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.

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 1/20(bw, bh)^T, where bw and bh 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 \([D(k_i^j − e − k_i^e)]\) ≥ 0.1, where ki^j and ki^e are the ith landmarks for the original and the shifted bounding box, respectively, and D denote the normalization matrix diag(bw, bh)^−1.

Training details: We pre-train the coarse model (i.e. E_c) for two epochs with a batch size of 64 with λ_mnfk = 1e−4, λ_ye = 1.0, λ_g = 1e−4, and λ_ye = 1e−4. Then, we train the coarse model for 1.5 epochs with a batch size of 32, with 4 images per subject with λ_θ = 2.0, λ_id = 0.2, λ_ye = 1.0, λ_mnfk = 1.0, λ_ye = 1.0, λ_g = 1e−4, and λ_ye = 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 λ_θ = 2.0, λ_mnf = 5e−2, λ_ye = 5e−3, λ_id = 1.0, and λ_ye = 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 than 3DDFA-V2, and it is more robust to occlusions than FaceScape and Extreme3D. Further, the DECA retargeting results appear realistic.

REFERENCES

Victoria Fernández Abreavay, Adnan Boukhayma, Philip HS Torr, and Edmond Boyer. 2020. Cross-modal Deep Face Normals with Deactivable Skip Connections. In Proceedings of the 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 Proceedings of the IEEE International Conference on Computer Vision (ICCV). 1021–1030.

ACM Trans. Graph., Vol. 40, No. 4, Article 88. Publication date: August 2021.
Fig. 1. Quantitative comparison to state-of-the-art on the NoW [Sanyal et al. 2019] challenge for female (left) and male (samples).
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. The expression in (i) is from the source expression E in Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstructed mesh.
Fig. 3. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.
Fig. 4. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.
Fig. 5. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.

ACM Trans. Graph., Vol. 40, No. 4, Article 88. Publication date: August 2021.
Fig. 6. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.
Fig. 7. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.
Fig. 8. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.
Fig. 9. Qualitative comparisons on random ALFW2000 [Zhu et al. 2015] samples. a) Input image, 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. The
reposing in (f) uses the source expression E from Figure 2 of the main paper. Blank entries indicate that the particular method did not return any reconstruction.

ACM Trans. Graph., Vol. 40, No. 4, Article 88. Publication date: August 2021.