Supplementary Material for MMFace: A Multi-Metric Regression Network for Unconstrained Face Reconstruction

Hongwei Yi[1]*, Chen Li[2]†, Qiong Cao[2], Xiaoyong Shen[2], Sheng Li[3]†, Guoping Wang[3], Yu-Wing Tai[2]
[1]Shenzhen Graduate School, Peking Univ. [2]Tencent
[3]School of Electronics Engineering and Computer Science, Peking Univ.
{hongweiyi, lisheng, wgp}@pku.edu.cn {chaselli, freyaqcao, dylanshen, yuwingtai}@tencent.com

Table 1. The specifications of the convolutional layers used in the parametric sub-network.

3D Conv		3dConv1	3dConv2	3dConv3	3dConv4	3dConv5
	Input	$1 \times 64 \times 64 \times 64$	$64 \times 32 \times 32 \times 32$	$128 \times 16 \times 16 \times 16$	$256 \times 8 \times 8 \times 8$	$512 \times 4 \times 4 \times 4$
	Output	$64 \times 32 \times 32 \times 32$	$128 \times 16 \times 16 \times 16$	$256 \times 8 \times 8 \times 8$	$512 \times 4 \times 4 \times 4$	$1024 \times \times 1 \times 1$
	Stride, Pad	2, 1	2, 1	2, 1	2, 1	2, 0
	Filter	$4 \times 4 \times 4$	$4 \times 4 \times 4$	$4 \times 4 \times 4$	$4 \times 4 \times 4$	$4 \times 4 \times 4$
FC		$FC^1_{id/exp/p}$	FC^2_{id}	FC^2_{exp}	FC^2_{p}	FC^3_{id}
	Input	1024	512	512	512	256
	Output	512	256	29	7	199

1. Network Architecture

Our method takes a facial image (cropped and scaled to 256×256) as input and estimates the corresponding 3D volume $\mathbb{V}$ and the 3DMM parameters $\mathbf{p} = [\mathbf{f}, \mathbf{r}, \mathbf{t}, \alpha_{\mathbf{id}}, \alpha_{\mathbf{exp}}]^{\mathbf{T}}$. It consists of two cascade sub-networks, namely the volumetric sub-network **VMN** and the parametric sub-network **PMN**.

The **VMN** has the same architecture with **VRN** [1] except an additional upsampling layer to regress the 3D volume in a different resolution. Two identical "hourglass" modules are stacked together to extract a $64\times64\times64$ feature map from the input image. In order to regress the 3D volumetric $\mathbb{V}$ in different resolutions, we first extend the channel of this feature map to the target resolution r as $64 \times 64 \times r$ and then employ one 2D upsample layer to estimate $\mathbb{V}$ in the $r\times r\times r$ resolution. Specifically, we use $r = \{64, 128, 192\}$ in our implementation.

The **PMN** takes the $64 \times 64 \times 64$ feature map of the **VMN** as input and estimates the corresponding 3DMM parameters $\mathbf{p}$. The specifications of the convolutional layers used in the **PMN** is illustrated in Table. 1.

2. VMN Evaluation and Results

Our multi-metric regression network not only estimates accurate 3DMM parameters, but also improves the intermediate volumetric geometry in turn by incorporating the parametric loss. A quantitative evaluation with **VRN** [1] by counting the voxel mismatching percentage between reconstructed volume and ground truth is listed in Table. 2. Some qualitative comparison are also shown in Fig. 1. Because **VRN** [1] has already estimated very accurate volumetric geometry, visually improvements such as producing more details or reducing artifacts change the quantitative evaluation little and Table. 2 demonstrates our result is slightly better than **VMN**. However, it is clear to see our results in Fig. 1 are visually more pleasing. The detailed structures around nose and mouth are reconstructed better in our results.

Table 2. The quantitative evaluation of volumetric regression.

Method	**AFLW2000-3D** [2]
VRN [1]	3.39%
MMFace-ICP-192	**3.13%**
MMFace-ICP-128	3.26%
MMFace-ICP-64	3.47%

3. Additional Results

We present more results of static images on **AFLW2000-3D** [2] in Fig. 2-Fig. 4.

[1]This work was done while Hongwei Yi was an intern at Tencent.
[2]Chen Li and Sheng Li are the joint corresponding authors.

References

[1] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, 2017. 1, 2

[2] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. In *CVPR*, 2016. 1, 2, 3, 4

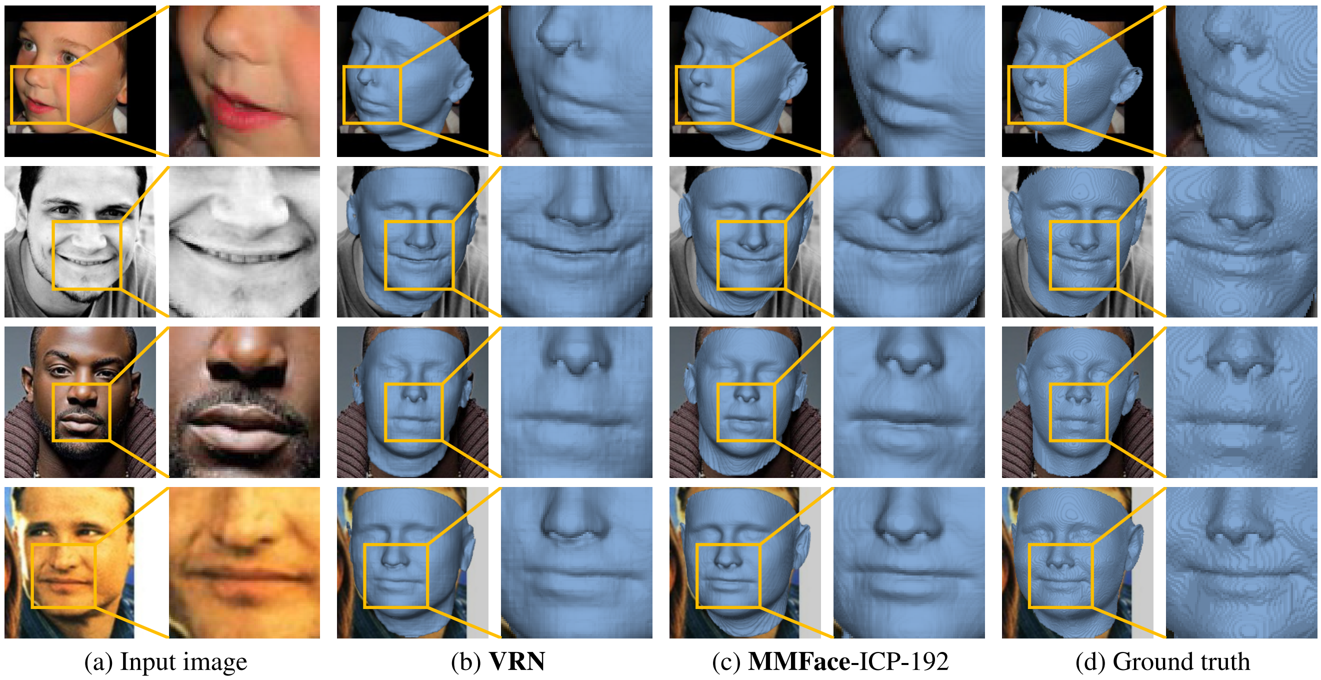

Figure 1. Comparison with **VRN** [1]. (a) The input image. (b-d) The result and close-up views of **VRN**, our **MMFace**-ICP-192 and the ground truth. Close-up views for better visualization are aligned right to their corresponding results.

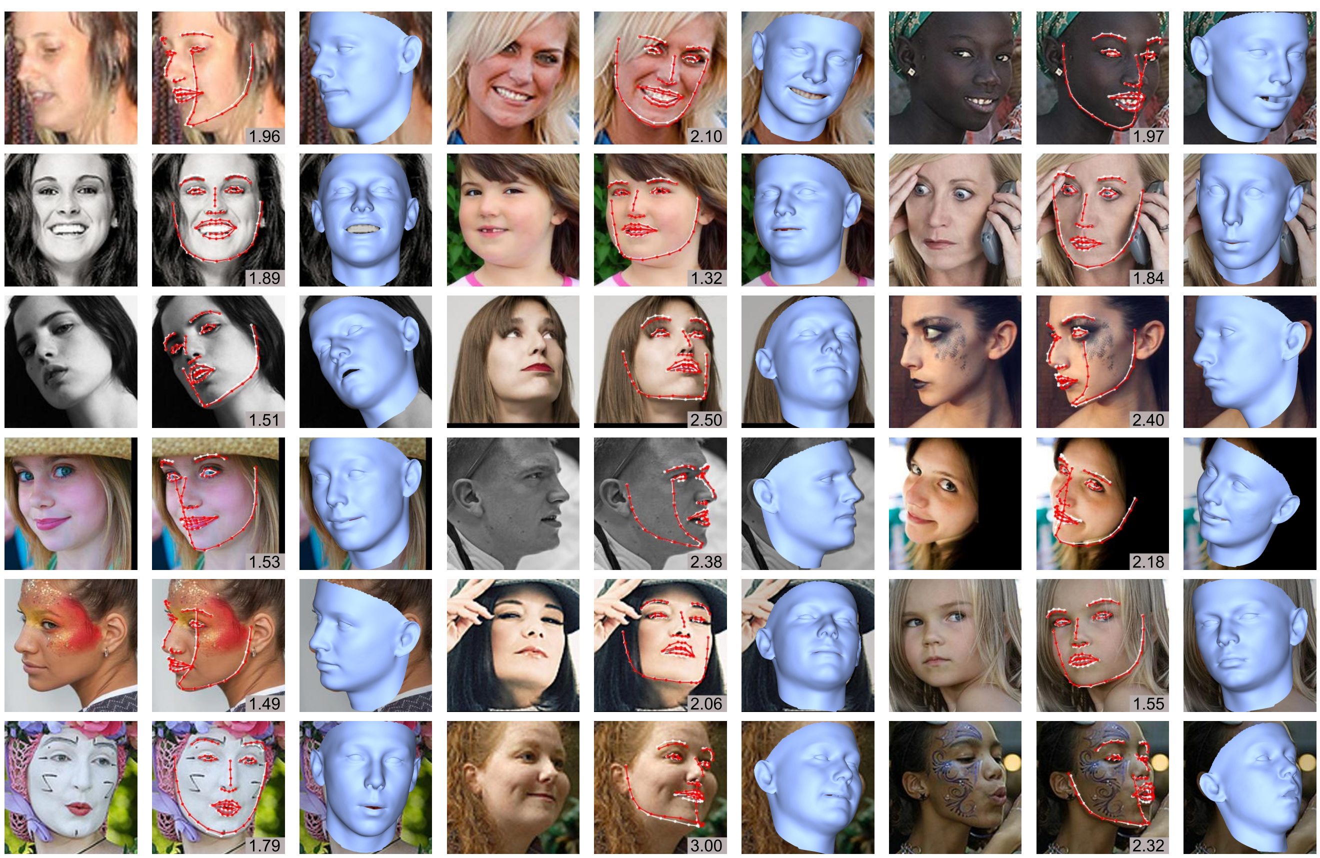

Figure 2. Face reconstruction and alignment results of our method on **AFLW2000-3D** [2].

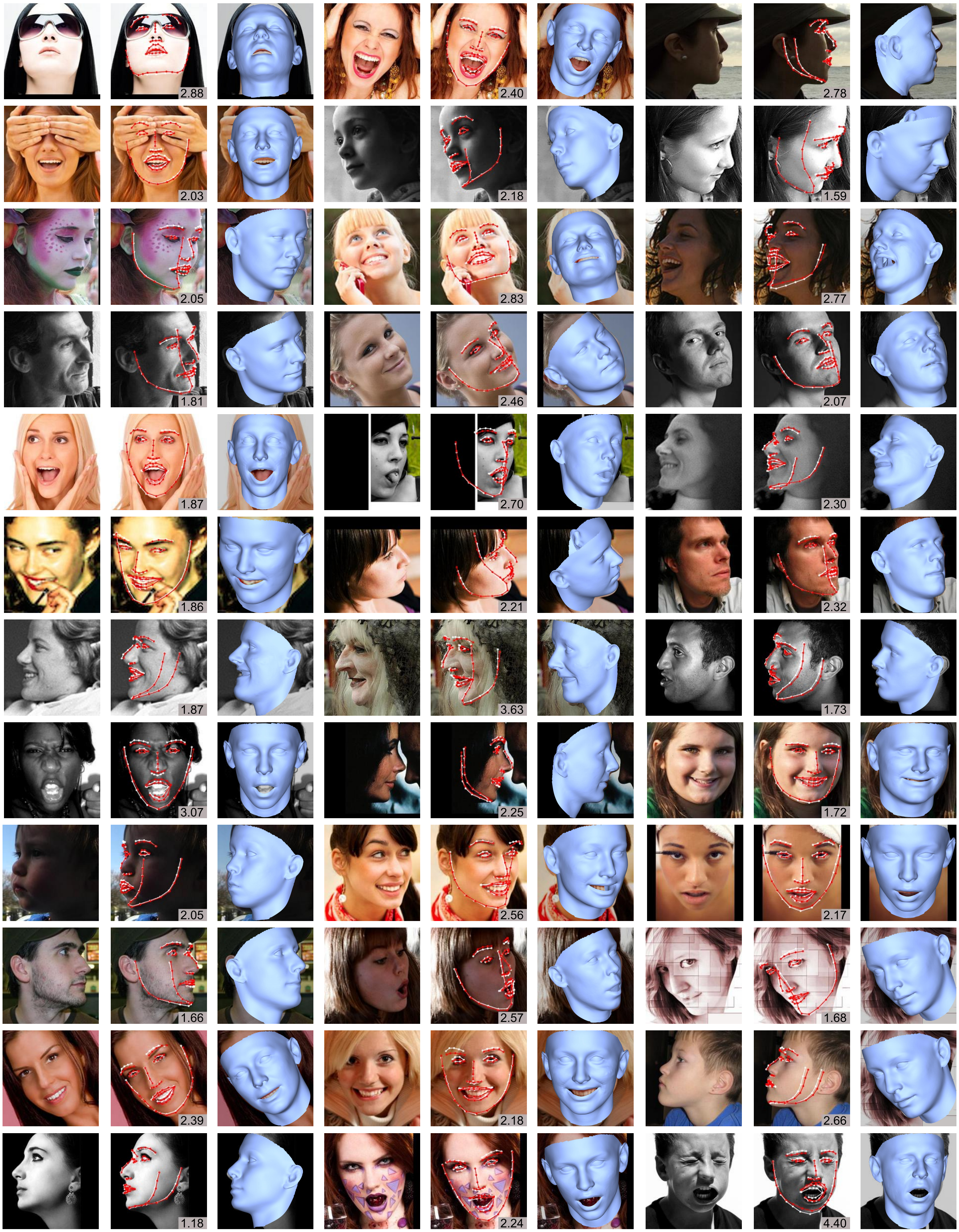

Figure 3. Face reconstruction and alignment results of our method on **AFLW2000-3D** [2].

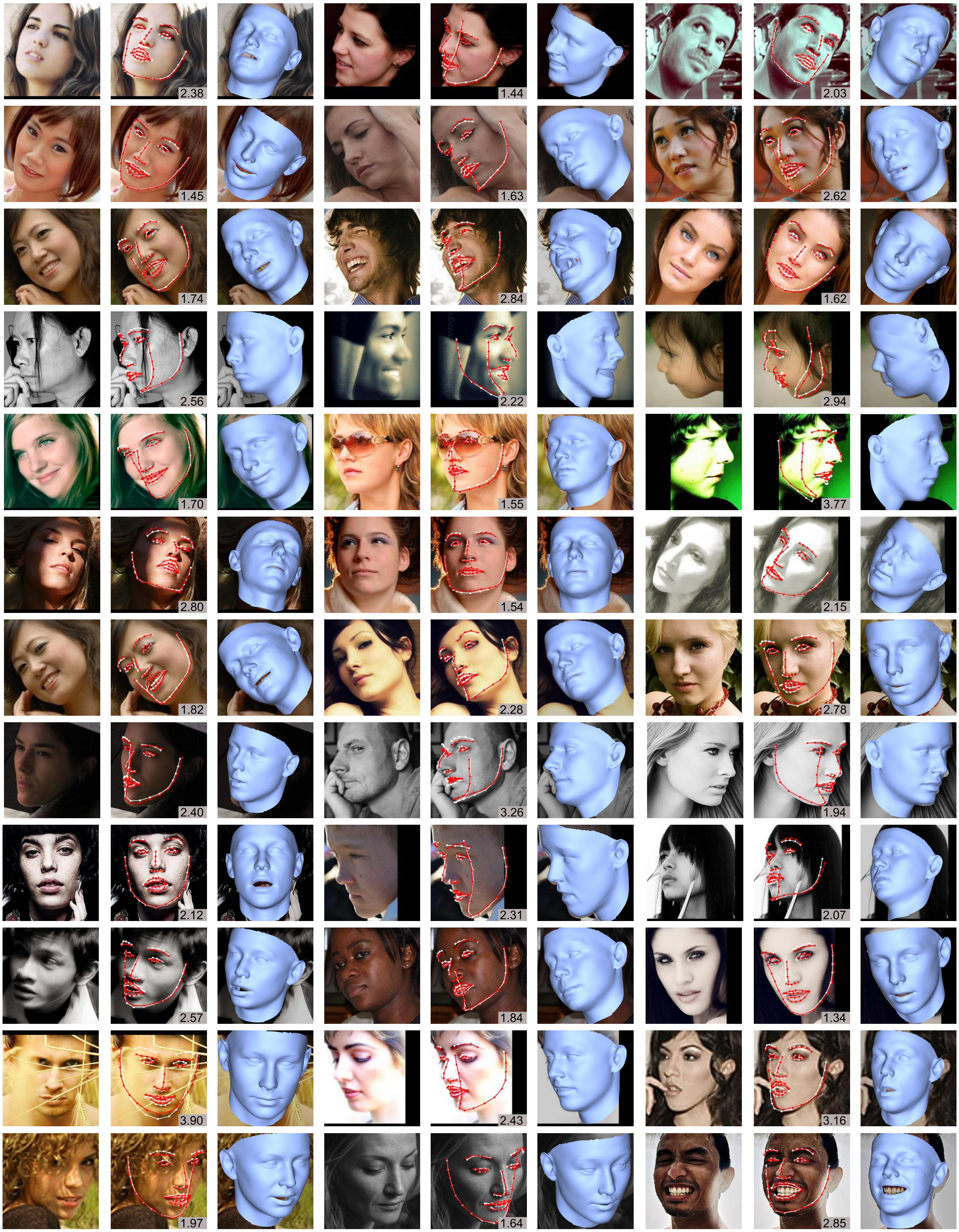

Figure 4. Face reconstruction and alignment results of our method on **AFLW2000-3D** [2].