Supplemental: Learning an Animatable Detailed 3D Face Model from In-The-Wild Images

YAO FENG[*], Max Planck Institute for Intelligent Systems and Max Planck ETH Center for Learning System, Germany
HAIWEN FENG[*], Max Planck Institute for Intelligent Systems, Germany
MICHAEL J. BLACK, Max Planck Institute for Intelligent Systems, Germany
TIMO BOLKART, Max Planck Institute for Intelligent Systems, Germany

CCS Concepts: • **Computing methodologies** → *Mesh models.*

Additional Key Words and Phrases: Detailed face model, 3D face reconstruction, facial animation, detail disentanglement

ACM Reference Format:
Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Supplemental: Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Trans. Graph.* 40, 4, Article 88 (August 2021), 10 pages. https://doi.org/10.1145/3450626.3459936

1 OVERVIEW

The supplemental material for our paper includes this document and a video. The video provides an illustrated summary of the method as well as animation examples. Here we provide implementation details and an extended qualitative evaluation.

2 IMPLEMENTATION DETAILS

Data: DECA is trained on 2 Million images from VGGFace2 [Cao et al. 2018], BUPT-Balancedface [Wang et al. 2019] and VoxCeleb2 [Chung et al. 2018]. From VGGFace2 [Cao et al. 2018], we randomly select $950k$ images such that $750K$ images are of resolution higher than 224×224, and $200K$ are of lower resolution. From BUPT-Balancedface [Wang et al. 2019] we randomly sample $550k$ with Asian or African ethnicity labels to reduce the ethnicity bias of VGGFace2. From VoxCeleb2 [Chung et al. 2018] we choose $500k$ frames, with multiple samples from the same video clip per subject to obtain data with variation only in the facial expression and head pose. We also sample $50k$ images from the VGGFace2 [Cao et al. 2018] test set for validation.

Data cleaning: We generate a different crop for the face image by shifting the provided bounding box by 5% to the bottom right (i.e. shift by $\boldsymbol{\epsilon} = \frac{1}{20}(b_w, b_h)^T$, where b_w and b_h denote the bounding box width and height). Then we expand the original and the shifted bounding boxes by 10% to the top, and by 20% to the left, right, and bottom. We run FAN [Bulat and Tzimiropoulos 2017], providing the expanded bounding boxes as input and discard all images with $\max_i \left\| \mathbf{D}(\mathbf{k}_i^2 - \boldsymbol{\epsilon} - \mathbf{k}_i^1) \right\| \geq 0.1$, where $\mathbf{k}_i^2$ and $\mathbf{k}_i^1$ are the ith landmarks for the original and the shifted bounding box, respectively, and $\mathbf{D}$ denote the normalization matrix $diag(b_w, b_h)^{-1}$.

Training details: We pre-train the coarse model (i.e. E_c) for two epochs with a batch size of 64 with $\lambda_{lmk} = 1e-4$, $\lambda_{eye} = 1.0$, $\lambda_{\boldsymbol{\beta}} = 1e-4$, and $\lambda_{\boldsymbol{\psi}} = 1e-4$. Then, we train the coarse model for 1.5 epochs with a batch size of 32, with 4 images per subject with $\lambda_{pho} = 2.0$, $\lambda_{id} = 0.2$, $\lambda_{sc} = 1.0$, $\lambda_{lmk} = 1.0$, $\lambda_{eye} = 1.0$, $\lambda_{\boldsymbol{\beta}} = 1e-4$, and $\lambda_{\boldsymbol{\psi}} = 1e-4$. The landmark loss uses different weights for individual landmarks, the mouth corners and the nose tip landmarks are weighted by a factor of 3, other mouth and nose landmarks with a factor of 1.5, and all remaining landmarks have a weight of 1.0. This is followed by training the detail model (i.e. E_d and F_d) on VGGFace2 and VoxCeleb2 with a batch size of 6, with 3 images per subject, and parameters $\lambda_{phoD} = 2.0$, $\lambda_{mrf} = 5e-2$, $\lambda_{sym} = 5e-3$, $\lambda_{dc} = 1.0$, and $\lambda_{regD} = 5e-3$. The coarse model is fixed while training the detail model.

3 EVALUATION

3.1 Qualitative comparisons

Figure 2 shows additional qualitative comparisons to existing coarse and detail reconstruction methods. DECA better reconstructs the overall face shape than all existing methods, it reconstructs more details than existing coarse reconstruction methods (e.g. (b), (e), (f)), and it is more robust to occlusions compared with existing detail reconstruction methods (e.g. (c), (d), (g)).

As promised in the main paper (e.g. Section 6.1), we show results for more than 200 randomly selected ALFW2000 [Zhu et al. 2015] samples in Figures 3, 4, 5, 6, 7, 8, and 9. For each sample, we compare DECA's detail reconstruction (e) with the state-of-the-art coarse reconstruction method 3DDFA-V2 [Guo et al. 2020] (see (b)) and existing detail reconstruction methods, namely FaceScape [Yang et al. 2020] (see (c)), and Extreme3D [Tran et al. 2018] (see (e)). In total, DECA reconstructs more details then 3DDFA-V2, and it is more robust to occlusions than FaceScape and Extreme3D. Further, the DECA retargeting results appear realistic.

REFERENCES

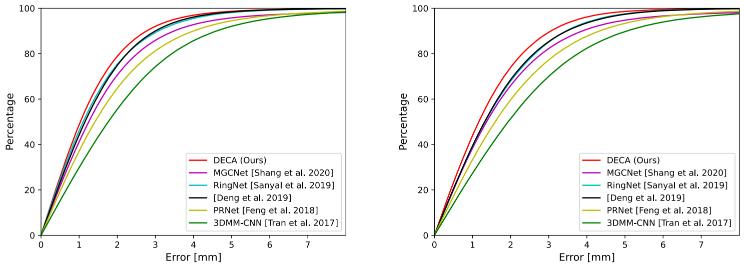
Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. 2020. Cross-modal Deep Face Normals with Deactivable Skip Connections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4979–4989.

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In

[*] Both authors contributed equally to the paper

Authors' addresses: Yao Feng, Max Planck Institute for Intelligent Systems, Tübingen, Max Planck ETH Center for Learning System, Tübingen, Germany, yfeng@tuebingen.mpg.de; Haiwen Feng, Max Planck Institute for Intelligent Systems, Tübingen, Germany, hfeng@tuebingen.mpg.de; Michael J. Black, Max Planck Institute for Intelligent Systems, Tübingen, Germany, black@tuebingen.mpg.de; Timo Bolkart, Max Planck Institute for Intelligent Systems, Tübingen, Germany, tbolkart@tuebingen.mpg.de.

0730-0301/2021/8-ART88
https://doi.org/10.1145/3450626.3459936



NoW (female) [Sanyal et al. 2019]

NoW (male) [Sanyal et al. 2019]

Fig. 1. Quantitative comparison to state-of-the-art on the NoW [Sanyal et al. 2019] challenge for female (left) and male (samples).

IEEE International Conference on Computer Vision (ICCV). 1021–1030.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition (FG)*. 67–74.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *INTERSPEECH*.

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *Computer Vision and Pattern Recognition Workshops*. 285–295.

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *European Conference on Computer Vision (ECCV)*. 534–551.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *European Conference on Computer Vision (ECCV)*. 152–168.

Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2011. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops (ICCV-W)*. 2144–2151.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 2019. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7763–7772.

Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. 2018. Extreme 3D face reconstruction: Seeing through occlusions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3935–3944.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In *IEEE International Conference on Computer Vision (ICCV)*.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 601–610.

Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. 2015. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 787–796.

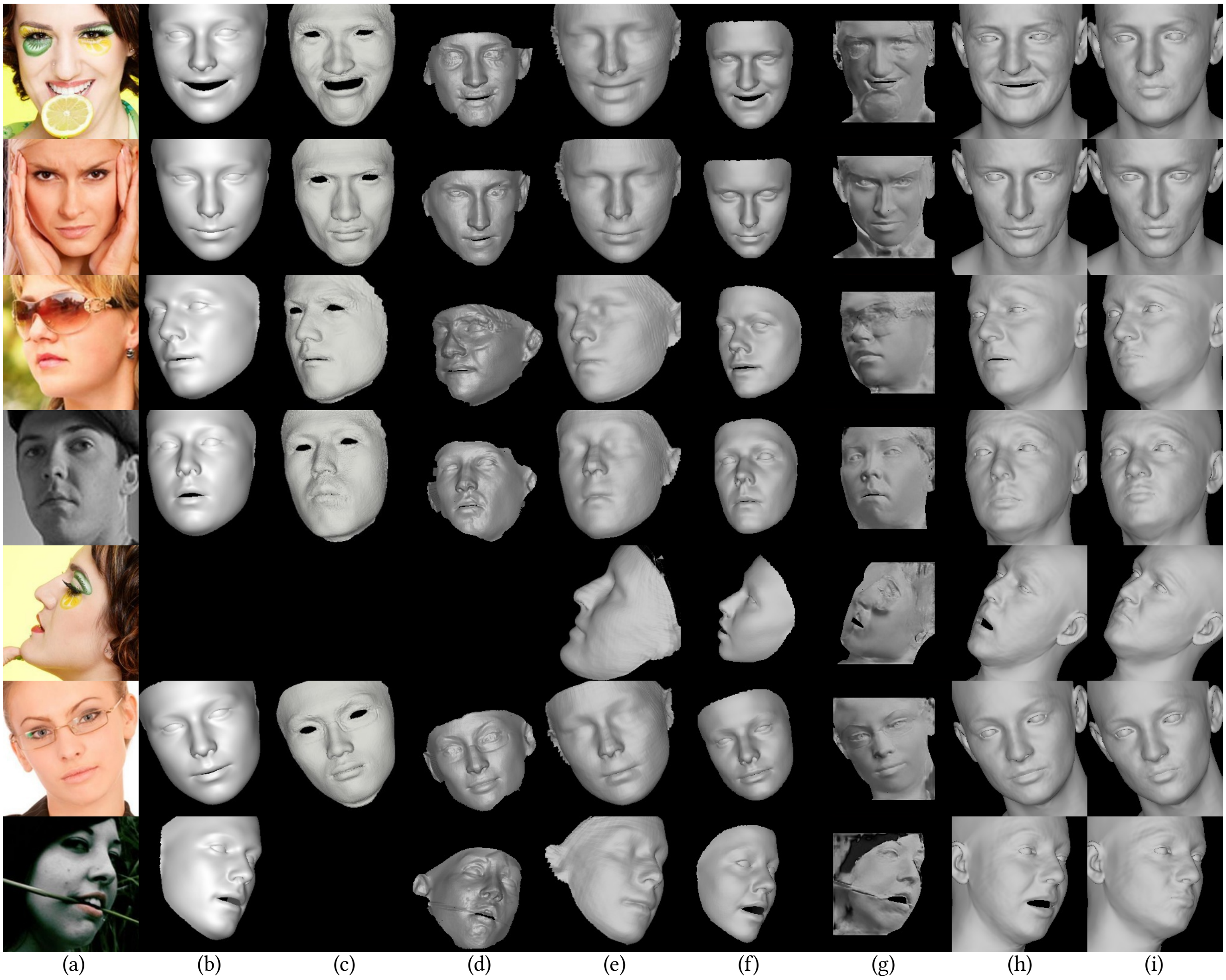


Fig. 2. Comparison to previous work, from left to right: (a) Input image, (b) 3DDFA-V2 [Guo et al. 2020], (c) FaceScape [Yang et al. 2020], (d) Extreme3D [Tran et al. 2018], (e) PRNet [Feng et al. 2018], (f) Deng et al. [2019], (g) Cross-modal [Abrevaya et al. 2020], (h) DECA detail reconstruction, and (i) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction. Input images are taken from ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015].

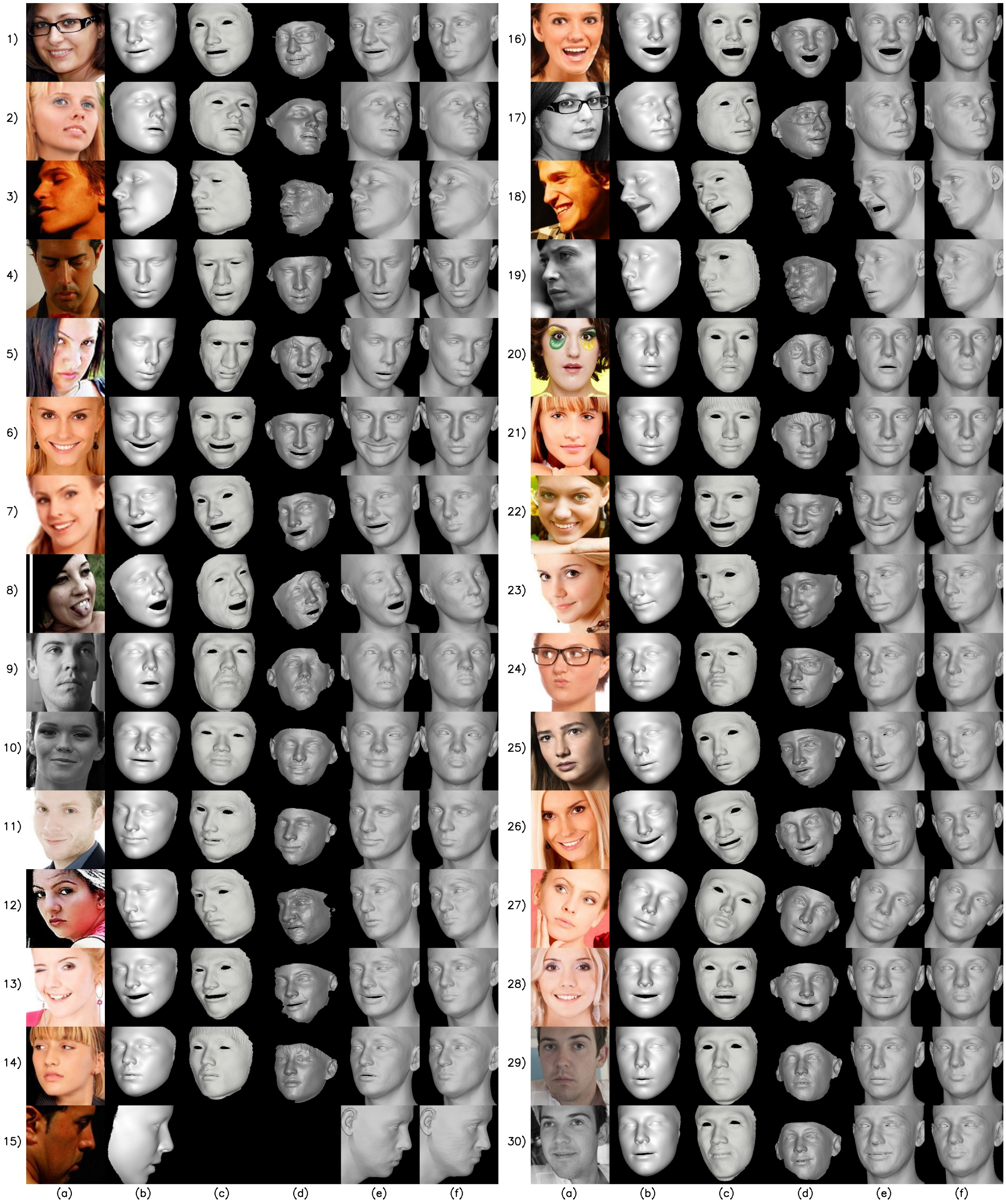


Fig. 3. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

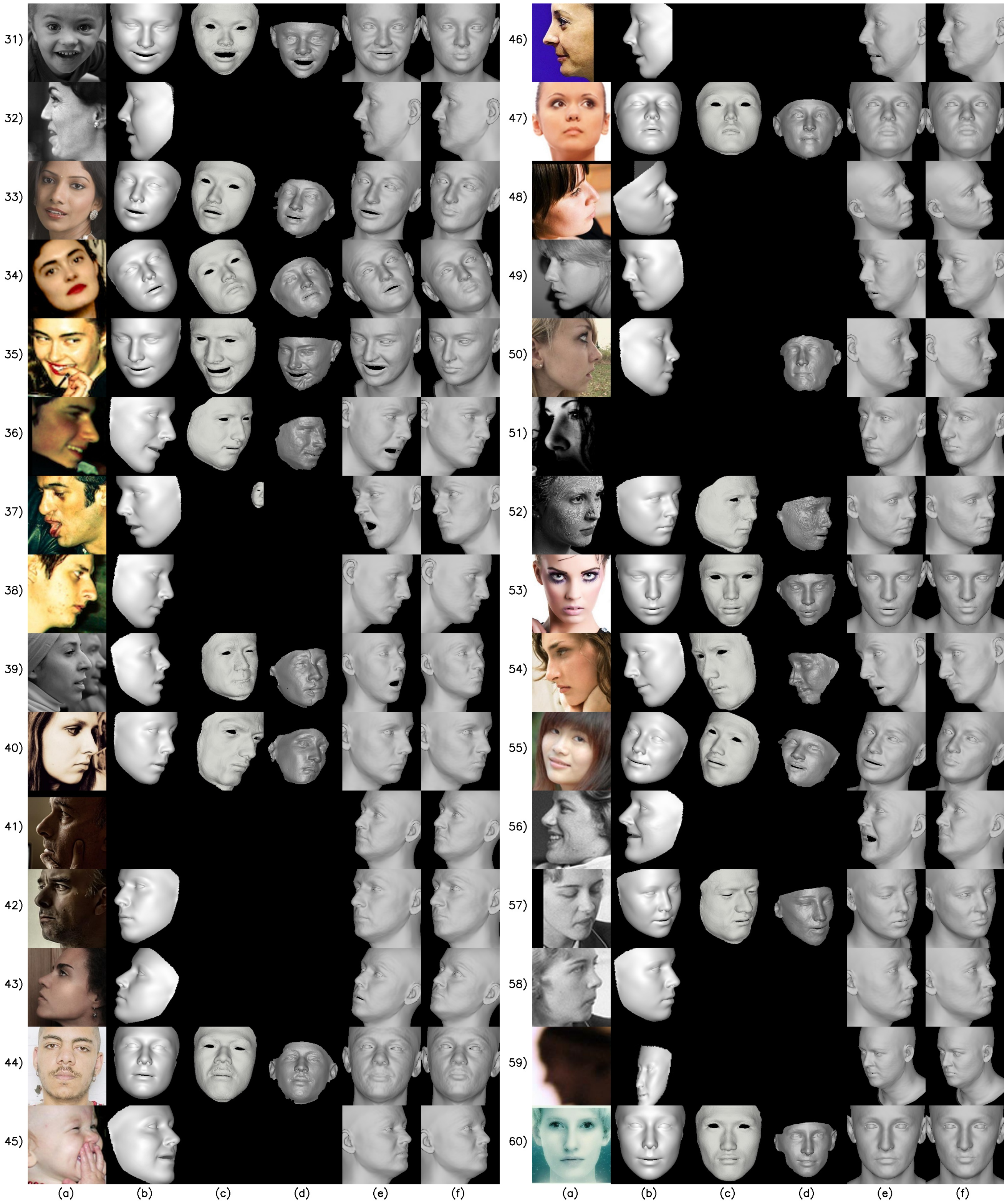


Fig. 4. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

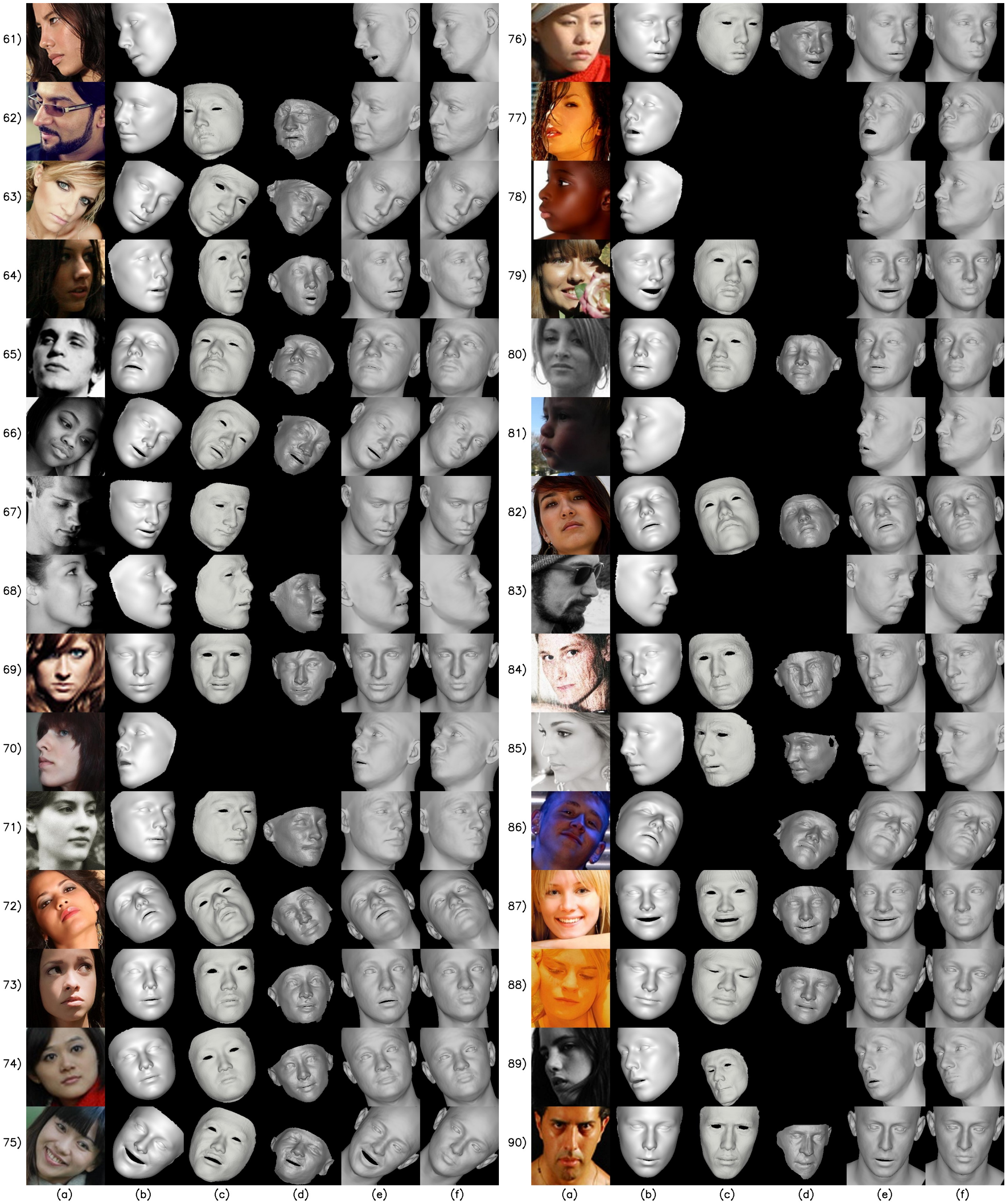


Fig. 5. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

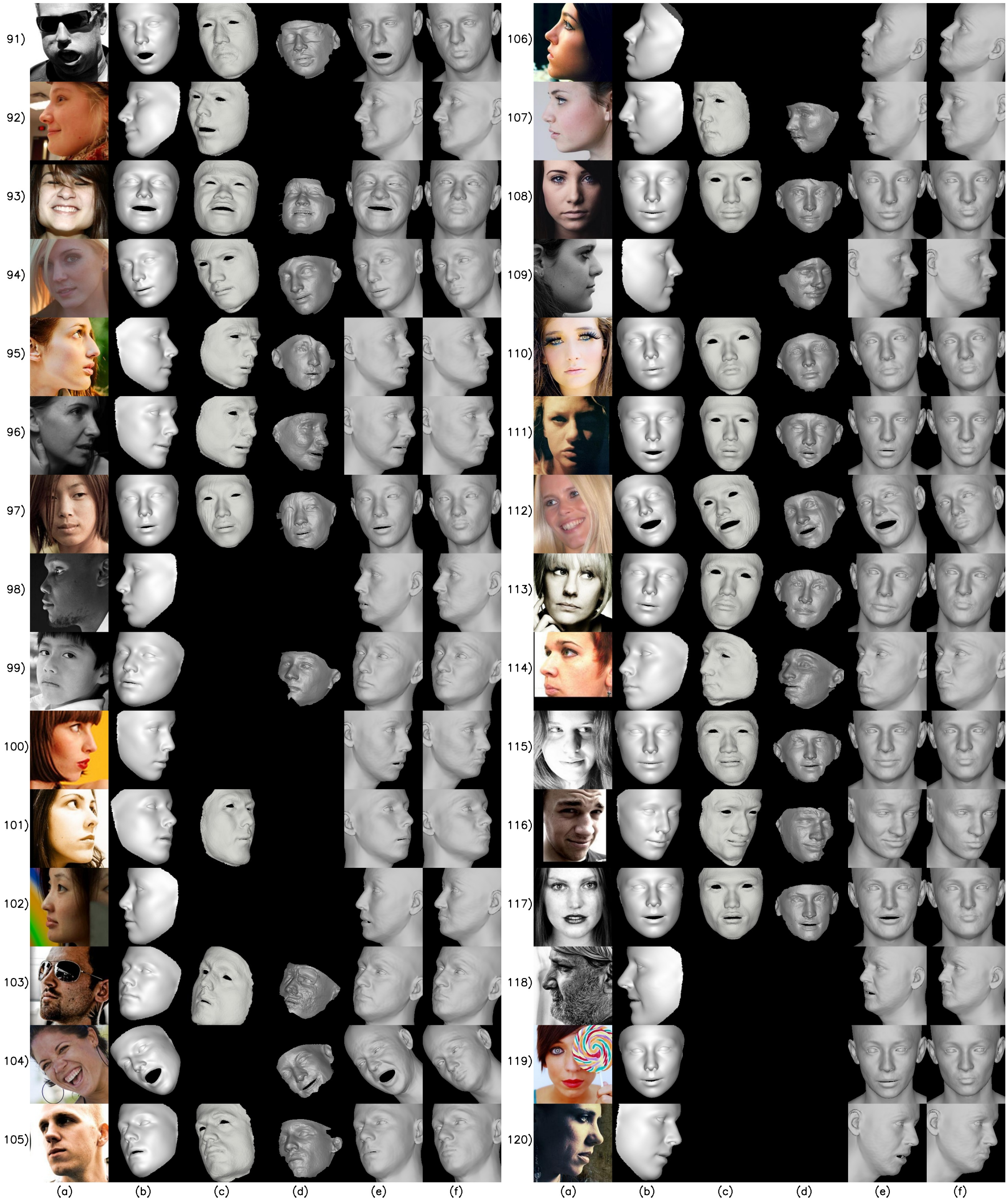


Fig. 6. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

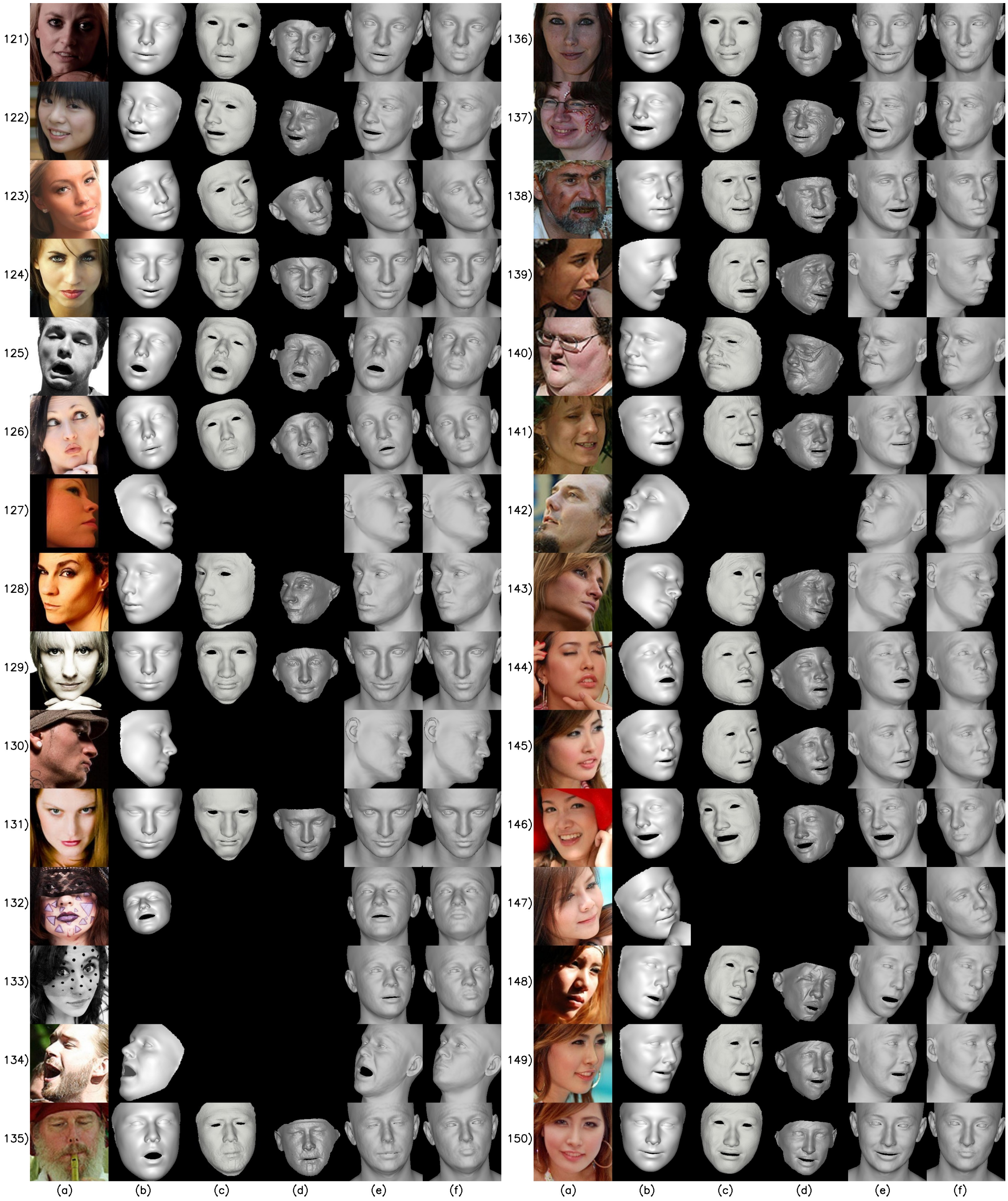


Fig. 7. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

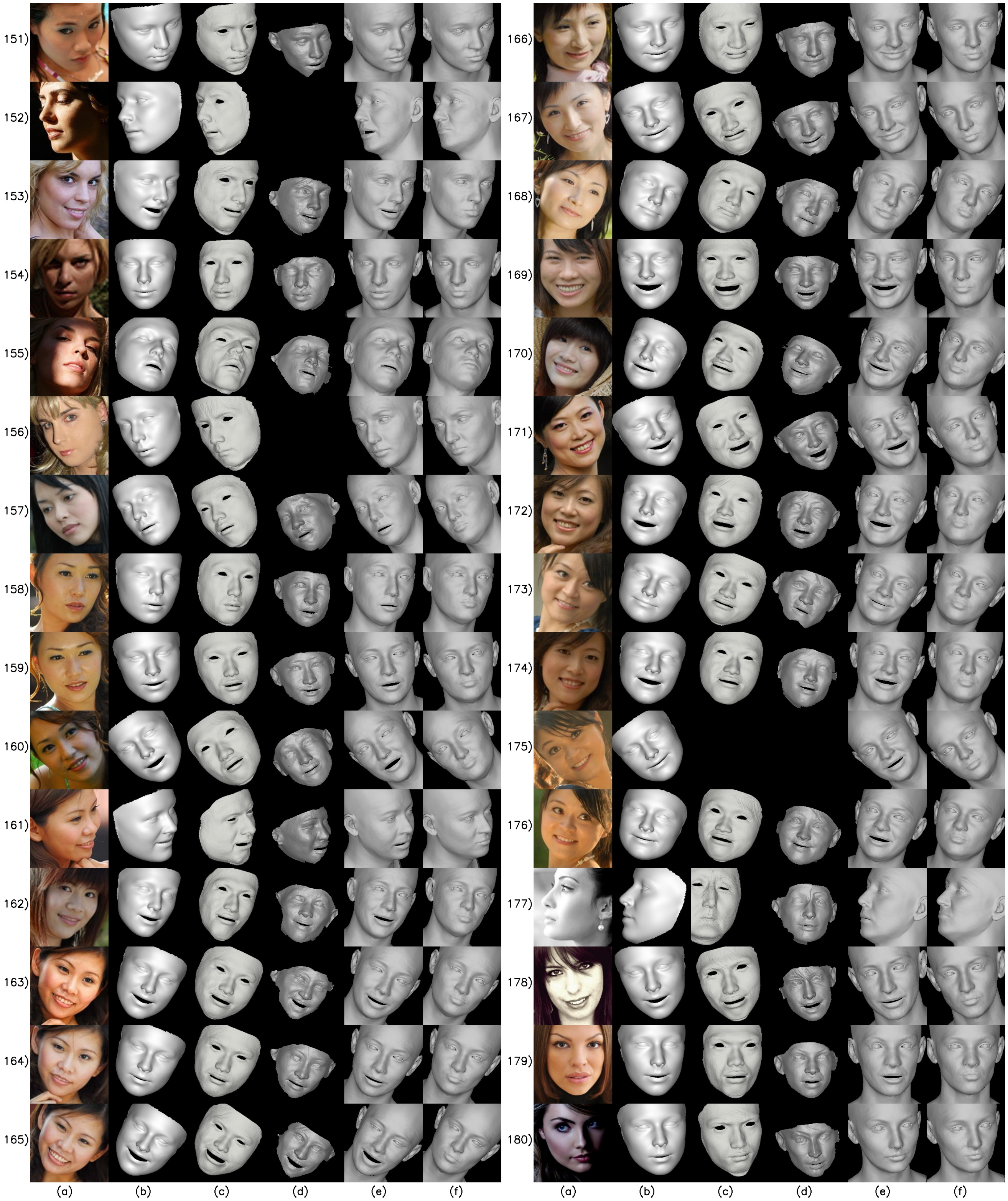


Fig. 8. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.

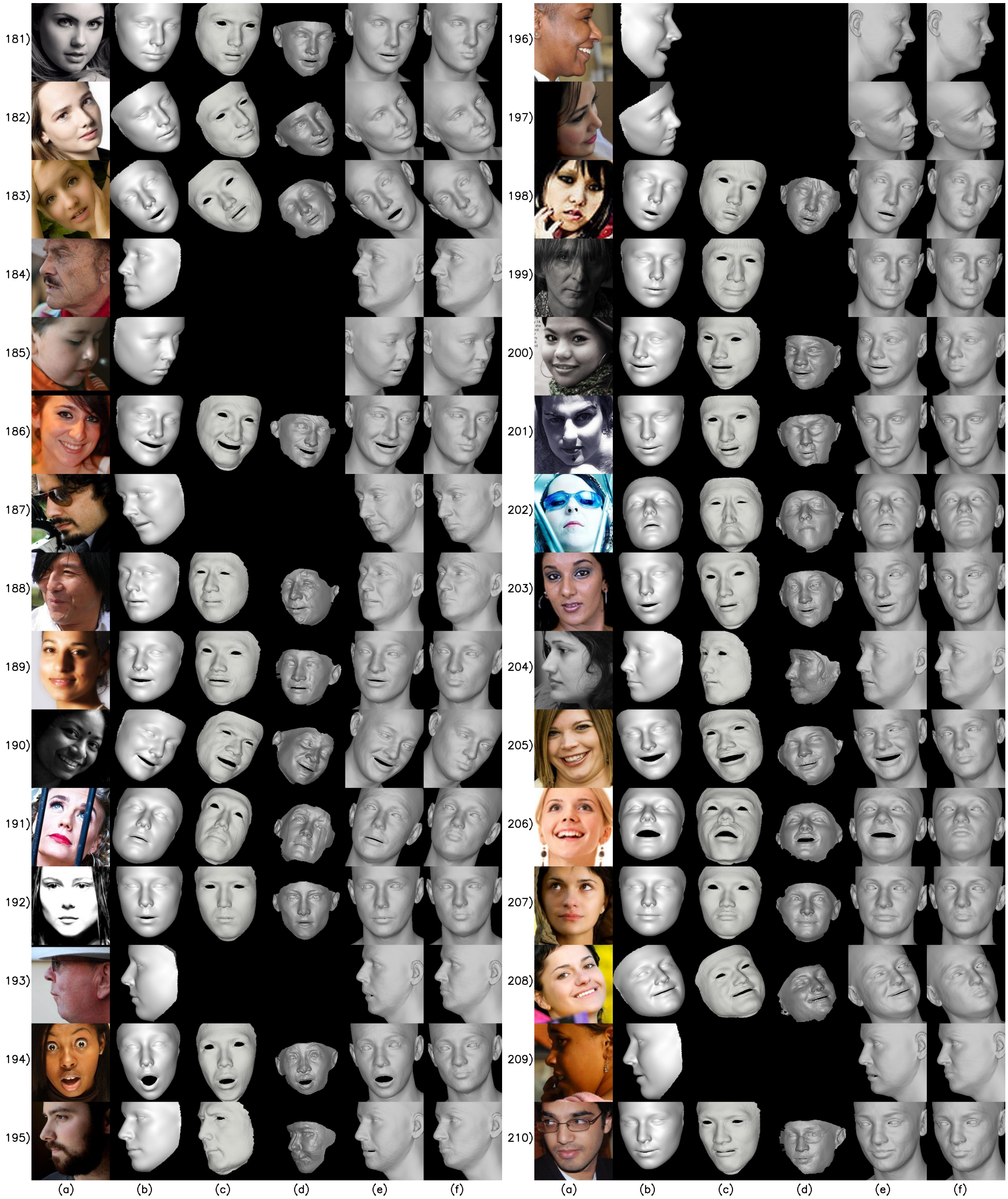


Fig. 9. Qualitative comparisons on random ALFW2000 [Köstinger et al. 2011; Zhu et al. 2015] samples. a) Input, b) 3DDFA-V2 [Guo et al. 2020], c) FaceScape [Yang et al. 2020], d) Extreme3D [Tran et al. 2018], e) DECA detail reconstruction, and f) reposing (animation) of DECA's detail reconstruction to a common expression. Blank entries indicate that the particular method did not return any reconstruction.