Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set (*Supplementary Material*)

1. Outline

In this supplementary document, we provide more details of our approach and evaluation which are omitted in the main paper due to space limitation. We also show some additional results from our method. The remaining structure of this document is organized as follows:

- Section 2: More details of our face model.
- Section 3: Math derivations of our analytic image generation process.
- Section 4: More details of our training losses.
- Section 5: Details of the evaluation protocol on FaceWarehouse dataset [2] and more comparison with [9].
- Section 6: Detailed analysis on confidence scores.
- Section 7: More visual results on in-the-wild images.

2. 3D Face Model

In this work, a cropped Basel 2009 3D face model [8] is used throughout our experiments. As shown in Fig 1 (a/b), we cut the mesh of Basel 2009 model along the outer boundary points from 68 facial landmarks [5] to exclude ear and neck regions. Inner mouth region is also excluded, leading to a final mesh with 35,709 vertices.

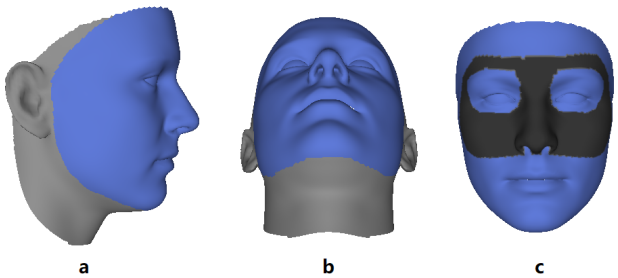

Figure 1. **(a/b)** The region (blue color) we use for 3D face reconstruction. **(c)** The region (dark color) we use for the texture flattening constraint.

3. Analytic Image Generation

Given a training RGB image I, we use R-Net to regress a coefficient vector $\mathbf{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p}) \in \mathbb{R}^{239}$ as described in the main paper. After getting $\mathbf{x}$, we calculate the image space position $\mathbf{u}_i(\mathbf{x})$, z-coordinate $z_i(\mathbf{x})$ (for occlusion handling via z-buffering) and color $\mathbf{c}_i(\mathbf{x})$ for each vertex $\mathbf{s}_i$ with surface normal $\mathbf{n}_i$ and texture $\mathbf{t}_i$[1]:

$$\begin{aligned} \mathbf{u}_i(\mathbf{x}) &= \Pi \circ \big(\mathbf{R}(\mathbf{p})\mathbf{s}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathbf{t}(\mathbf{p})\big) \\ z_i(\mathbf{x}) &= \big(\mathbf{R}(\mathbf{p})\mathbf{s}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathbf{t}(\mathbf{p})\big)_z \\ \mathbf{c}_i(\mathbf{x}) &= \mathbf{C}(\mathbf{R}(\mathbf{p})\mathbf{n}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}), \mathbf{t}_i(\boldsymbol{\delta})|\boldsymbol{\gamma}) \end{aligned} \tag{1}$$

where Π denotes the full perspective projection and $\mathbf{C}(\cdot)$ is the illumination model defined in the main paper. Then, similar to [4], we do a rasterization with $\mathbf{u}_i(\mathbf{x})$ and $\mathbf{z}_i(\mathbf{x})$ to get reconstructed image I'.

4. More Details of Our Losses

4.1. Skin Attention

In this work, we adopt a robustified photometric loss to train the network. As descried in the main paper, the photometric error is weighted by a skin attention mask derived from a skin probability map P for the pixels on I.

To obtain the skin probability P, we train a naive Bayes classifier with Gaussian Mixture Models (GMMs) on a skin image dataset from [6]. This dataset contains 4,671 human images with skin region labeled and 8,965 natural images without human. To obtain the likelihood functions for skin and non-skin color, we fit GMMs onto skin and non-skin pixels (YCbCr color space) in the training set respectively using the Expectation Maximization algorithm. Four Gaussian components are used for both. Consequently, the posterior skin probability for a new pixel can be computed with the prior probabilities and likelihood functions following

[1]Note the slight abuse of notation here: the vertex texture $\mathbf{t}_i$ with subscript i should not be confused with $\mathbf{t}$ which is the translation vector; similarly, the vertex color notation $\boldsymbol{c}_i$ should not be confused with the identity confidence vector $\boldsymbol{c}$ predicted by C-Net.

Figure 2. Examples of the skin probability maps used in our robust photometric loss. Note that these skin probability maps are used only for training; they are not required in the testing stage.

Bayes rule. Figure 2 presents some examples of the resultant skin color probability maps P for our training images.

4.2. Texture Flattening

As mentioned in the main paper, to favor constant skin albedo, we add a texture flattening constrain to penalize texture map variance over a pre-defined region $\mathcal{R}$: $L_{tex}(\mathbf{x}) = \sum_{c\in\{r,g,b\}} var(\mathbf{T}_{c,\mathcal{R}}(\mathbf{x}))$. The region we use is shown in Fig. 1 (c), which covers part of cheek, nose, and forehead. The texture flattening loss helps to remove shading from generated face texture. Some examples of our generated texture are shown in Fig. 5.

5. Evaluation on FaceWarehouse

Evaluation Protocol. To conduct a fair comparison with [9, 10, 7, 3] on the Facewarehouse dataset [2], we use the evaluation protocal of [9]. Specifically, the topology of reconstruction meshes is first transferred to the one defined by [9], which contains 60K vertices evenly distributed across the whole head, via non-rigid registration. The ground truth meshes of Facewarehouse [2] is also subdivided to obtain a denser topology. Then, with a point-to-point correspondence pre-computed by [9], a 3D similarity transformation is applied to align reconstructions with ground truth. Finally, a mean closest point error is calculated as the geometry error and reported. All the numerical results on FaceWarehouse presented in the main paper follows the same evaluation protocol on the same 9 subjects selected by [9].

More Result Comparison. Figure 3 shows the 9 FaceWarehouse subjects used for evaluation. Compared to [9], our shape reconstructions are more faithful and exhibit higher variance across different people. Besides, the recovered texture from [9] seem to be over-smooth and contain obvious shading components whereas ours better represent the raw skin reflectance.

6. Detailed Analysis on Confidence Scores

In this section, we expand on the analysis of confidence score statistics in the main paper (Section 6.2.1). In this analysis, we collect one frontal and one profile face image for each of the 53 subjects on the MICC dataset [1] and compute their confidence vectors via C-Net. Then we calculate the average relative confidence vectors for frontal and

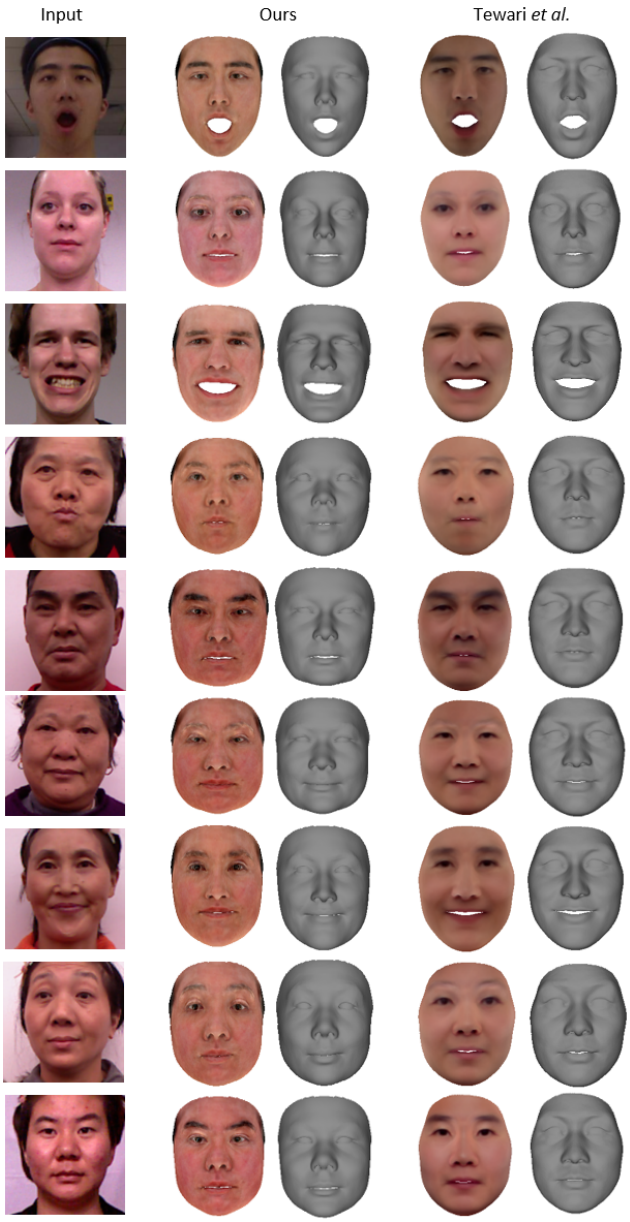

Figure 3. Comparison with [9] (fine results) on the 9 subjects in Facewarehouse [2]. Our shape reconstructions are more faithful and exhibit higher variance across different people. Besides, the recovered textures from [9] seem to be over-smooth and contain obvious shading components, whereas ours better represent the raw skin reflectance.

profile faces as:

$$\begin{aligned} \boldsymbol{c}^{frontal} &= \tfrac{1}{53}\textstyle\sum_{j=1}^{53}(\boldsymbol{c}^{j,frontal} \oslash \boldsymbol{c}^{j,all}) \\ \boldsymbol{c}^{profile} &= \tfrac{1}{53}\textstyle\sum_{j=1}^{53}(\boldsymbol{c}^{j,profile} \oslash \boldsymbol{c}^{j,all}) \end{aligned} \tag{2}$$

where $\oslash$ denotes Hadamard division and $\boldsymbol{c}^{j,all} = \boldsymbol{c}^{j,frontal} + \boldsymbol{c}^{j,profile}$. Figure 4 (left) shows the first 20 entries of $\boldsymbol{c}^{frontal}$ and $\boldsymbol{c}^{profile}$ with largest PCA energy

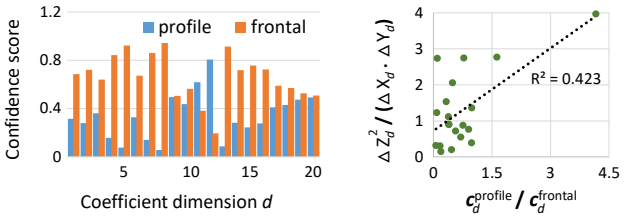

Figure 4. Confidence statistics on frontal and profile images. We show the first 20 entries with largest PCA energy (standard derivation). Left: average relative confidence scores of 53 subjects. Right: Z-direction shape influence *w.r.t.* profile-to-front confidence ratio. Each dot represents a coefficient vector entry. Entries having larger influence on face depth (Z-direction) tend to get relatively larger confidence scores on profile faces than on frontal ones (linear regression $R^2 = 0.423$).

(according to identity basis $\mathbf{B}_{id}$). As can be seen, the confidence scores for profile faces are lower on the majority of the entries, but higher on a few entries such as the 11th and 12th ones.

We further analyze the influence of each entry d on the face depth or Z-direction 3D face component (controlling nose height *etc.*). For each of them, we compute an indicator $\Delta Z_d^2/(\Delta X_d \cdot \Delta Y_d)$ defined as:

$$\Delta Z_d^2/(\Delta X_d \cdot \Delta Y_d) = \frac{\| \mathbf{B}_{id}^{d,z} \|_1^2}{\| \mathbf{B}_{id}^{d,x} \|_1 \| \mathbf{B}_{id}^{d,y} \|_1} \qquad (3)$$

where $\mathbf{B}_{id}$ is the identity basis matrix of 3DMM, $\mathbf{B}_{id}^{d,z}$ denotes the vector formed by the Z-coordinates of the vertices in the d-th basis (similarly for $\mathbf{B}_{id}^{d,x}$ and $\mathbf{B}_{id}^{d,y}$), and $\| \cdot \|_1$ denotes the l_1 norm. This indicator represents a basis's influence on Z-components of a face relative to X- and Y-components. A larger value indicates the corresponding identity coefficient has more contribution on Z-direction deformation and vise versa. As shown in Fig. 4 (right), we evaluate the correlations between profile-to-frontal confidence ratio $\boldsymbol{c}_d^{profile}/\boldsymbol{c}_d^{frontal}$ and the above indicator. We conduct linear regression and get $R^2 = 0.423$, indicating that coefficient entries having larger influence on face depth (Z-direction) tend to have larger profile-to-frontal confidence ratio (*i.e.*, getting relatively larger confidence scores on profile faces than on frontal ones). This is consistent with our intuition and suggests that our network learns to exploit information from different views for better reconstruction.

7. More Visual Results

7.1. Single Image Reconstruction

Here, we show more visual results of our method on in-the-wild images. Figure 5 shows that our method achieves high quality reconstructions across *different races and ages*. Figure 6 demonstrates the robustness of our method under *challenging conditions* including large occlusions, heavy make ups, large poses, and extreme expressions.

7.2. Multi Image Aggregation

We further show more aggregation results based on image sets. As shown in Figure 7, our method is able to produce quality results for unconstrained image sets. Note these unconstrained images may contain challenging pose, lighting and visibility conditions. Our confidence aggregation strategy allows the network to favor high quality images and fuse the complementary information to achieve accurate reconstruction.

References

[1] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In *The Joint ACM Workshop on Human Gesture and Behavior Understanding*, pages 79–80, 2011. 2

[2] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2014. 1, 2

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

[4] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing (IVC)*, 28(5):807–813, 2010. 1

[6] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision (IJCV)*, 46(1):81–96, 2002. 1

[7] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4625–4634, 2018. 2

[8] 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 (AVSS)*, pages 296–301, 2009. 1

[9] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 1, 2

[10] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. MoFa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017. 2

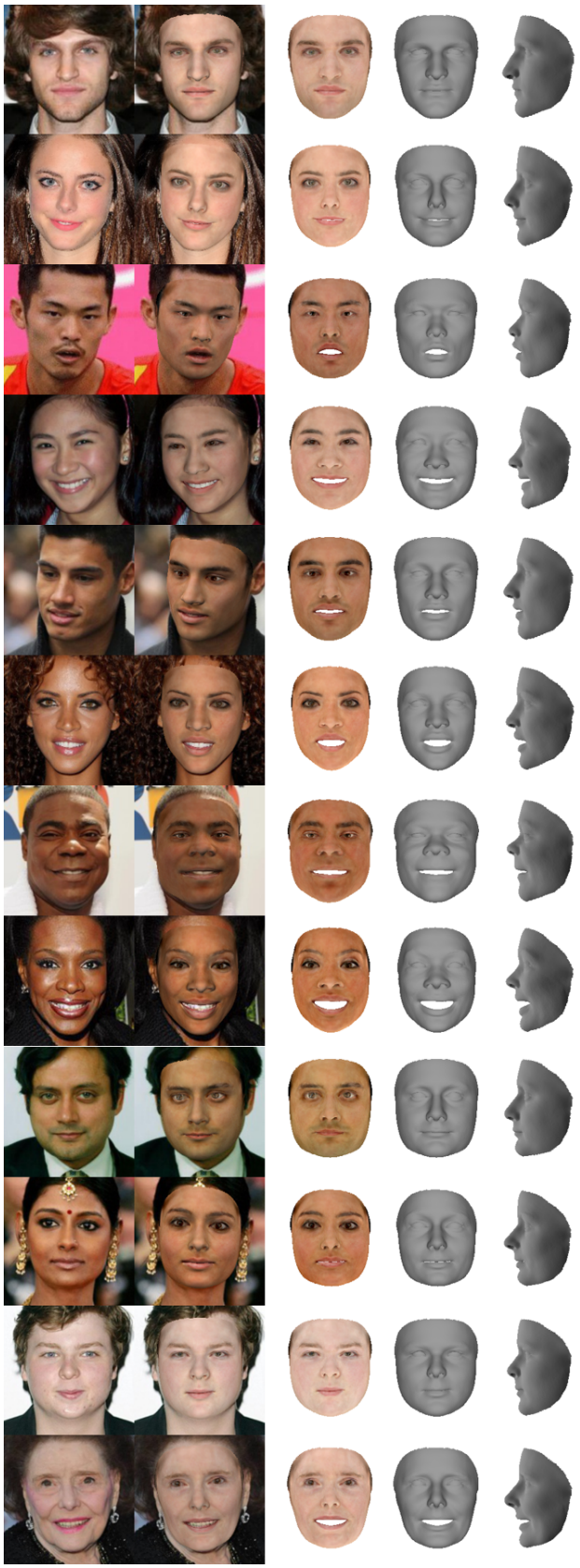

Figure 5. Results across different races and ages.

Figure 6. Results under challenging conditions such as occlusions, large poses and extreme expressions.

Figure 7. Results on in-the-wild image sets. The leftmost bar charts shows the sorted value of confidence vector summation of each image in the set. (Note in practice we use element-wise coefficient aggregation; to ease presentation, confidence vector summation is shown here.) Five images sampled from a set are shown in the following five columns, with their confidence vector summations displayed in top left and the reconstructed images shown below. The last column shows our aggregated results. Textures are simply averaged across each set.