Supplementary Material for "VoLux-GAN: A Generative Model for 3D Face Synthesis with HDRI Relighting"

Feitong Tan[1,2,*] Sean Fanello[1] Abhimitra Meka[1] Sergio Orts-Escolano[1] Danhang Tang[1]
Rohit Pandey[1] Jonathan Taylor[1] Ping Tan[2] Yinda Zhang[1]

[1] Google [2] Simon Fraser University

In this supplementary material, we provide more details regarding the proposed data augmentation strategy, network architecture and additional results. Finally, we also discuss the limitations of the model. We also provide a supplementary HTML page showing animated results of generated face under various camera viewpoints and environmental illuminations.

1. Data Augmentation via Portrait Relighting

We provide additional information regarding our data augmentation strategy which uses the portrait relighting method of [12] to produce pseudo ground truth albedo, normals, a relit image and the associated light maps (diffuse and specular components) on the CelebA [9] and the FFHQ [7] datasets. Specifically, we generate 5 and 10 relit images for each image in CelebA and FFHQ datasets. The HDRI map is randomly sampled from a collection of 400 maps sourced from public repository [14] and randomly rotated horizontally. We show more example images of the augmented CelebA images and FFHQ images in Figure 1 and Figure 2. For each identity, we visualize the relit image and the associated light maps with two different HDRI images.

2. Network Architecture

The details of the proposed architecture are shown in Figure 3. As detailed in the main paper, the framework consists of four modules: a neural implicit intrinsic field (NeIIF) network, upsampling blocks, a relighting network and a mapping network. Similar to StyleGAN2 [8], the mapping network consists of 8 fully-connected layers with 512 units, that maps the latent code to a style vector. The output vectors are then broadcast to every fully-connected layer in the NeIIF network and the upsampling blocks. For each vector, there is an affine transformation layer to map it to an affine-transformed style, which is used to modulate the feature maps of the NeIIF network and upsampling block. The NeIIF network consist of a positional encoder (the Fourier feature dimension is set to 10) and a 6-layer MLP with 256 units. The feature maps of each fully-connected layer are modulated by an affine transformation from the mapping network. Each upsampling block consists of two fully-connected layers modulated by the latent code z, a pixelshuffle upsampler and a BlurPool with stride 1 which increases the resolution by 2x. The relighting network is a residual U-Net with skip connections.

3. Additional Results

Here we show more results from our method.

3.1. Intermediate Intrinsic Images

In Figure 4, we visualize the albedo, relit image, normal map, diffuse map and specular map from our generator trained on FFHQ dataset. Note that since normal and shading maps are directly generated from the neural implicit field, we render them in low resolution for efficient training, which is a strategy adopted and demonstrated to be successful by other works [5]. We then rely on the generated feature map F to produce high frequency details in the albedo and the final relit results.

3.2. Relighting Accuracy

We show a qualitative comparison of our relighting method with environmental relighting of a real person captured in a dense high-resolution Light Stage in Figure 5, which is very close to ground truth relighting. Note that as the environment map rotates, our method produces plausible shadows and specularities that spatially match the pseudo-ground-truth setup, indicating that our underlying 3D volumetric geometry and skin reflectance is stable. While there is some dampening of specularities and cast shadows, the overall identity of the generated person is

[*] Work done while the author was an intern at Google.

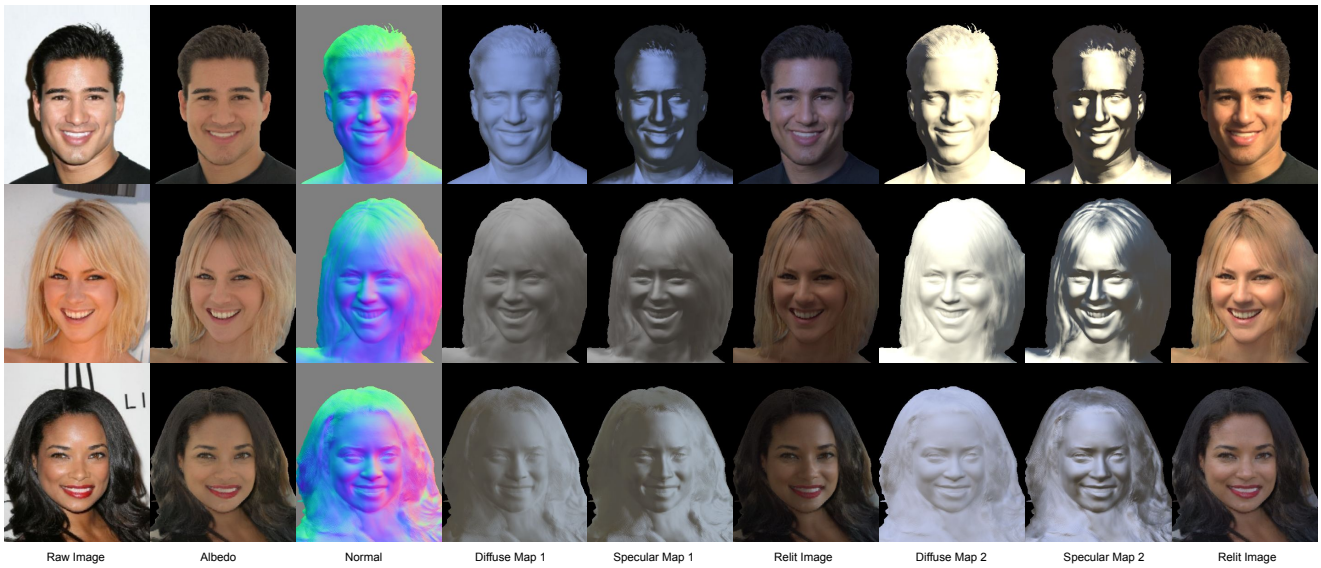

Figure 1. Relighting augmentation on CelebA [9] using [12] to generate albedo, normal, shading, and relit images with different HDRI Relighting, which supervise the training via adversarial losses.

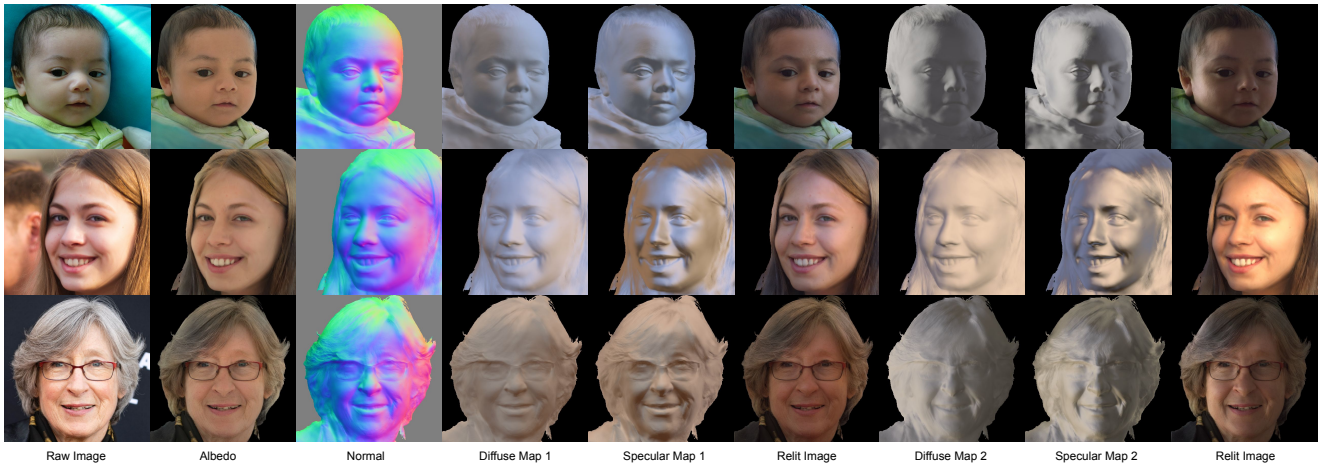

Figure 2. Relighting augmentation on FFHQ [7] using [12] to generate albedo, normal, shading, and relit images with different HDRI Relighting, which supervise the training via adversarial losses.

well preserved, which is a significant improvement over the state-of-the-art [11].

3.3. Rotate Camera and Lighting

We show more subjects generated from the model trained on the FFHQ dataset with randomly sampled latent codes in Figure 6. For each identity (*i.e.* latent code), we show the rendering under the same HDRI map but different camera pose, and the rendering under a fixed camera pose with rotating HDRI map. The results indicate that our method provides controllability over camera viewpoint and illumination, and deliver faithful rendering results.

4. Animated Results in Companion HTML Page

We provide a supplementary HTML page to show animated rendering results. Please open with your local browser. In the HTML page, we show 1) our intermediate intrinsic results and final relighting results in a continuous camera trajectory, 2) comparison on the relighting faithfulness to ShadeGAN [11] under rotating HDRI, using image based relighting with a Light Stage [6] as the reference, 3) relighting of the same or different subjects under same or different environment map, 4) multi-view synthesis, 5) a comparison on albedo stability with the baseline of pi-

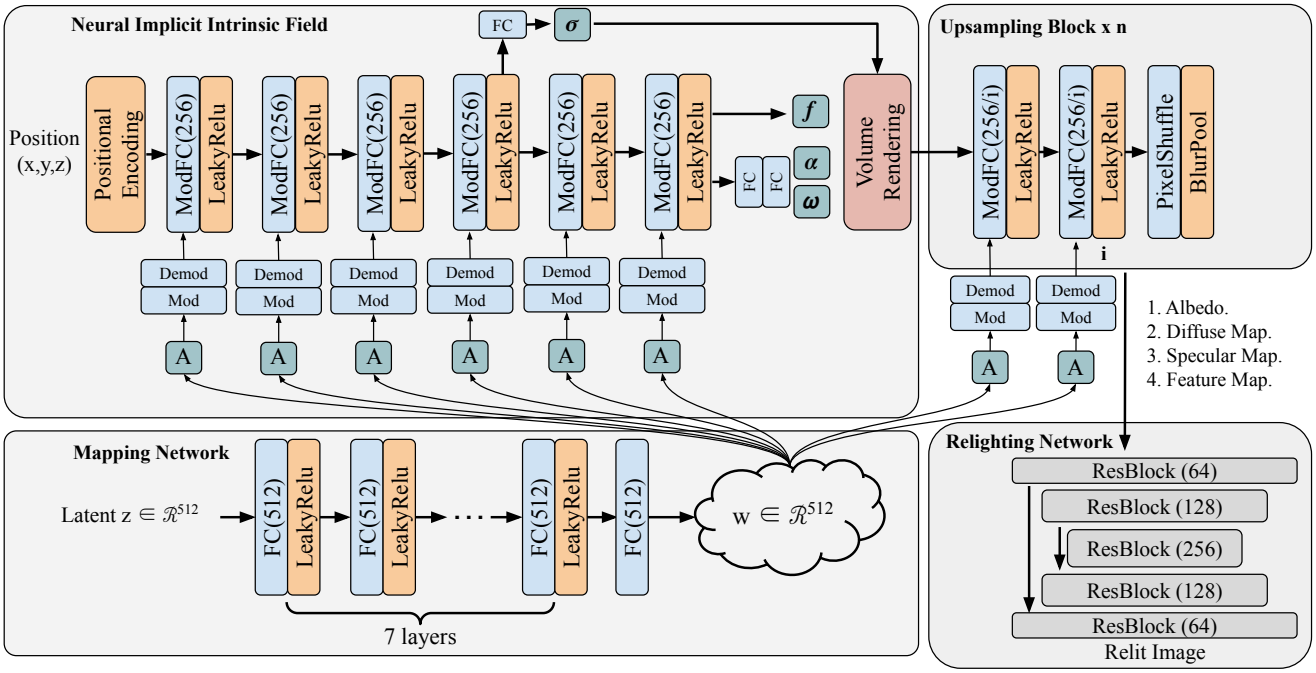

Figure 3. Proposed architecture of our neural generator, which consists of a neural implicit intrinsic field network, upsampling blocks, a relighting network and a mapping network.

GAN [4] + TR [12].

5. Limitations

Although the proposed approach is a step forward towards generative relightable 3D faces, it still has limitations. First, it lacks high frequency details on geometry and albedo when rendered at high resolutions (see Figure 4), despite our high quality supervision: we believe that using intuition from previous work [3, 7, 8] could help address this.

At more extreme viewpoint changes, the identity similarity scores drop as demonstrated in Table 1 in the main paper, indicating that stronger pose/viewpoint changes may result in distortion of identity. This is likely due to skewed distribution of our in-the-wild training data which is mostly frontal, with very few side facing views. We believe that this can be improved by more carefully curating the training data using importance sampling to have a more even distribution of facial poses. Yet, please note that our method outperforms other state-of-the-art 3D synthesis methods [4, 11], which in turn are significantly better than 2D based generative view synthesis methods [1, 10, 13].

Furthermore, aliasing effects are noticeable when changing viewpoints especially around the teeth and hair. An approach similar to [2] could potentially mitigate these effects.

Additionally, although our model shows impressive relighting results, it still cannot capture the same details of specular highlights when compared to image based relighting using a Light Stage as shown in Figure 5. Additional losses that focus on specularities may help mitigate this issue.

Finally, the lack of supervision on the actual facial expression, makes the model unconstrained, leading to different face gestures when changing the viewpoint (see animated results in the provided HTML page). Adding semantic information such as keypoints or per-pixel labels could be an effective way to enable control over the expressions and ensure more consistency across views.

References

[1] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021. 3

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021. 3

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 3

[4] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Pro-*

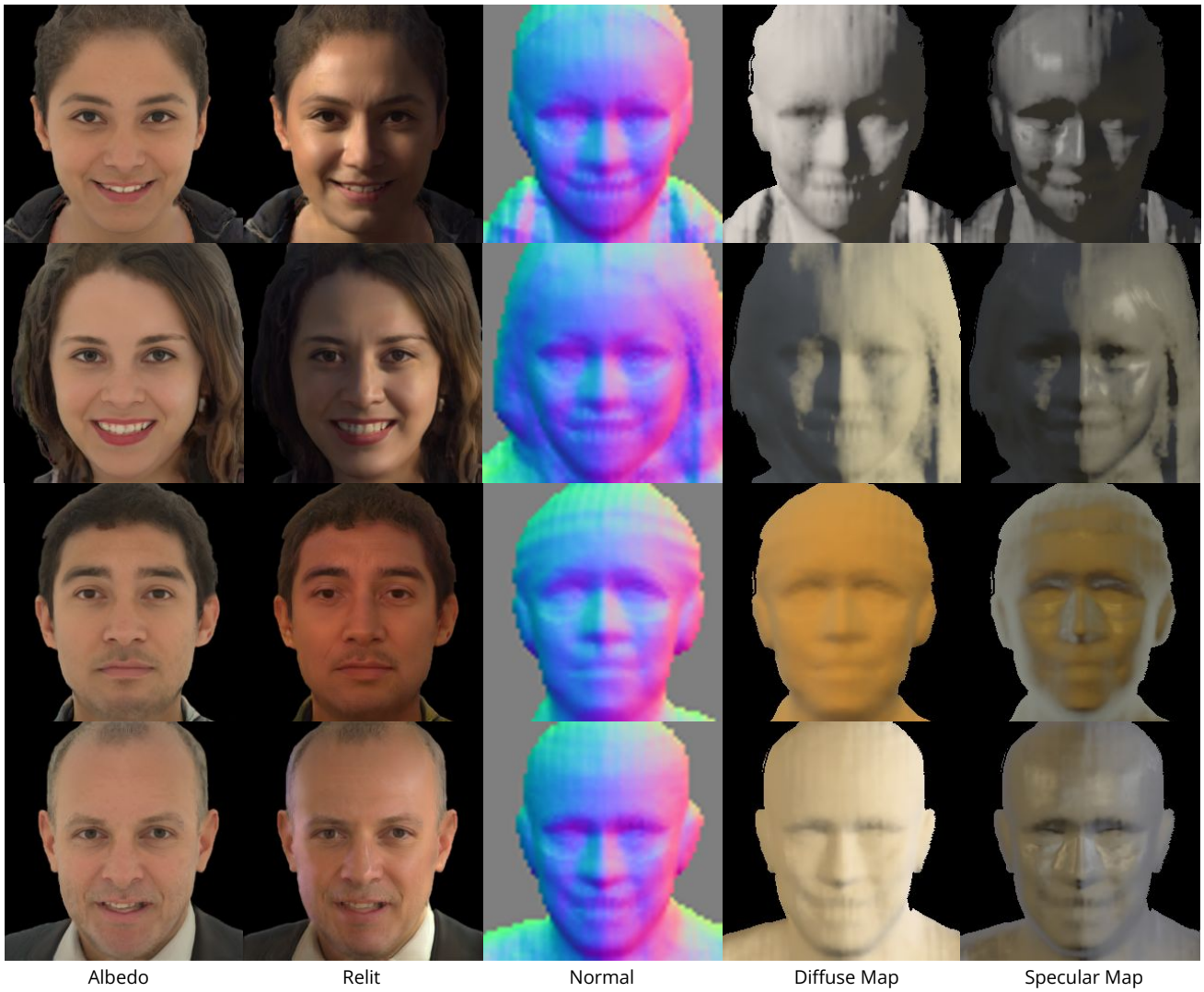

Figure 4. Results of intermediate intrinsic images from our model trained on FFHQ [7]. From left to right, we show the albedo, relit image, normal map, diffuse map and specular map.

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5799–5809, 2021. 3

[5] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 1

[6] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 38(6):1–19, 2019. 2, 5

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1, 2, 3, 4

[8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1, 3

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 2

[10] B R Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. Photoapp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics*, 40(4):1–16, 2021. 3

[11] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. A shading-guided generative im-

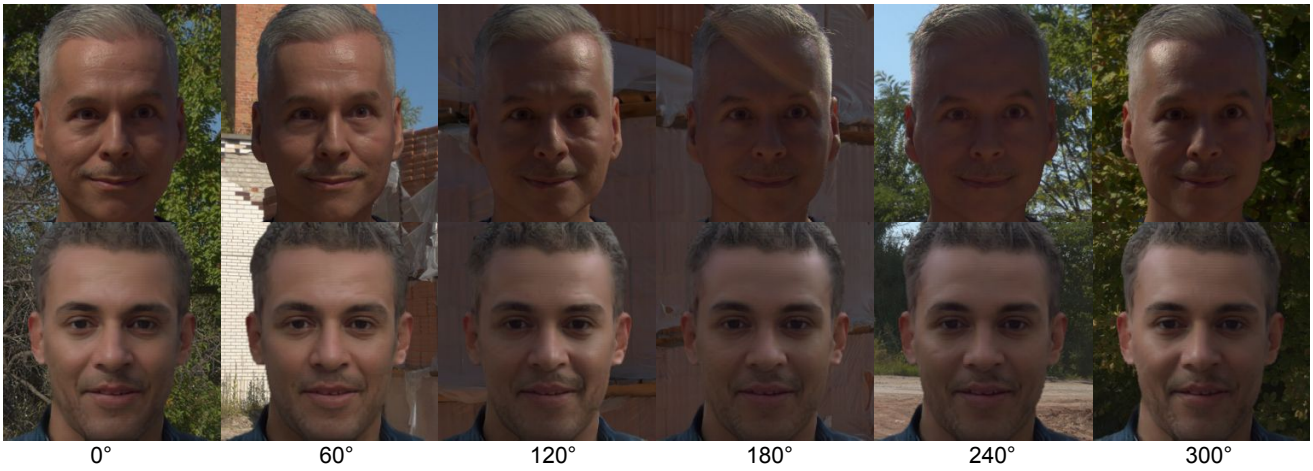

Figure 5. We compare our relighting result to image based relighting (IBR) using a Light Stage [6] with the same HDRI illumination. Note that our method produces consistent and plausible shading, soft shadows and specularities.

plicit model for shape-accurate 3d-aware image synthesis. In *NeurIPS*, 2021. 2, 3

[12] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 1, 2, 3

[13] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, june 2020. 3

[14] Greg Zaal, Sergej Majboroda, and Andreas Mischok. Hdri haven. https://www.hdrihaven.com/, 2020. Accessed: 2021-11-13. 1

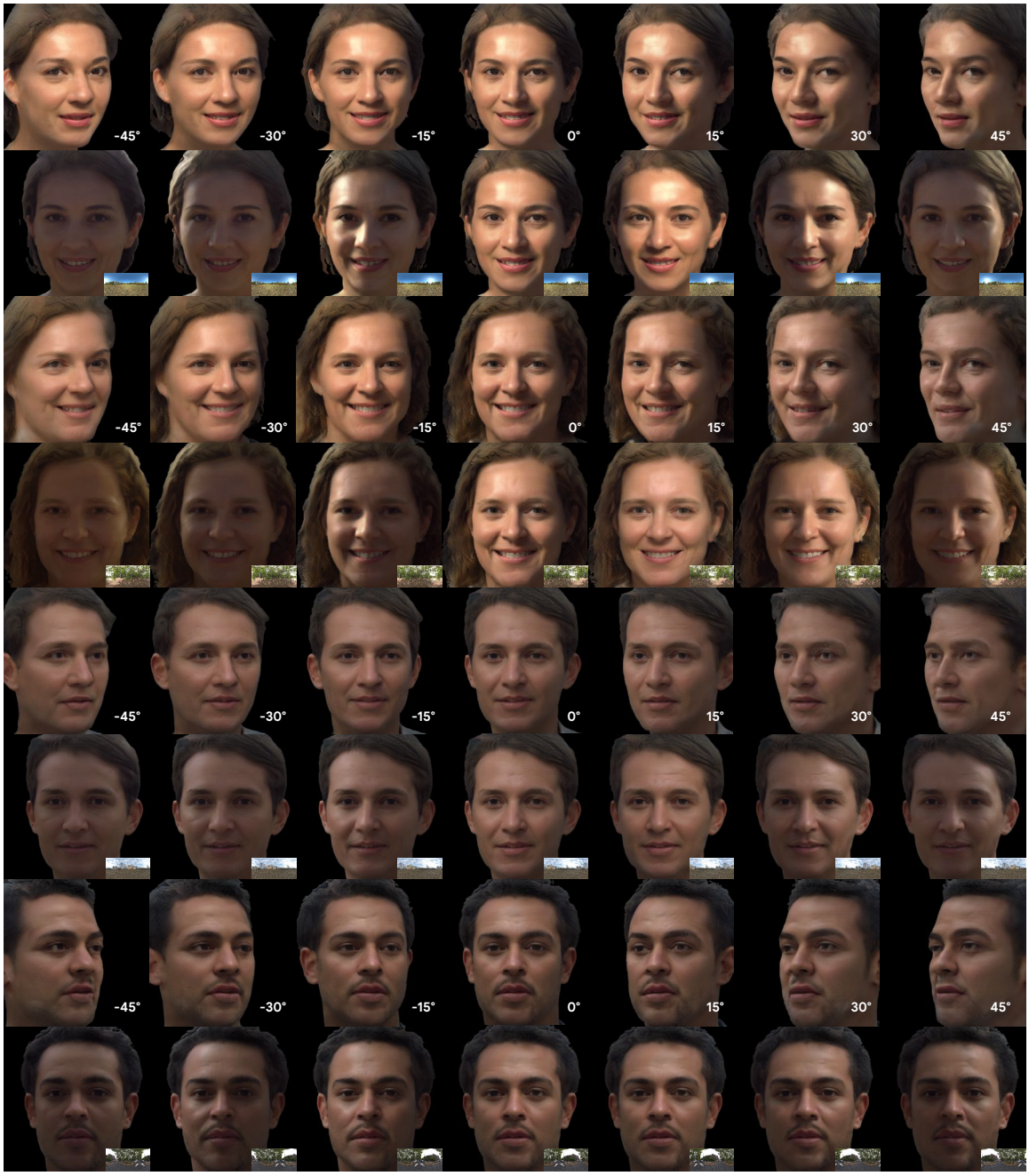

Figure 6. More synthesized images under rotating camera or rotating lighting. Note the relighting consistency and view-dependent effects.