Supplementary material for paper: MoSAR: Monocular Semi-Supervised Model for Avatar Reconstruction using Differentiable Shading

Abdallah Dib[1*] Luiz Gustavo Hafemann[1*] Emeline Got[1] Trevor Anderson[1] Amin Fadaeinejad[1,2]
Rafael M. O. Cruz[3] Marc-André Carbonneau[1]

[1]Ubisoft LaForge York University[2] Ecole de Technologie Supérieure[3]

1. Hyperparameters

We use the following hyperparameters for training the 3D Geometry Reconstruction networks: $lr = 0.00005$, $\lambda_{\text{photo}} = 0.1$, $\lambda_{\text{landmark}} = 0.001$, $\lambda_{\text{lap}} = 0.1$, $\lambda_{\text{light}} = 0.001$, $\lambda_{\text{exp}} = 0.1$, $\lambda_{\text{alb}} = 0.1$, $\lambda_{\text{supervised}} = 1.$, $\lambda_{\text{nrm}} = 0.1$.

The texture completion and light normalization networks are both trained with $lr = 0.001$.

The displacement map estimation is trained with $lr = 0.0001$. The texture estimation networks are trained with: $lr = 0.0001$, $\lambda_{\text{shading}} = 0.5$, $\lambda_{\text{sup}} = 1$, $\lambda_{\text{GAN}} = 1$.

All modules were implemented in Pytorch, and trained on two CUDA-enabled GPUs with 24 GB RAM. Used the Adam optimizer [4] for training all networks.

2. FFHQ-UV-Intrinsics

In this section, we describe the process of generating our new dataset named *FFHQ-UV-Intrinsics*, built from the publicly available dataset FFHQ-UV [1]. The FFHQ-UV dataset is composed of texture maps of $1K$ resolution, for subjects sampled from the latent space of StyleGAN. These texture contains evenly illuminated face images. However, light, geometry and skin reflectance information are entangled in the same texture making them less suitable for relighting.

To obtain the intrinsic face attributes, we first re-targeted the texture maps to our own topology and resize them to 512×512. Next, we apply the proposed *light normalization* and *Intrinsic texture maps estimation* steps. We then upscale these texture maps to 1K resolution and retarget them back to their original topology.

The resulting dataset, *FFHQ-UV-Intrinsics*, is being publicly released for the research community. The dataset contains diffuse, specular, ambient occlusion, translucency and normal maps for 10K subjects. This is the first dataset that offer rich intrinsic face attributes at high resolution and at large scale, with the aim of advancing research in this field.

*Equal contribution

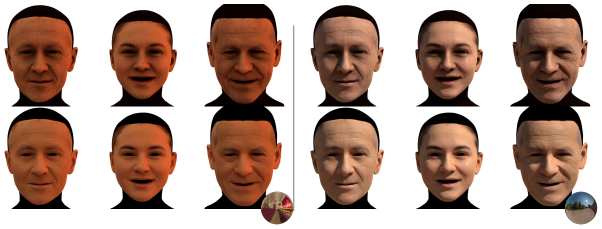

Figure 1. Impact of using AO and Translucency maps: with (top) and without these maps (bottom).

Figure 2. Results on images with strong directional light. Top: Input. Middle: Completed Texture. Bottom: Light-normalized.

3. Impact of AO and Translucency

Ambient Occlusion (AO) and Translucency maps are commonly used in modern rendering engines to improve realism. Figure 1 shows renders for the same subject, with and without these maps.

4. Challenging light conditions

Figure 2 shows results of the proposed light normalization (LN) on challenging light conditions. We noticed that light normalization effectively remove strong lights.

Figure 3 shows additional comparisons of the full model vs the supervised(+LN) model. It depicts the same 7 subjects of Figure 2. It shows that the full model is better at removing residual light that the LN step could not remove. This is most noticeable around the eyes and nose - see the

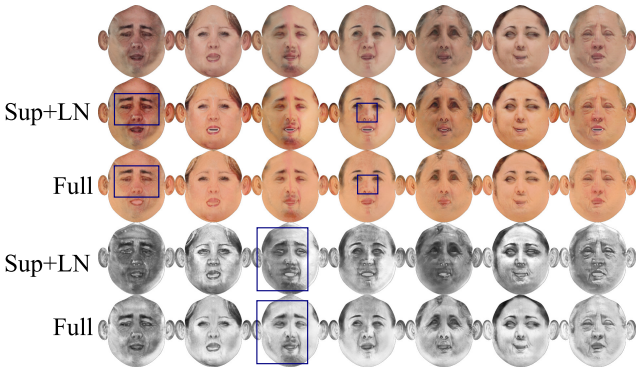

Figure 3. Additional comparison of estimated Diffuse and specular albedos (same subjects as Fig. 2).

highlighted regions. These important improvements are localized around salient areas of the face, which explains why quantitative results (Table 2 in the paper) are not largely different despite a significant qualitative improvement. These results generalize across subjects and support the claim that semi-supervised training improves quality over only supervised training.

5. Comparisons on additional subjects

Figures 4 to 10 show additional comparisons between our method, FitMe [5] (second row) and Relightify [7] (third row). For every method, we show the estimated geometry and the rendering under 4 different environment maps.

Additionally, we perform geometry comparison on the same 20 subjects for methods that only estimates geometry. Figures 11 and 12 show comparison of our estimated geometry against DECA [3] , HRN [6] and Deep3D [2].

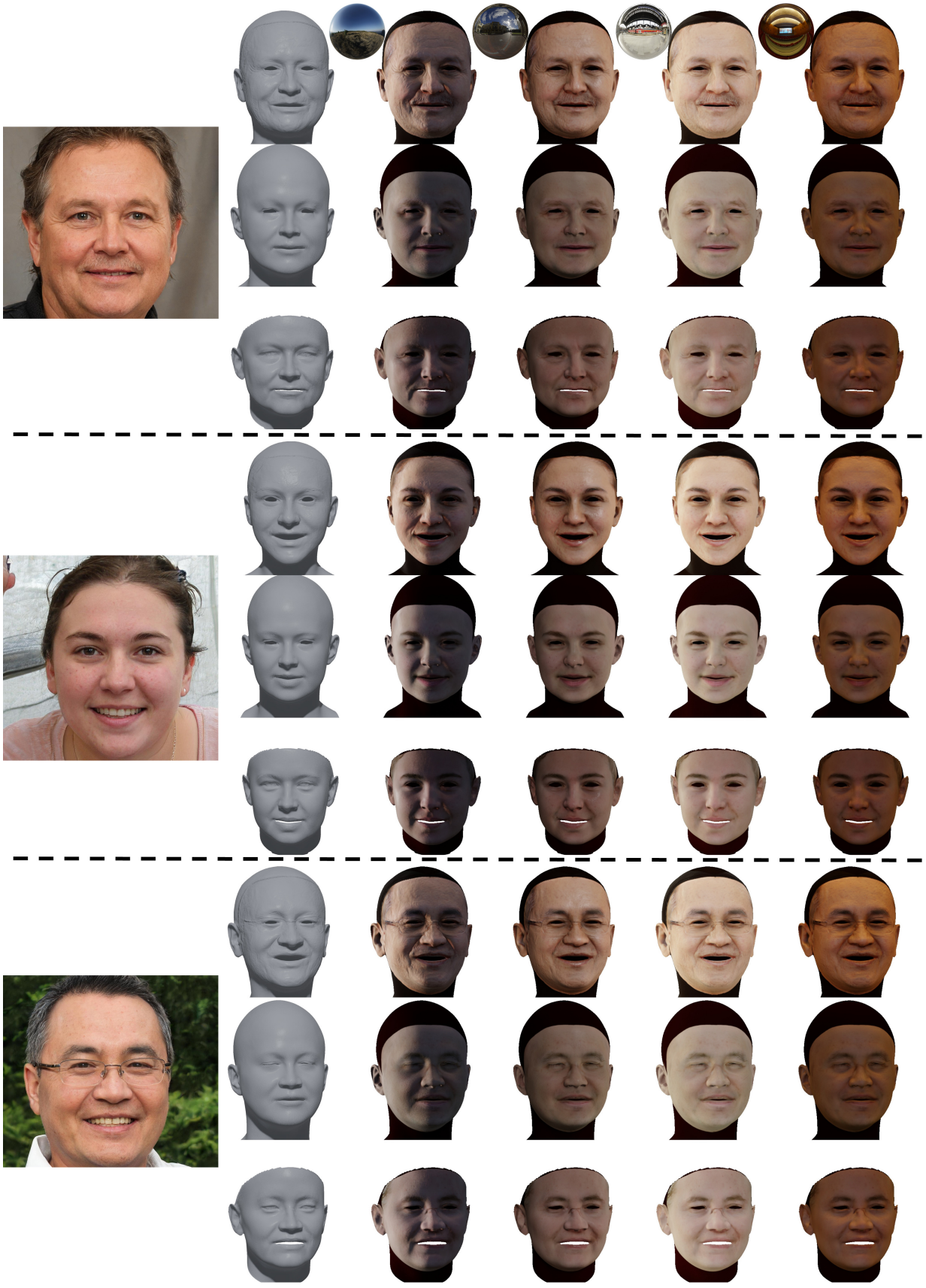

Figure 4. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

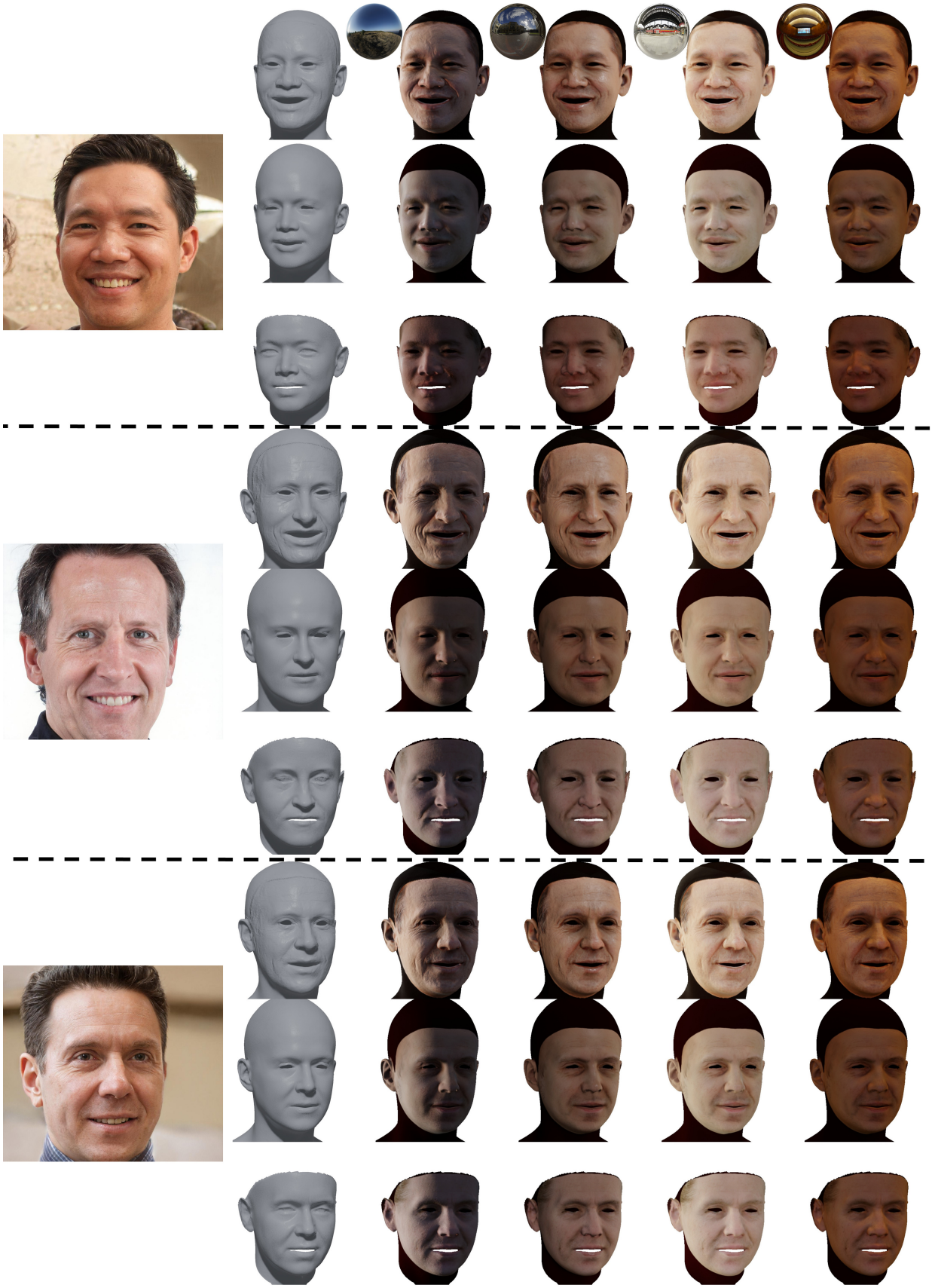

Figure 5. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

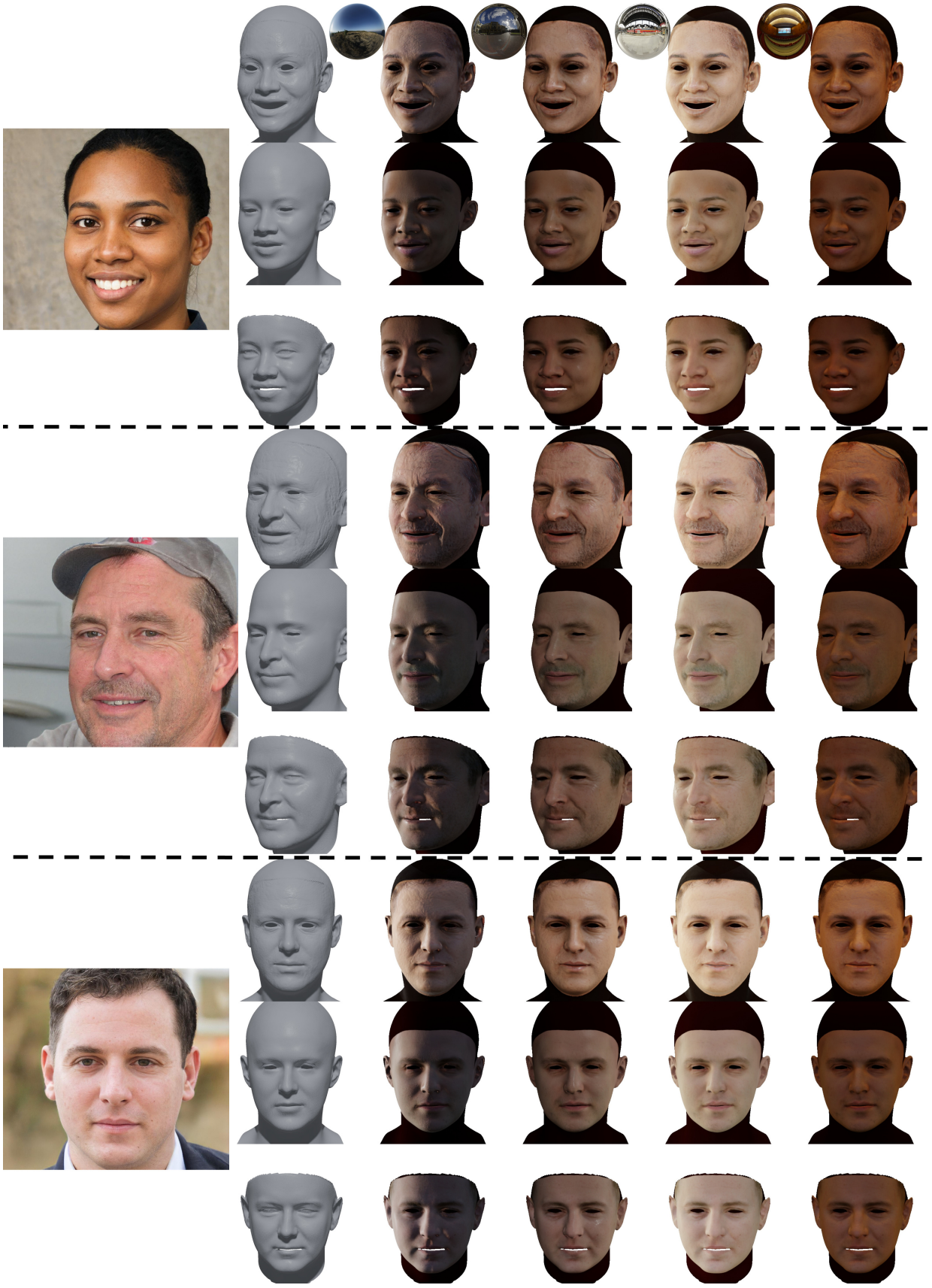

Figure 6. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

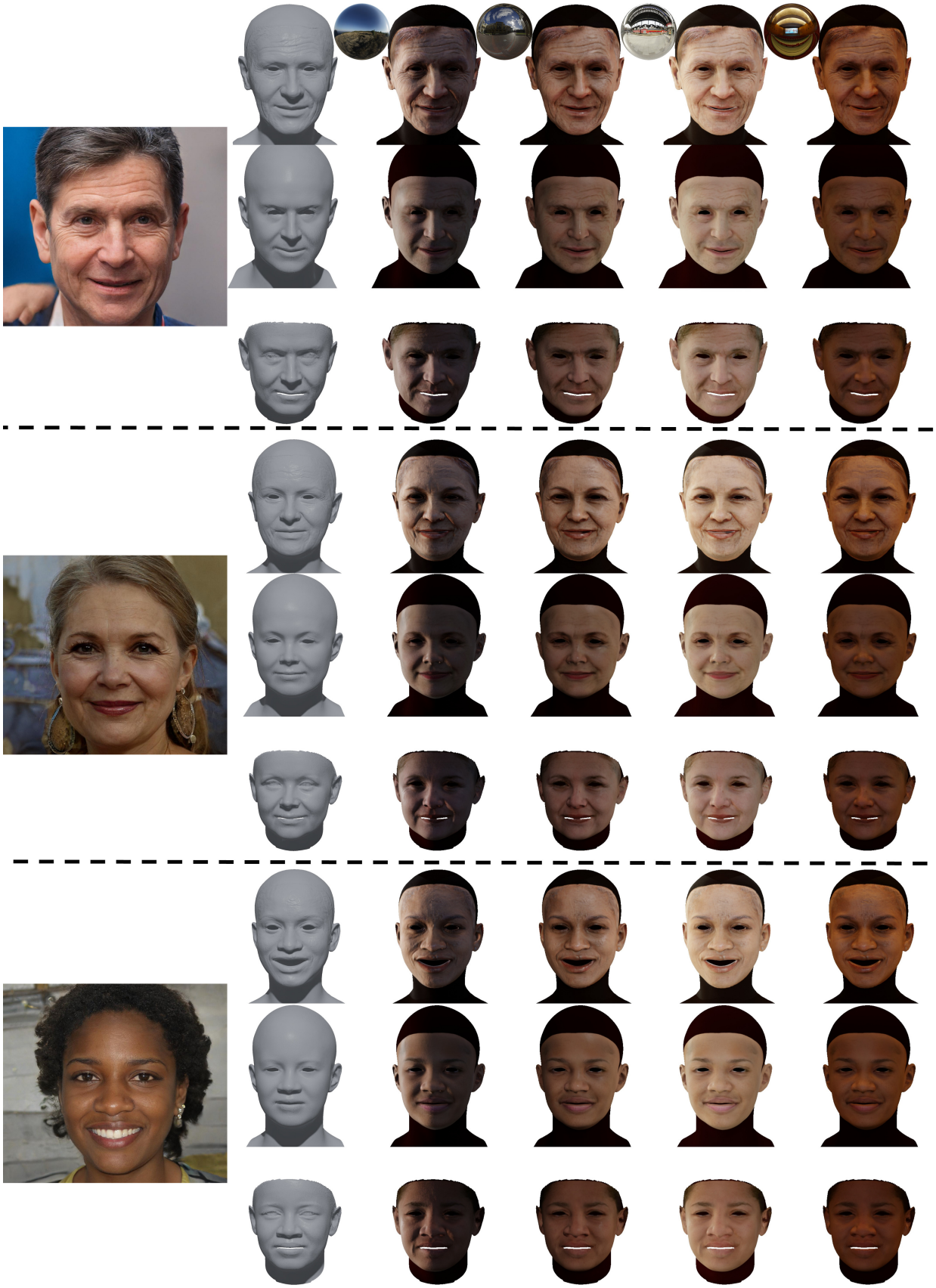

Figure 7. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

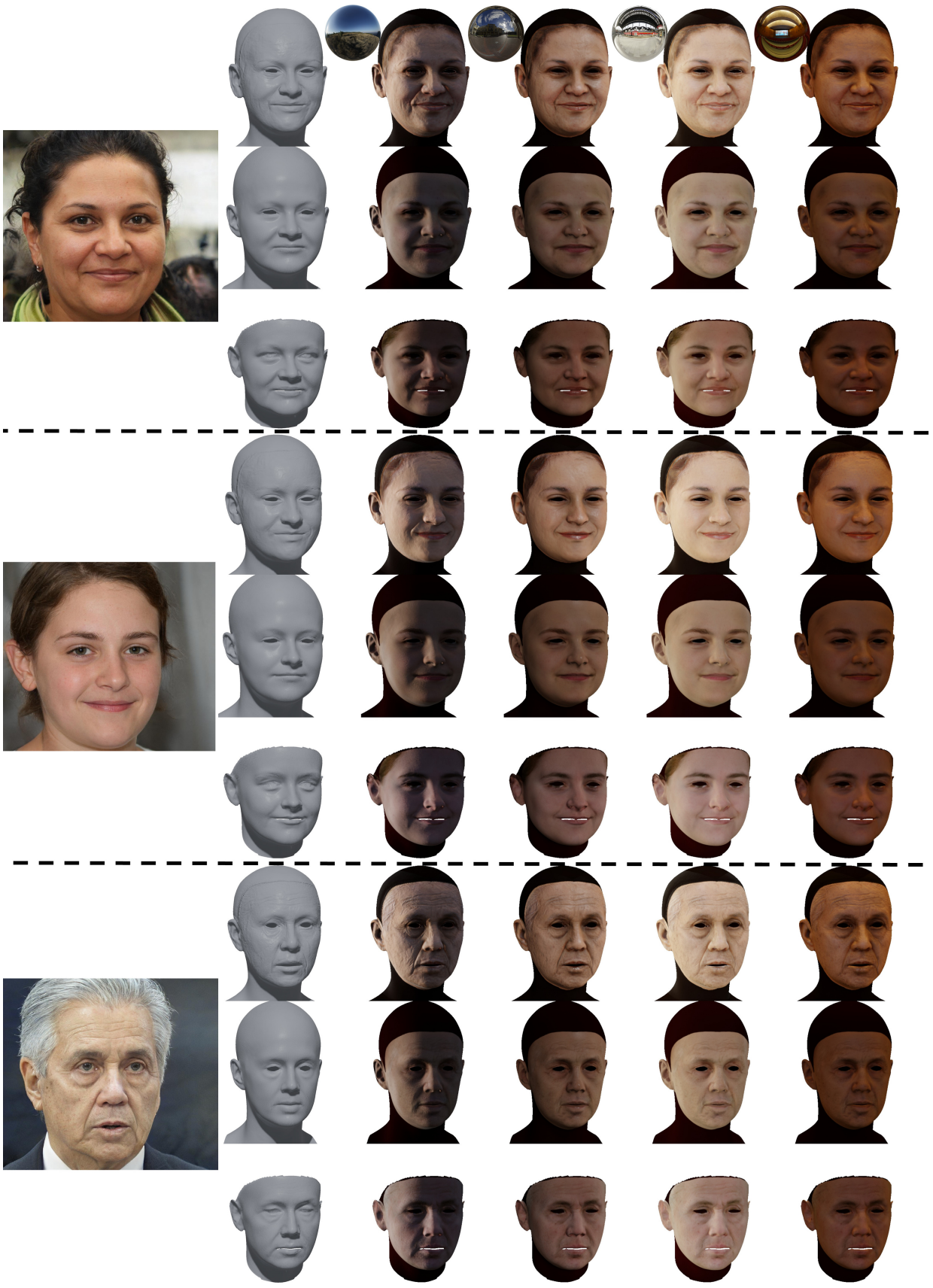

Figure 8. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

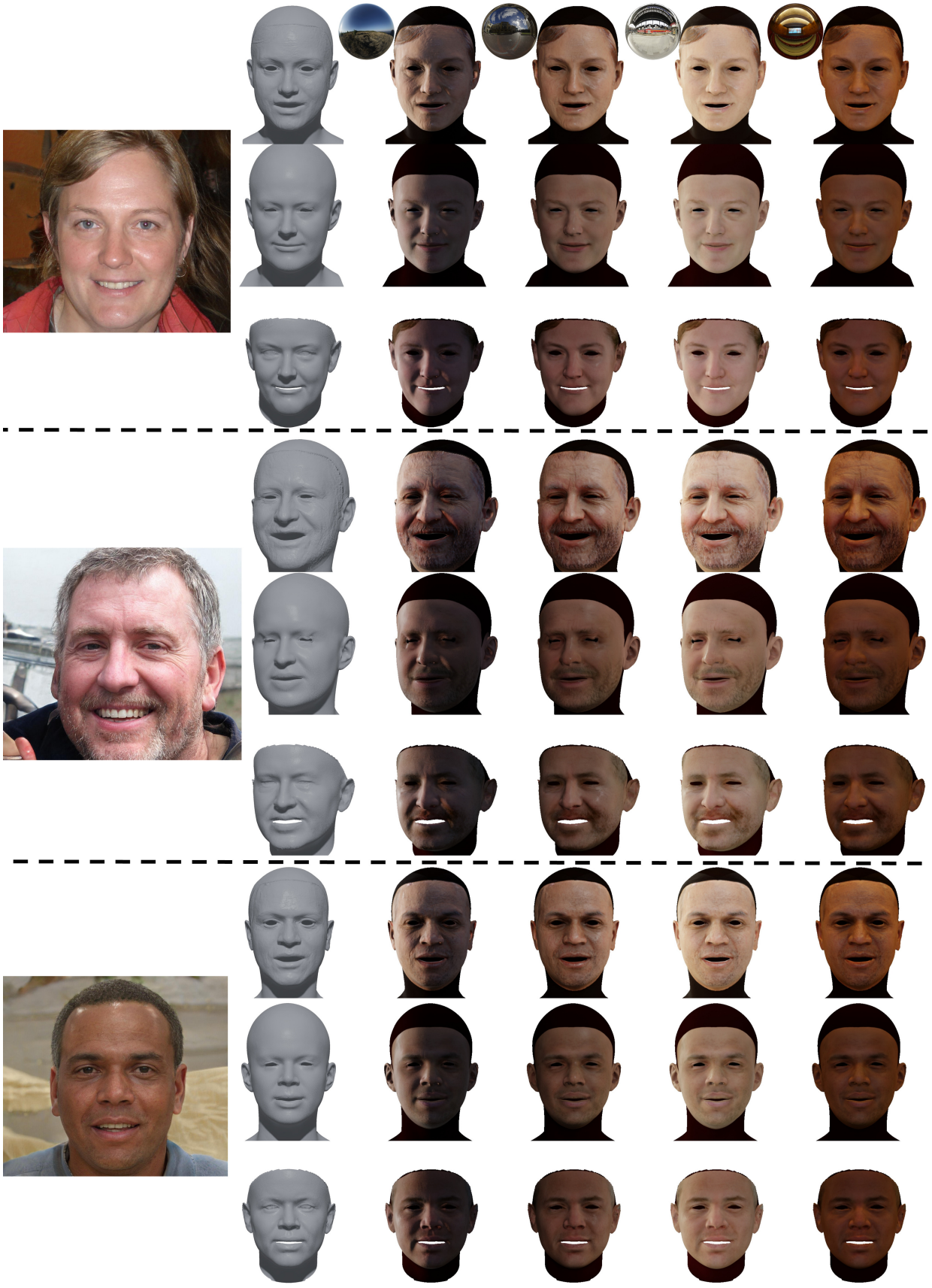

Figure 9. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

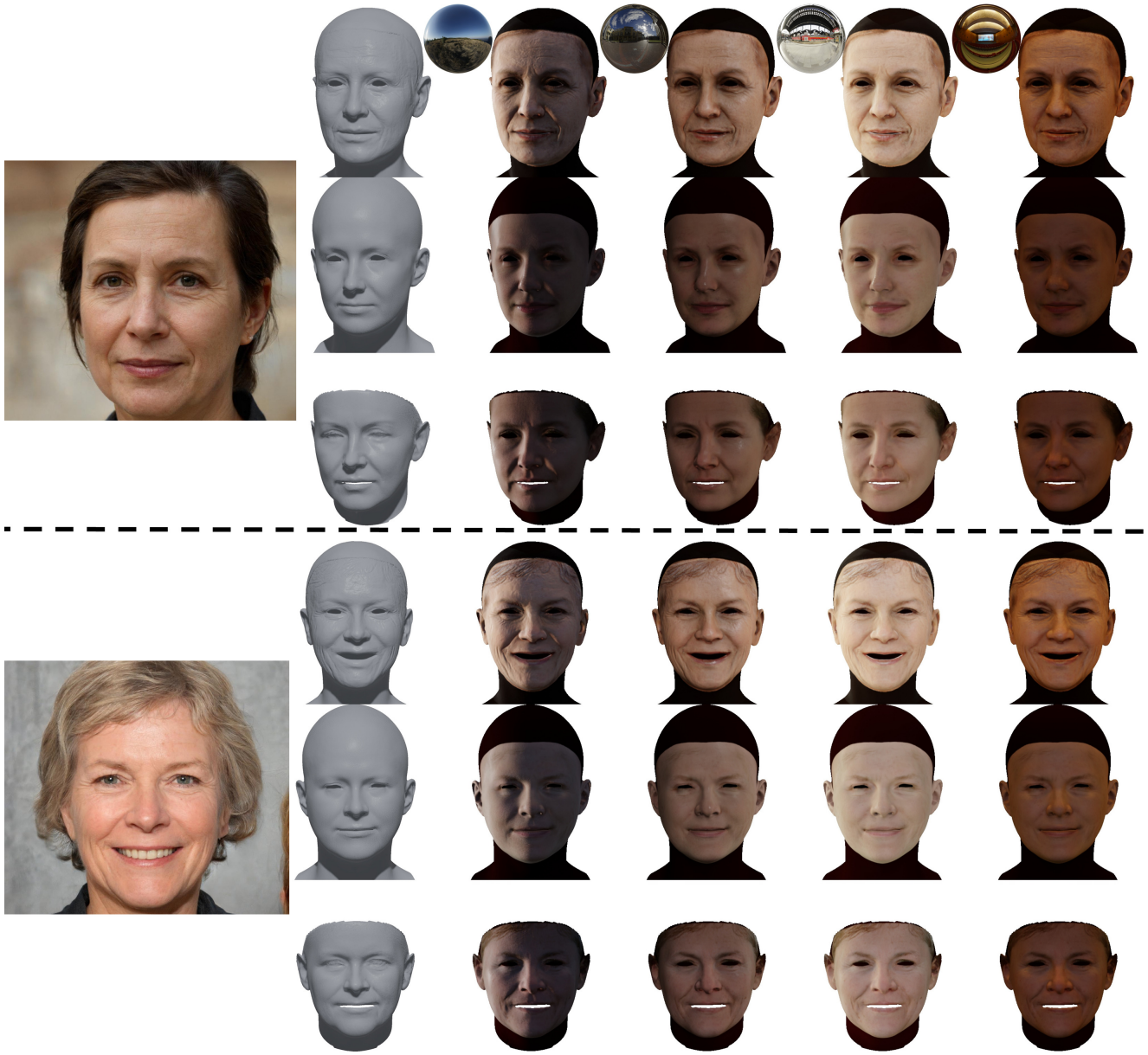

Figure 10. Comparison of the estimated geometry and renders, under 4 lighting conditions, between our method (first row), FitMe [5] (second row) and Relightify [7] (third row)

Figure 11. Comparison of the estimated geometry between our method (second row), DECA [3] (third row), HRN [6] (fourth row) and Deep3D [2] (last row)

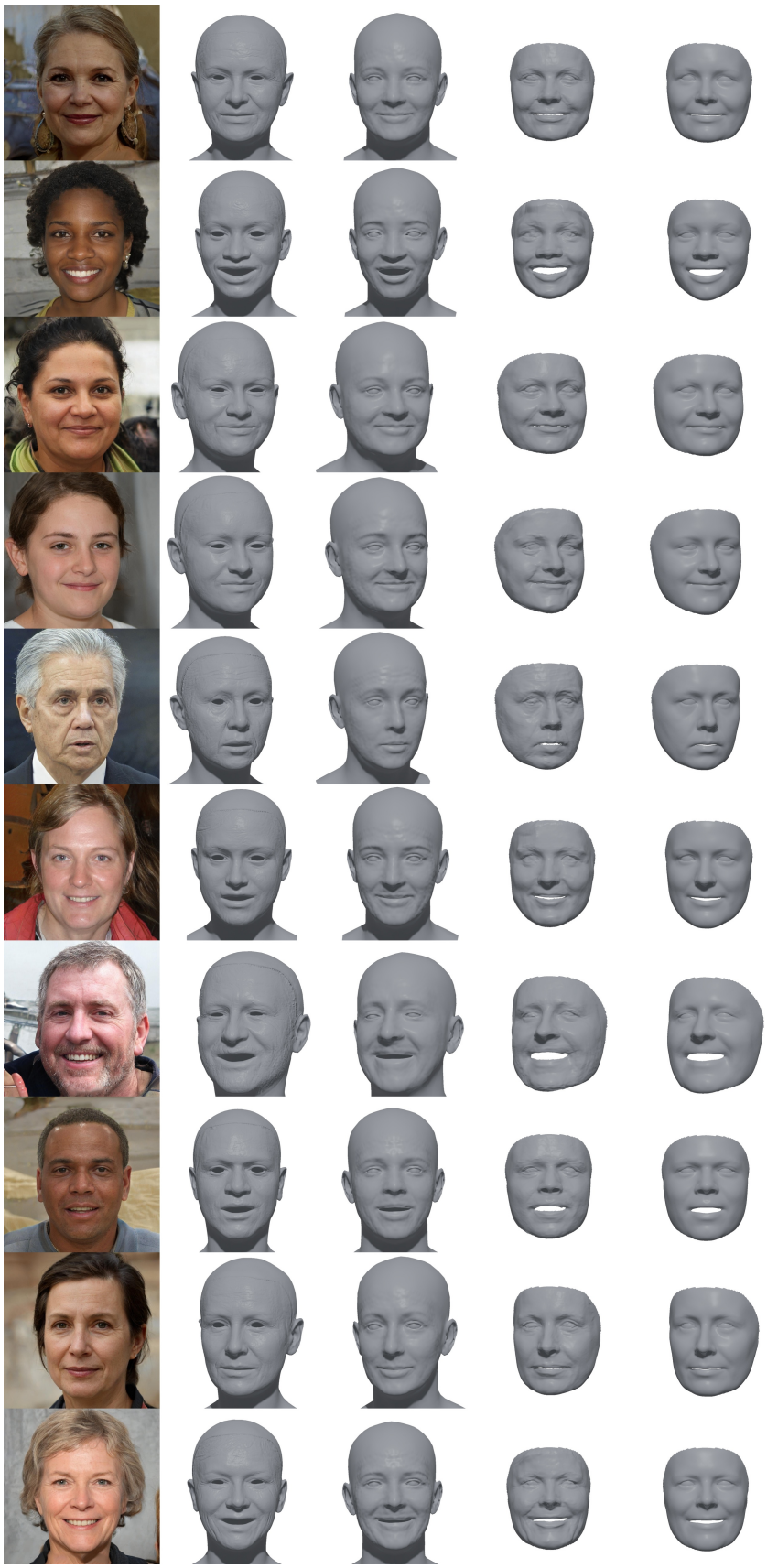

Figure 12. Comparison of the estimated geometry between our method (second row), DECA [3] (third row), HRN [6] (fourth row) and Deep3D [2] (last row)

References

[1] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 10, 11

[3] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13, 2021. 2, 10, 11

[4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 1

[5] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8629–8640, 2023. 2, 3, 4, 5, 6, 7, 8, 9

[6] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 2, 10, 11

[7] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 5, 6, 7, 8, 9