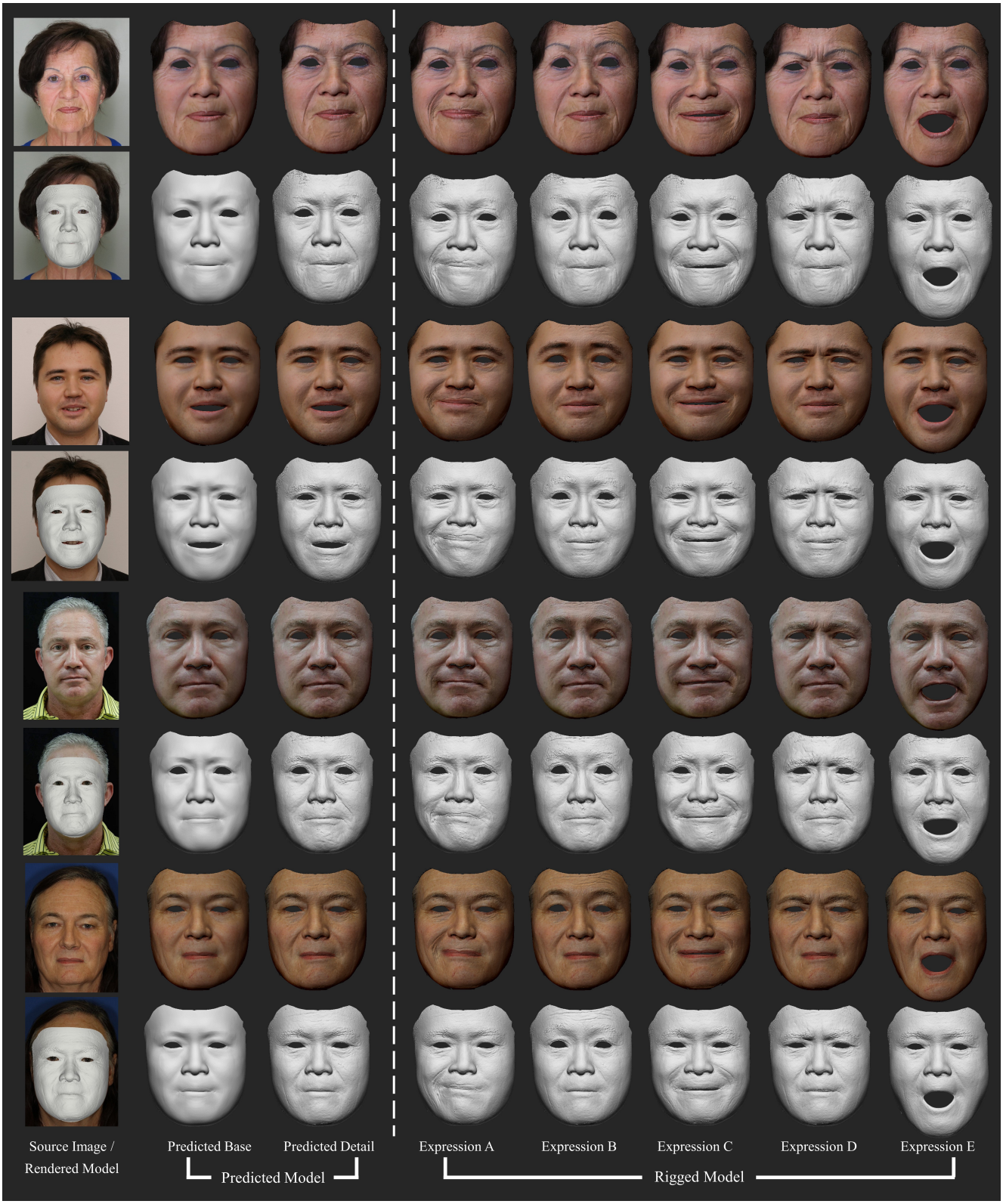

Figure 5: We show more results as the extension of Figure 6 in our main paper.

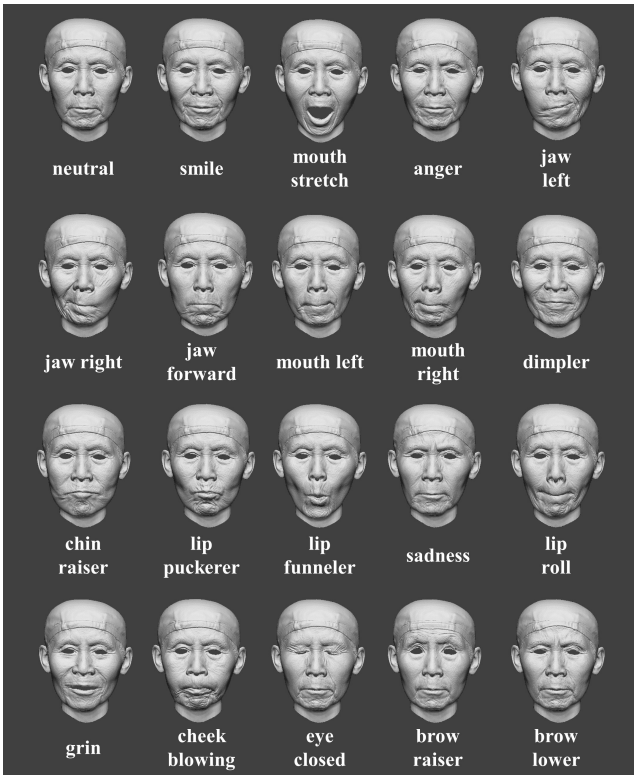

Figure 6: The 20 specified expressions which the subjects are asked to perform.

7. Photo-realistic Image Synthesis

Similar to [6, 2], given a facial image, our bilinear model can be used to synthesize images in other expressions. Specifically, we use the base model fitting method to estimate the face model. Then we change the expression parameter to generate the face model in the target expression and warp the image pixels guided by translations of vertices on the 3D face model. The details caused by the expression changing are further synthesized by adjusting the pixel shading. New pixel value is calculated based on the new normal from the predicted displacement map and estimated illumination in model fitting procedure. The synthesized images are shown in Figure 2.

8. Failure Case

We show some failure cases of our method in Figure 3.

References

[1] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *CVPR*, pages 1–8, 2007. 1

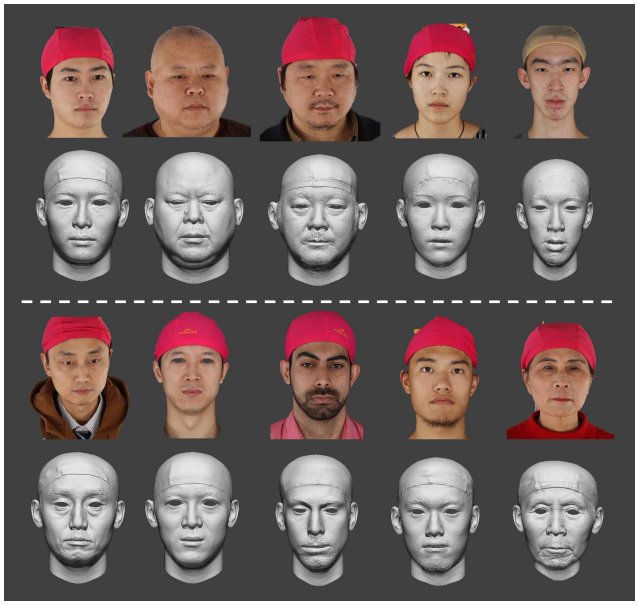

Figure 7: More models with different identities are shown in this figure. The upper part is the images of the subjects, and lower part is the processed topologically uniformed models.

[2] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *TVCG*, 20(3):413–425, 2013. 1, 4

[3] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 1

[4] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 1

[5] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ToG*, 23(3):399–405, 2004. 1

[6] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ToG*, volume 24, pages 426–433, 2005. 4

[7] Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. Realtime performance-based facial animation. In *ToG*, volume 30, page 77, 2011. 1

[8] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, pages 787–796, 2015. 2