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

Feitong Tan
Google
USA
feitongtan@google.com

Sean Fanello
Google
USA
seanfa@google.com

Abhimitra Meka
Google
USA
abhim@google.com

Sergio Orts-Escolano
Google
USA
sorts@google.com

Danhang Tang
Google
USA
danhangtang@google.com

Rohit Pandey
Google
USA
rohitpandey@google.com

Jonathan Taylor
Google
USA
jontaylor@google.com

Ping Tan
Simon Fraser University
Canada
pingtan@sfu.ca

Yinda Zhang
Google
USA
yindaz@google.com

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Computer vision**; **Rendering**.

KEYWORDS

Neural Rendering, Relighting, Generative Model

ACM Reference Format:
Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. 2022. Supplementary Material for "VoLux-GAN: A Generative Model for 3D Face Synthesis with HDRI Relighting". In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3528233.3530751

In this supplementary material, we provide more details regarding the proposed data augmentation strategy, implementation details and additional results. Finally, we also discuss the limitations of the model. We also provide a supplementary video 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 [Pandey et al. 2021] to produce pseudo ground truth albedo, normals, a relit image and the associated light maps (diffuse and specular components) on the CelebA [Liu et al. 2015] and the FFHQ [Karras et al. 2019] 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 [Zaal et al. 2020] and randomly rotated horizontally. We show example images of the augmented FFHQ images in Figure 1. For each identity, we visualize the relit image and the associated light maps with two different HDRI images.

2 IMPLEMENTATION DETAILS

We implement our framework in TensorFlow 2. To ensure stable training, we first train the neural implicit field and the upsampling network for high quality albedo and geometry. We only enable albedo and geometry adversarial loss and adopt progressive growing training strategy [Karras et al. 2017] with the path loss to train the upsampling network. Once the network converges, we enable all the loss terms and train the whole network end-to-end. During training, the Adam optimizer is applied with $\beta_1 = 0$, $\beta_2 = 0.99$ and the learning rates for the generator and the discriminator are set to 0.0022 and 0.0025 respectively. We train the VoLux-GAN model on 8 Tesla V100 GPUs with batch size 16 for 800k million iterations with albedo and geometry adversarial losses, which takes 5 days, then it is trained for additional 400k iterations to generate the relit images for around 3 days. The inference time is 0.272 seconds for rendering a relit image in 256×256 resolution.

2.1 Network Architecture

The details of the proposed architecture are shown in Figure 2. 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 [Karras et al. 2020], the mapping network consists of 8 fully-connected layers with 512 units, that maps the latent code

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08.
https://doi.org/10.1145/3528233.3530751

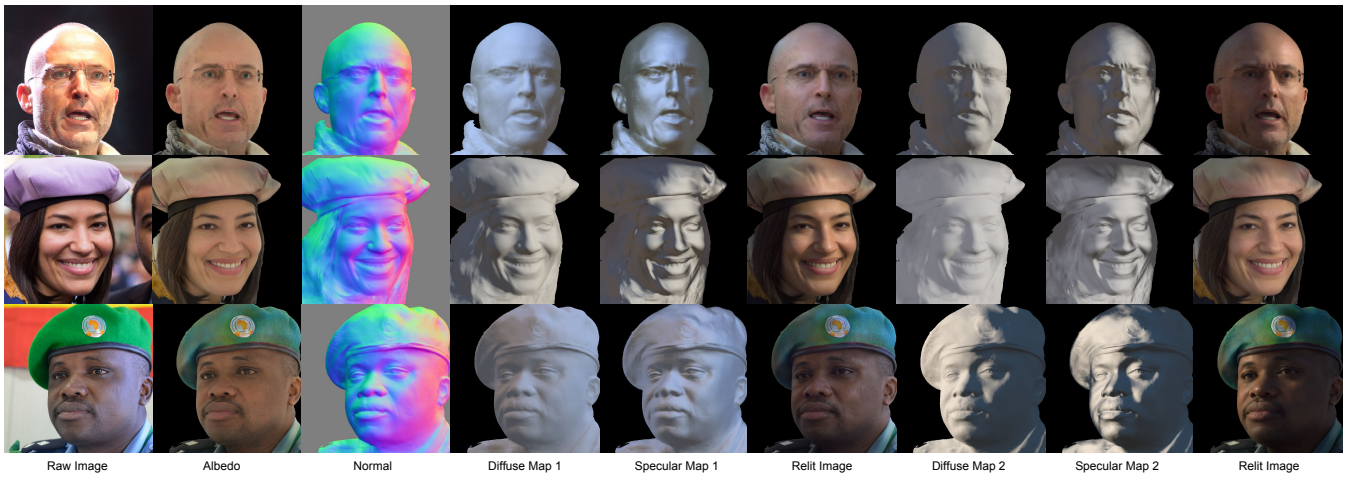

Figure 1: Relighting augmentation on FFHQ [Karras et al. 2019] using [Pandey et al. 2021] to generate albedo, normal, shading, and relit images with different HDRI Relighting, which supervise the training via adversarial losses. (License information: 1. GEPA-10011644021 from Special Olympics 2017 (CC0); 2. Carolina's Graduation Ceremony from Richard Isla Rodas (Public Domain Mark); 3. 2016_08_03_Ugandan_IPO's-6 from AMISOM Public Information (CC0))

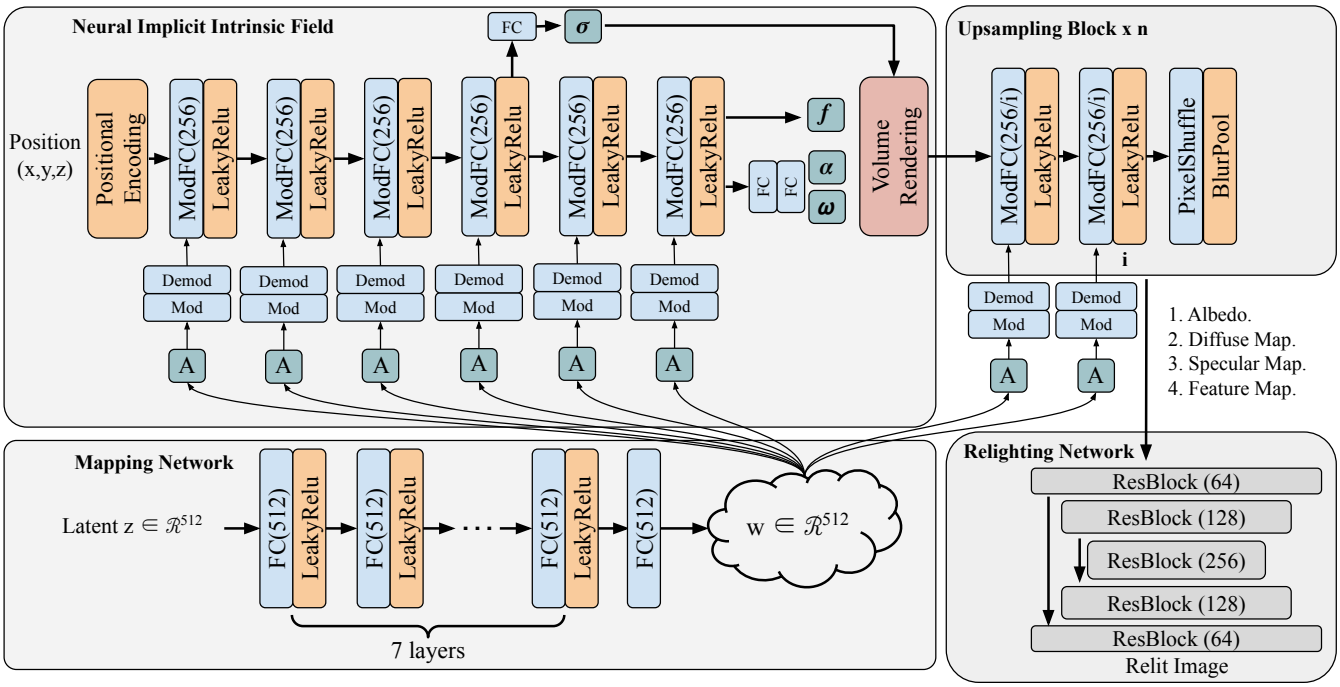

Figure 2: Proposed architecture of our neural generator, which consists of a neural implicit intrinsic field network, upsampling blocks, a relighting network and a mapping network.

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. And we provide the target HDRI images in Figure 3 for comparing our method with ShadeGAN [Pan et al. 2021] and pi-GAN [Chan et al. 2021b] coupled with Total Relighting [Pandey et al. 2021].

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 [Gu et al. 2021]. 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 well preserved, which is a significant improvement over the state-of-the-art [Pan et al. 2021].

3.3 Rotate Camera and Lighting

We show four 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 VIDEO

We provide a supplementary video to show animated rendering results. In the video, we show 1) our intermediate intrinsic results and final relighting results in a continuous camera trajectory, 2) comparison on the relighting faithfulness to ShadeGAN [Pan et al. 2021] under rotating HDRI, using image based relighting with a Light Stage [Guo et al. 2019] 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-GAN [Chan et al. 2021b] + TotalRelighting [Pandey et al. 2021].

5 LIMITATIONS AND FUTURE WORK

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 [Brock et al. 2019; Karras et al. 2019, 2020] 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. Our upsampling and relighting convolutional modules could be contributors to this problem since they do not guarantee multi-view consistency in the generated details. We mitigate this problem through several design choices, e.g. path regularization and conservative convolution layer which reduce this inconsistency. Besides, this also 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 [Chan et al. 2021b; Pan et al. 2021], which in turn are significantly better than 2D based generative view synthesis methods [Abdal et al. 2021; Mallikarjun et al. 2021; Tewari et al. 2020].

The rendering speed of our model at inference is about 0.3 seconds on a NVIDIA V100 GPU, which is not applicable for real-time applications. But this could be improved by adapting more efficient 3D representations, e.g. the tri-plane representation [Chan et al. 2021a] and ray-based representation [Sitzmann et al. 2021]. With efficient representations, we could also produce higher-resolution rendering.

Furthermore, aliasing effects are noticeable when changing viewpoints especially around the teeth and hair. An approach similar to [Barron et al. 2021] 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 video). 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

Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. 2021. StyleFlow: Attribute-Conditioned Exploration of StyleGAN-Generated Images Using Conditional Continuous Normalizing Flows. *ACM Trans. Graph.* 40, 3, Article 21 (May 2021), 21 pages. https://doi.org/10.1145/3447648

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. arXiv:2103.13415 [cs.CV]

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2021a. Efficient Geometry-aware 3D Generative Adversarial Networks. In *arXiv*.

Figure 3: Qualitative comparisons on CelebA with ShadeGAN [Pan et al. 2021] and pi-GAN [Chan et al. 2021b] + portrait relighting [Pandey et al. 2021]. Note that HDR maps for each example are visualized to the left of relit images

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021b. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5799–5809.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. StyleNeRF: A Style-based 3D-Aware Generator for High-resolution Image Synthesis. *arXiv preprint arXiv:2110.08985* (2021).

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

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

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

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

B R Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. 2021. PhotoApp: Photorealistic Appearance Editing of Head Portraits. *ACM Transactions on Graphics* 40, 4 (2021), 1–16.

Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. 2021. A Shading-Guided Generative Implicit Model for Shape-Accurate 3D-Aware Image Synthesis. In *NeurIPS*.

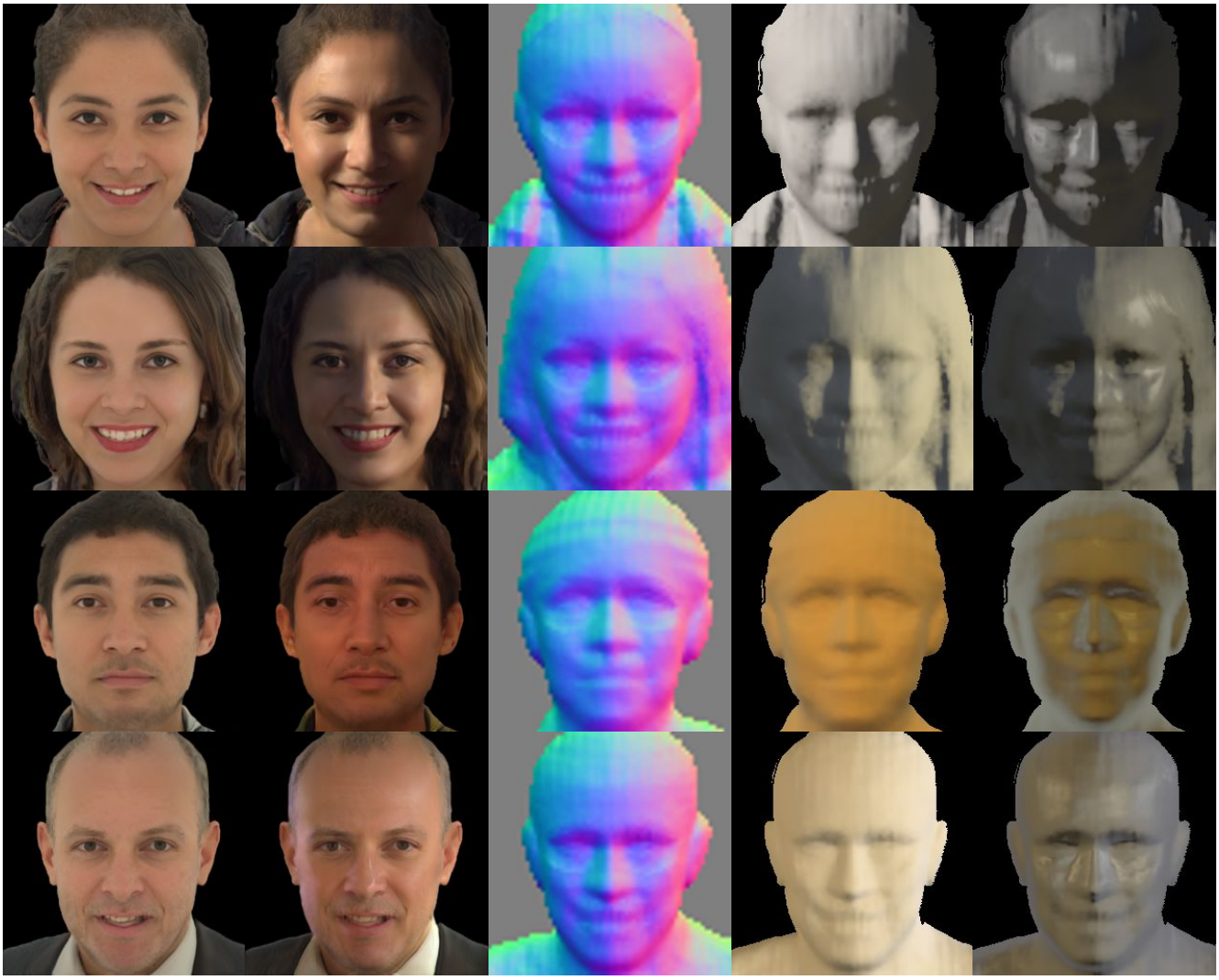

Figure 4: Results of intermediate intrinsic images from our model trained on FFHQ [Karras et al. 2019]. From left to right, we show the albedo, relit image, normal map, diffuse map and specular map.

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

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. 2021. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems* 34 (2021).

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020. StyleRig: Rigging StyleGAN for 3D Control over Portrait Images, CVPR 2020. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Greg Zaal, Sergej Majboroda, and Andreas Mischok. 2020. HDRI Haven. https://www.hdrihaven.com/. Accessed: 2021-11-13.

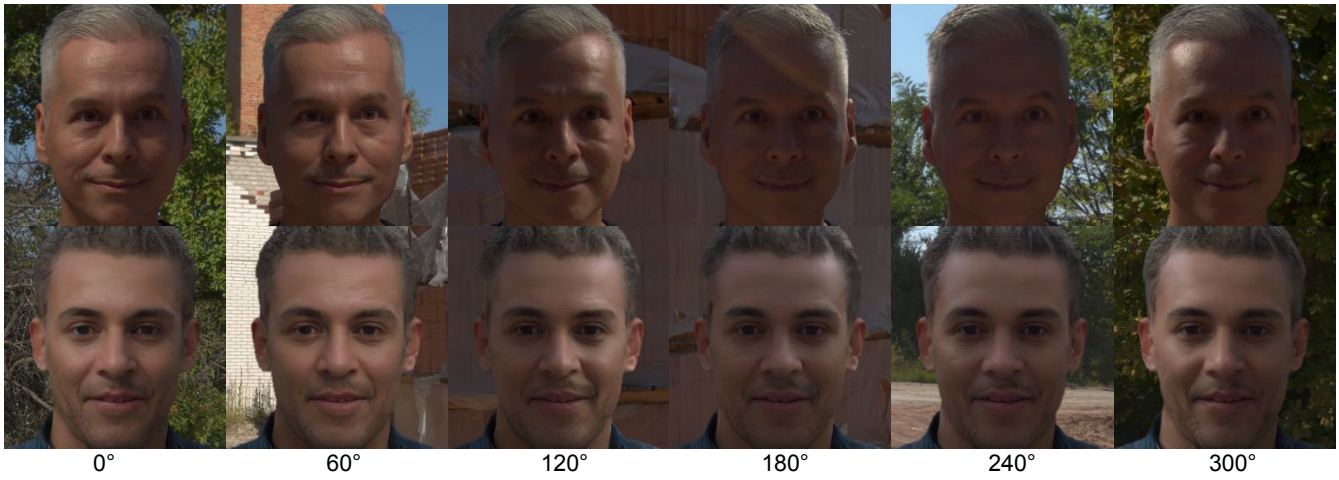

Figure 5: We compare our relighting result to image based relighting (IBR) using a Light Stage [Guo et al. 2019] with the same HDRI illumination. Note that our method produces consistent and plausible shading, soft shadows and specularities.

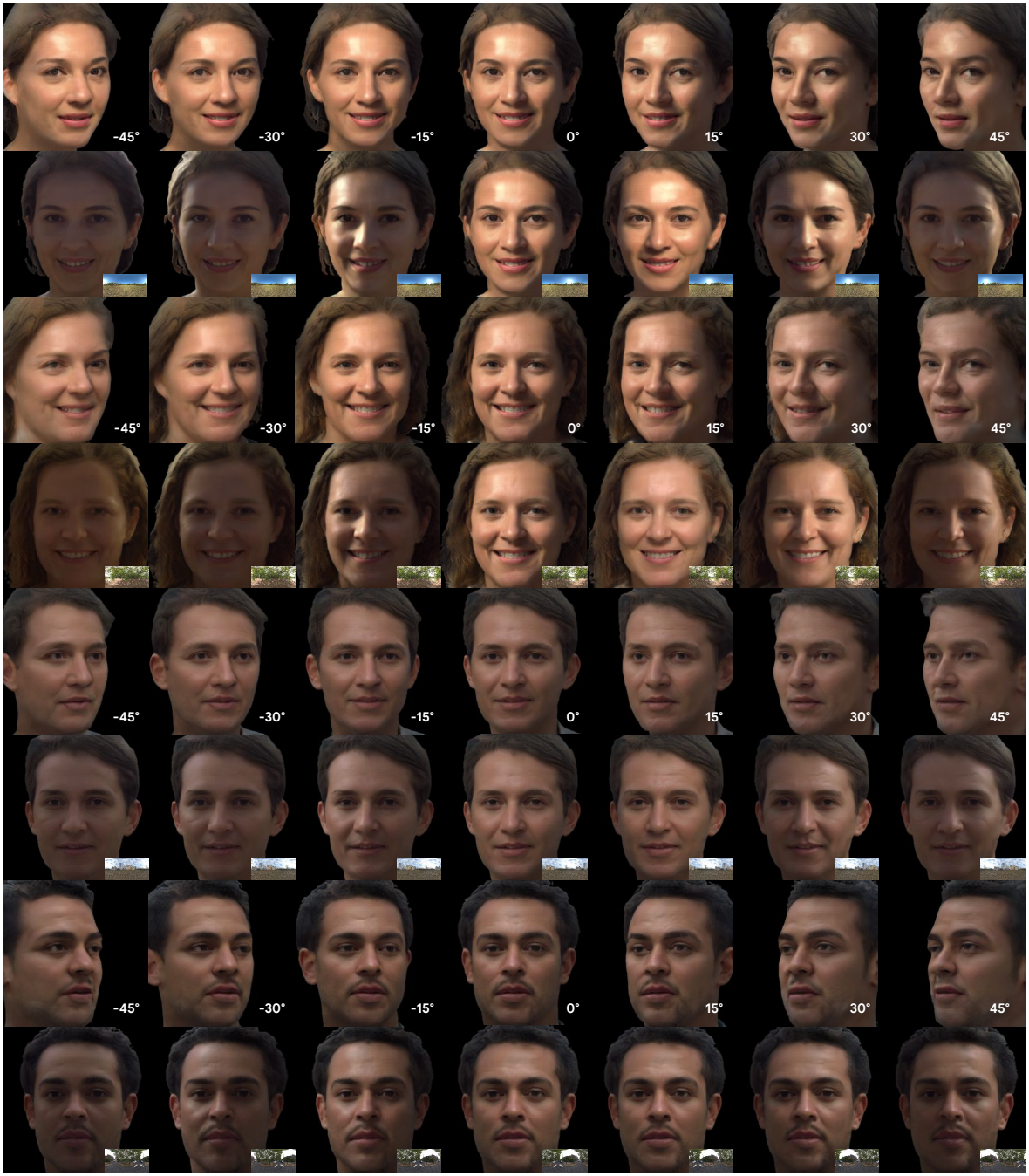

Figure 6: More synthesized images under rotating camera or rotating lighting. Note the relighting consistency and view-dependent effects.