Supplemental Material for Normalized Avatar Synthesis Using StyleGAN and Perceptual Refinement

Huiwen Luo Koki Nagano Han-Wei Kung Qingguo Xu
Zejian Wang Lingyu Wei Liwen Hu Hao Li
Pinscreen

Appendix I. Additional Comparisons

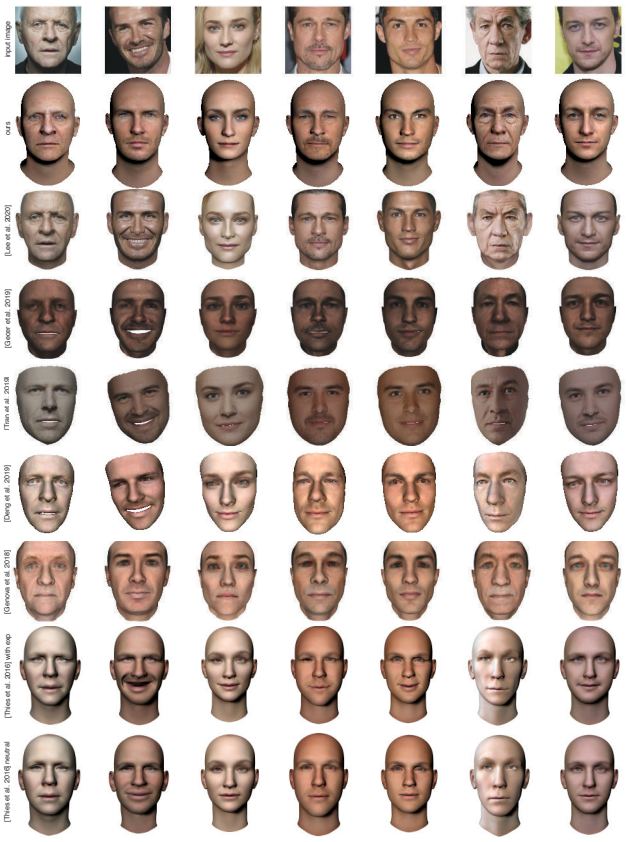

Figure 1: Additional comparisons. The first row shows the input images and the second row our results. The remaining rows are the reconstructed 3D faces obtained by [5, 3, 7, 2, 4, 6], respectively.

In Fig. 1, we compare our method with several recent state-of-the-art single view face reconstruction approaches. Thies et al. [6] extend the seminal work of Blanz and Vetter [1] with facial expression blendshapes and iteratively optimize for shape, texture, and lighting condition by minimizing energy terms based on facial landmark and pixel color constraints. We visualize the avatars with and without the facial expressions of the corresponding input photo. Neutralizing facial expressions is straightforward by setting all the blendshape coefficients to 0. We notice that the linear morphable face model is unable to recover features such as facial hair, as well as high-frequency geometry and appearance details. As a result, the face renderings often lack the likeness of the original subject and often fall within the so called "uncanny valley". Genova et al. [4] predict identity coefficients of linear 3DMM using a deep neural network and Deng et al. [2] predict the lights and face poses simultaneously using additionally linear 3DMM coefficients. Their models are still restricted to the linear subspace which has limited capabilities for representing facial details. Gecer et al. [3] introduce an unsupervised training approach to regress linear 3DMM coefficients for geometry and adopt a Generative Adversarial Network model for generating nonlinear texture. Tran et al. [7] present an approach to learn additional proxies as means to avoid strong regularization, which efficiently captures high level details for geometry and texture with a simple decoder architecture. They do not separate identity and expressions in the training. Lee et al. [5] demonstrate the latest work for generating 3D face models from a single input photograph using non-linear 3DMMs and an uncertainty-aware mesh decoder. The resulting 3D faces are very faithful to the input image, but the lighting and expressions are baked into the texture and mesh. As a result, neither Lee et al. [5] nor the above non-linear 3DMM techniques produce normalized results as shown in our paper. Notice that the results in Fig. 1 from row 3 to row 7 were taken directly from the paper of [5], and the renderings may have slight inconsistencies.

Appendix II. Additional Evaluations

In Sec. 3.1, we adopt a two step training method by first training G and then freezing G in order to compute the code inversion and to train R. Fig. 2 shows that the latent codes can be effectively found out with our choice of loss function in Eq. 2. Specifically, while pixel loss and adversarial

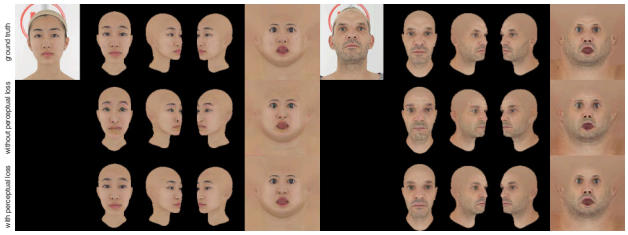

Figure 2: Algorithmic choice justification on the loss function for GAN-inversion. From top to bottom: Ground truth geometry and texture; Reconstruction results optimized by pixel loss and adversarial loss; Reconstruction results with perceptual loss in addition.

loss cannot preserve the overall similarity, adding the perceptual loss improves the high-level appearance in the rendering views.

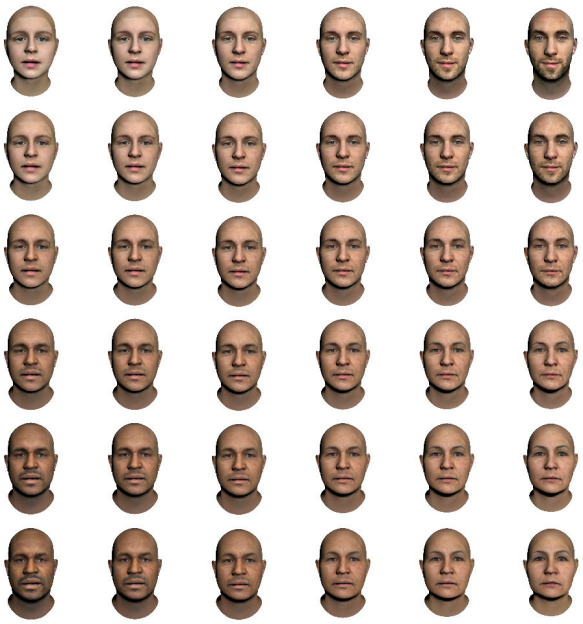

Figure 3: Illustration of latent vector interpolation. The four input 3D avatars are shown at the corners, while all the in-between interpolations are based on bi-linear interpolated weights.

Face Interpolation. In Fig. 3, we show interpolation results of multiple 3D avatars. The four input avatars are shown at the corners. All the interpolation results are obtained via bi-linearly interpolation of the embedding **w** computed from the four images. As shown in the results, realistic, plausible, and artifact free avatar assets can be generated using our method, which can be useful for a wide range of avatar manipulation and synthesis tasks.

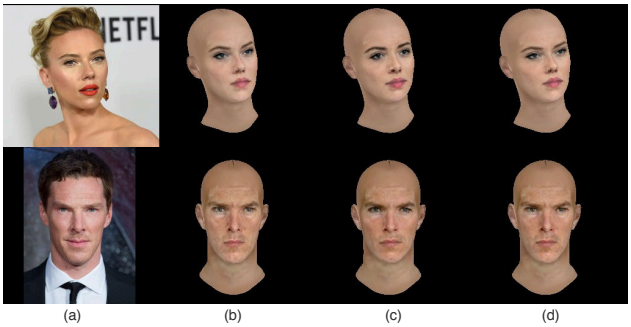

Figure 4: Visual comparison illustrating the effects of losses in the perceptual refinement step, where the full model leads to better results. From left to right: (a) input image; (b) refinement result with identity loss and **w** regularization; (c) refinement result with perceptual loss and **w** regularization; (d) refinement result with all three losses.

Optimization Loss. Fig. 4 shows the benefit of each loss term in L_{refine} for the perceptual refinement. Combining identity loss, perceptual loss, and **w** regularization allows us to generate clean assets, where the resulting subject preserves the likeness of the subject in the original input photo, but at the same time, ensures consistent and detailed assets with normalized lighting and neutral expressions.

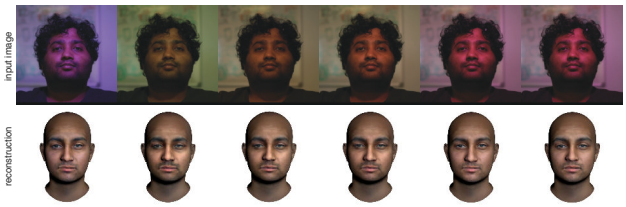

Figure 5: Consistent reconstructions of albedo texture under varying extreme illuminations.

Illumination Consistency. Fig. 5 demonstrates consistent face reconstructions of albedo textures from varying illuminations conditions. In this experiment we move around a light with different extreme colors around the subjects and demonstrate how a consistent 3D avatar with a nearly identical dark skin tone is correctly reconstructed for each input photo.

Expression Consistency. We demonstrate how consistent faces are reconstructed from input images with different expressions in Fig. 6. In particular, our method digitizes consistent 3D avatars with neutral expressions despite a wide range of diverse and extreme facial expressions of the same person as shown in the first row and the third row. While some amount of the input expressions are reflected in the normalized results, the overall neutralization is significantly

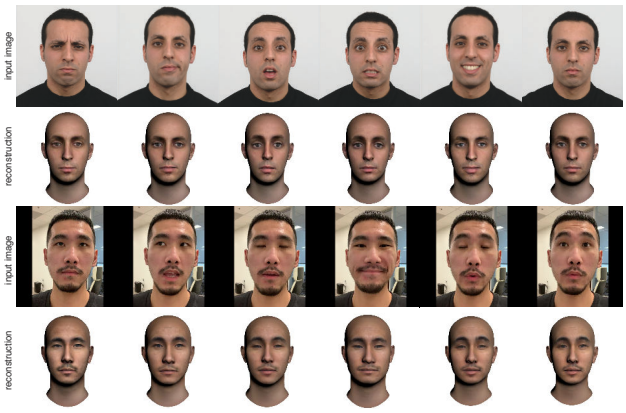

Figure 6: Consistent reconstructions of 3D avatars from images with different expressions.

superior than existing techniques, especially for extreme input facial expressions.

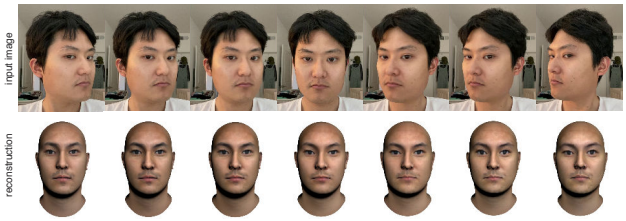

Figure 7: Consistent reconstructions under different poses.

Pose Consistency. Fig. 7 shows consistent reconstructions from varying head poses. For side views, our method can still generate highly consistent textures and geometries despite non-visible face regions in the input image.

Appendix III. Additional Results

To demonstrate the robustness of the our technique, we provide 156 additional examples with a wider range of extremely challenging input photographs in Fig. 8, Fig. 9, Fig. 10, and Fig. 11. These figures illustrate input pictures, successful normalized 3D face reconstructions, as well as renderings using HDRI-based lighting environments. Our results include diverse ethnicity, both genders, and varying age groups, ranging from children to old people. We also showcase a wide range of complex lighting conditions, stylized photographs, black and white portraits, drawings and paintings, facial occlusions, as well as a wide range of extreme head poses and facial expressions. Notice that we also show several results of the same person, but reconstructed from entirely different input images.

References

[1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[2] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Computer Vision and Pattern Recognition Workshops*, 2019.

[3] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[5] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, 2016.

[7] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019.

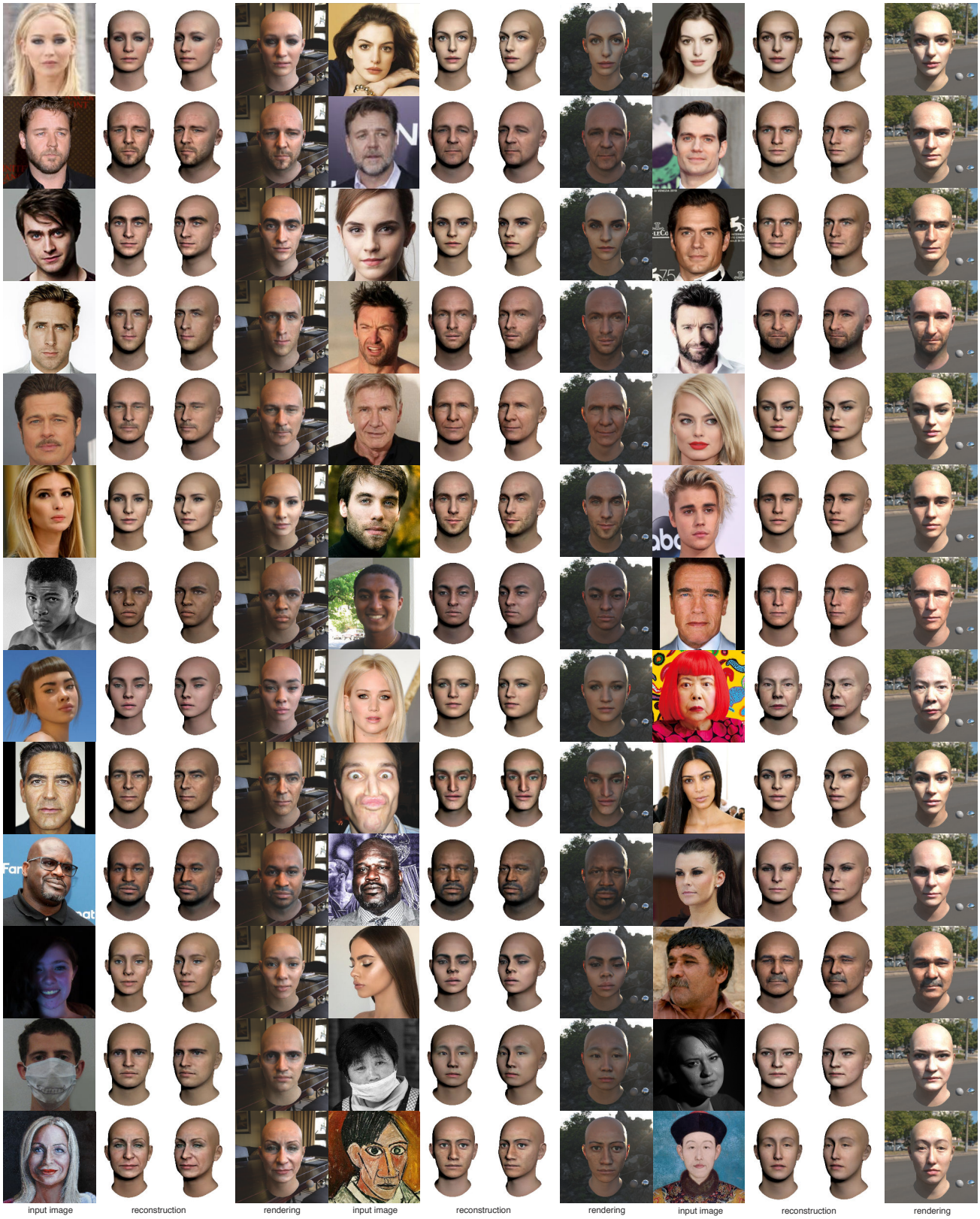

Figure 8: Batch 1 additional results of normalized 3D avatars from a single input image. None of these subjects have been used in training for our networks.

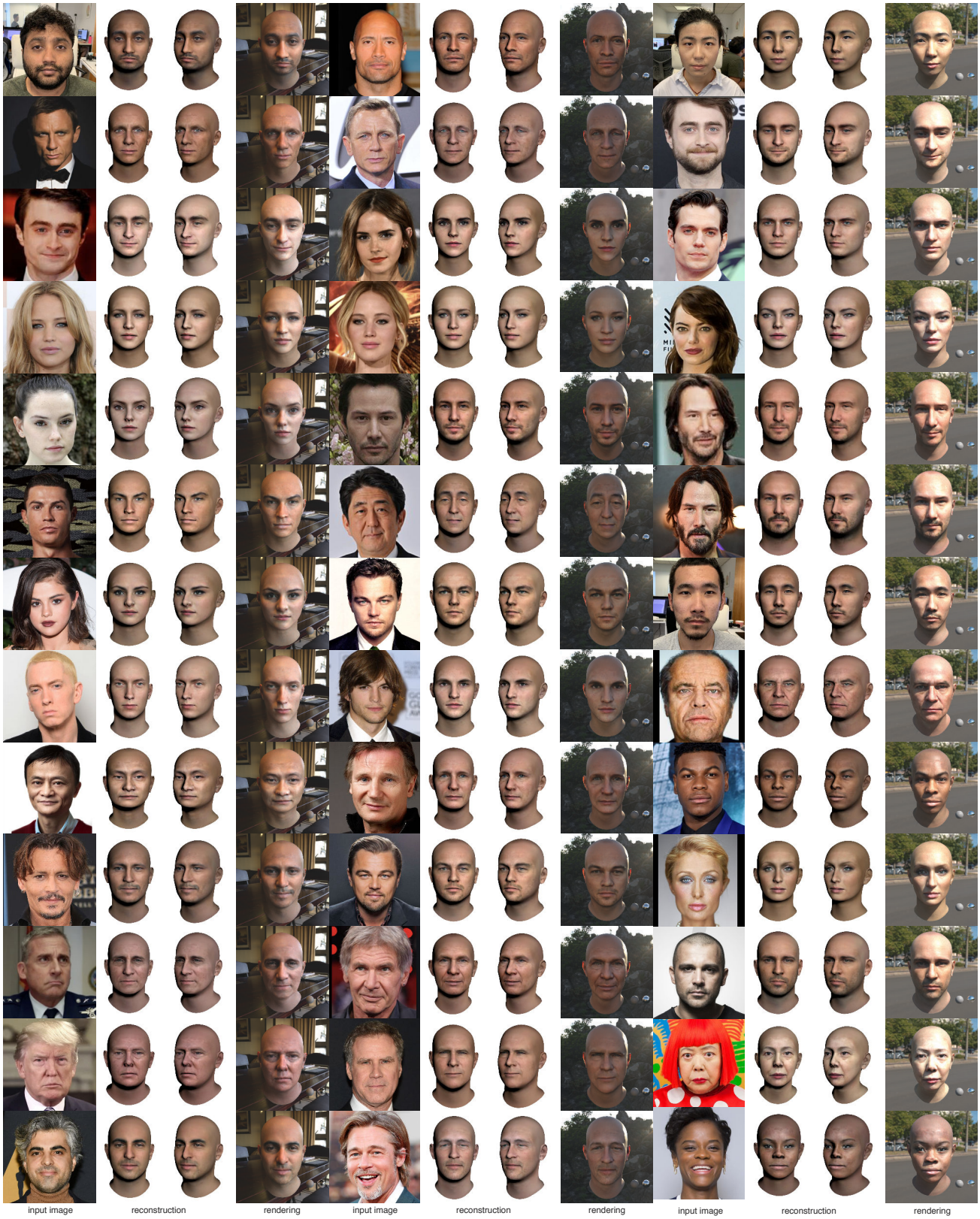

Figure 9: Batch 2 additional results of normalized 3D avatars from a single input image. None of these subjects have been used in training for our networks.

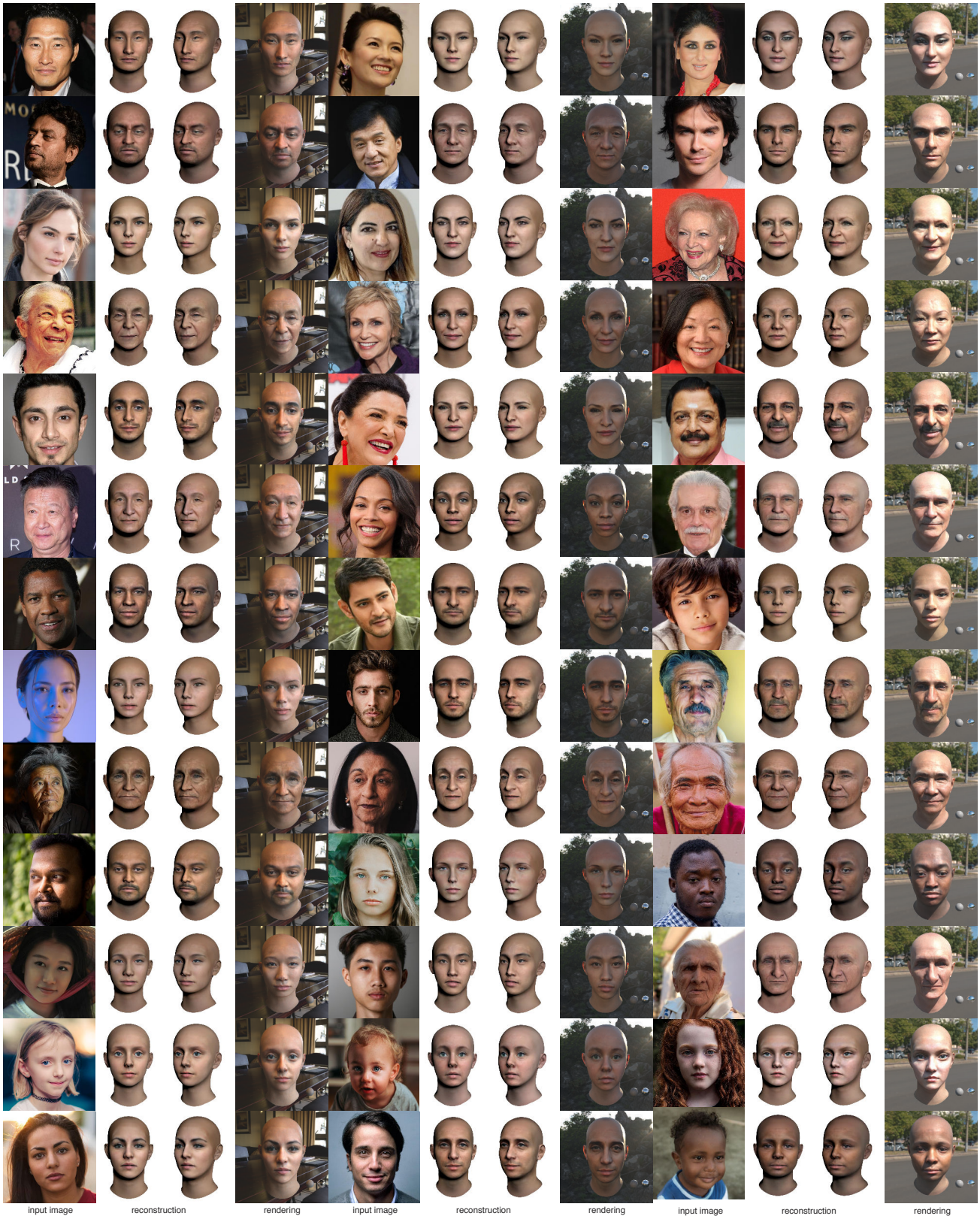

Figure 10: Batch 3 additional results of normalized 3D avatars from a single input image. None of these subjects have been used in training for our networks.

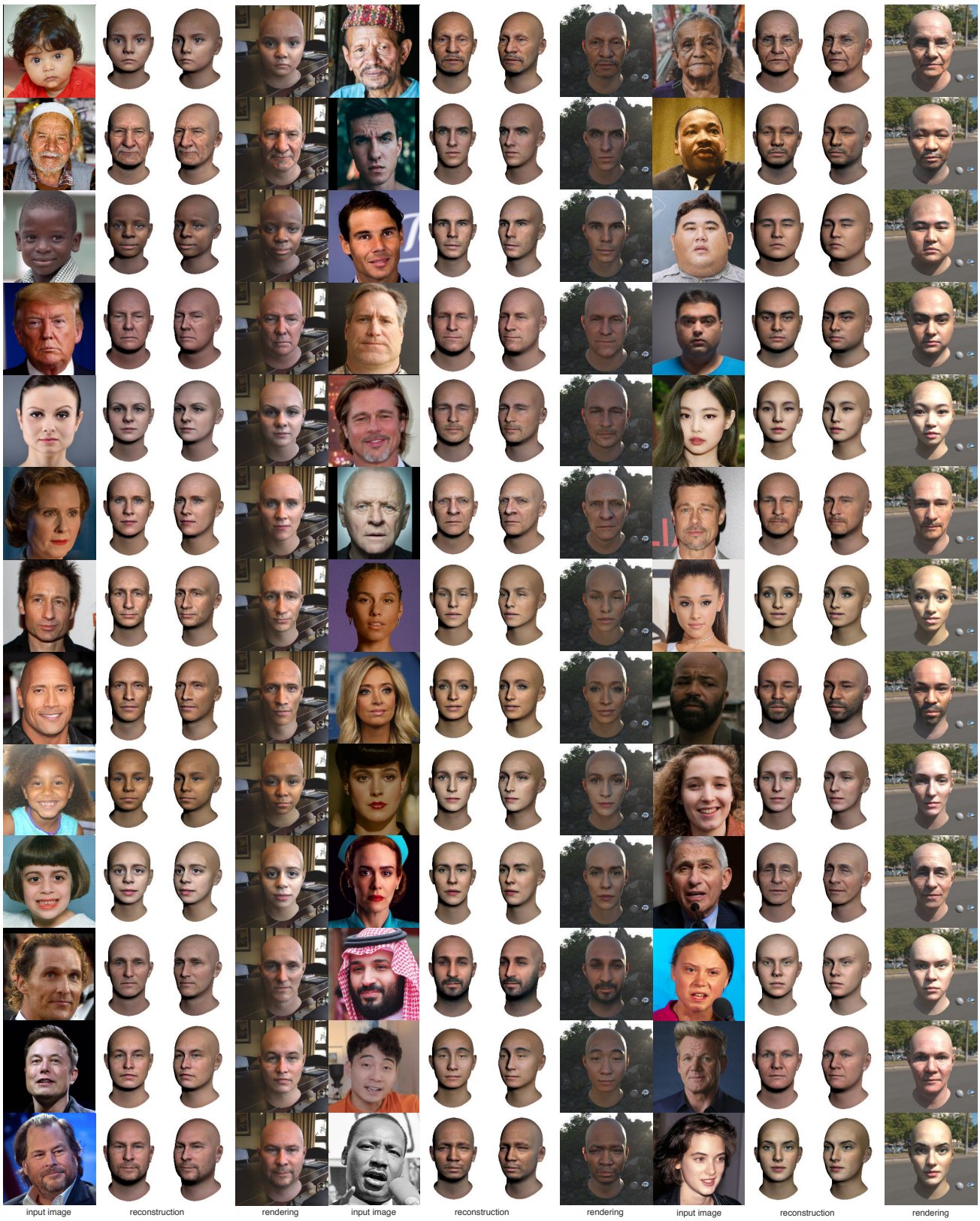

Figure 11: Batch 4 additional results of normalized 3D avatars from a single input image. None of these subjects have been used in training for our networks.