Real-time 3D-aware Portrait Video Relighting

Supplementary Material

Ziqi Cai[1,2] Kaiwen Jiang[3] Shu-Yu Chen[1] Yu-Kun Lai[4] Hongbo Fu[5,6] Boxin Shi[8,9] Lin Gao[*1,7]

[1]Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences
[2]Beijing Jiaotong University [3]University of California San Diego [4]Cardiff University [5]City University of Hong Kong
[6]The Hong Kong University of Science and Technology [7]University of Chinese Academy of Sciences
[8]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[9] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

{zqtsai,kevinjiangedu}@gmail.com, chenshuyu@ict.ac.cn, Yukun.Lai@cs.cardiff.ac.uk
hongbofu@cityu.edu.hk, shiboxin@pku.edu.cn, gaolin@ict.ac.cn

A. Additional Results

A.1. Additional Qualitative Results

Figure B provides additional qualitative results generated by our relighting technique from portrait videos. We encourage readers to refer to the supplementary video for a more comprehensive visualization of our method.

A.2. Additional Comparisons

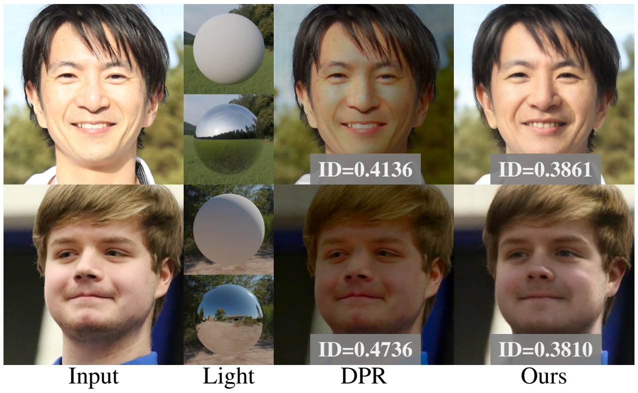

Figure A. Additional qualitative comparison between DPR and ours. Our method consistently delivers portraits with more faithful adherence to desired lighting conditions and enhanced realism.

We conduct a quantitative evaluation for both DPR [13] and our method. It should be noted that DPR demonstrates a considerably lower Fréchet Inception Distance (FID) and a higher Identity Perseverance (ID). Specifically, the FID is 14.98 for DPR and 45.08 for our method, whereas the ID is 0.8531 for DPR and 0.7711 for our method. However, these metrics alone do not unequivocally indicate that DPR generates more realistic results, as evidenced in Figure A. The observed differences in FID and ID metrics can be attributed to DPR's underlying approach, which operates on the LAB color format. Specifically, DPR selectively modifies only the L channel while keeping the A and B channels unchanged. This strategy results in pixel-wise aligned outputs that closely match the original input. This pixel-wise alignment may contribute to the lower FID and higher ID. However, this approach may lead to a loss of color diversity and realism. Additionally, fine-grained details, particularly color-dependent features, may be inadequately captured, as it neglects changes in the A and B channels.

B. Implementation Details

Shading Encoder. The shading encoder appends three layers of Convolutional Neural Network (CNN) with LeakyReLU activation on top of two StyleGAN2 blocks, enabling the synthesis of shading tri-planes directly conditioned on both an albedo tri-plane and a specified lighting condition.

Temporal Consistency Network. The proposed Temporal Consistency Network leverages a combination of self-attention within a branch and cross-attention across two branches to enhance temporal consistency in processing tri-plane sequences. The network is specifically designed to operate on n pairs of albedo tri-planes and shading tri-planes, where we empirically set n to be 5 for optimal performance. The architecture of the Temporal Consistency Network comprises four 8-headed transformers with 4 layers, with each transformer block having a hidden size of 512. To introduce non-linearities and enhance the expressive power of the network, additional CNN blocks with ReLU activation are placed before and after the transformer-based processing. These CNN blocks employ a kernel size of 1×1.

Superresolution Module. We augment the superresolution module [1] by incorporating an extra convolutional neural network (CNN) conditioned on a predicted albedo tri-plane. In contrast to approaches such as [9], we refrain from fine-tuning the backbone model. This strategic choice not only saves time but also mitigates the risk of potential model degradation during the fine-tuning process.

Data Preprocessing. For calculating camera pose, we employ the method proposed by Deng *et al*. [2]. We imple-

[*]Corresponding author is Lin Gao

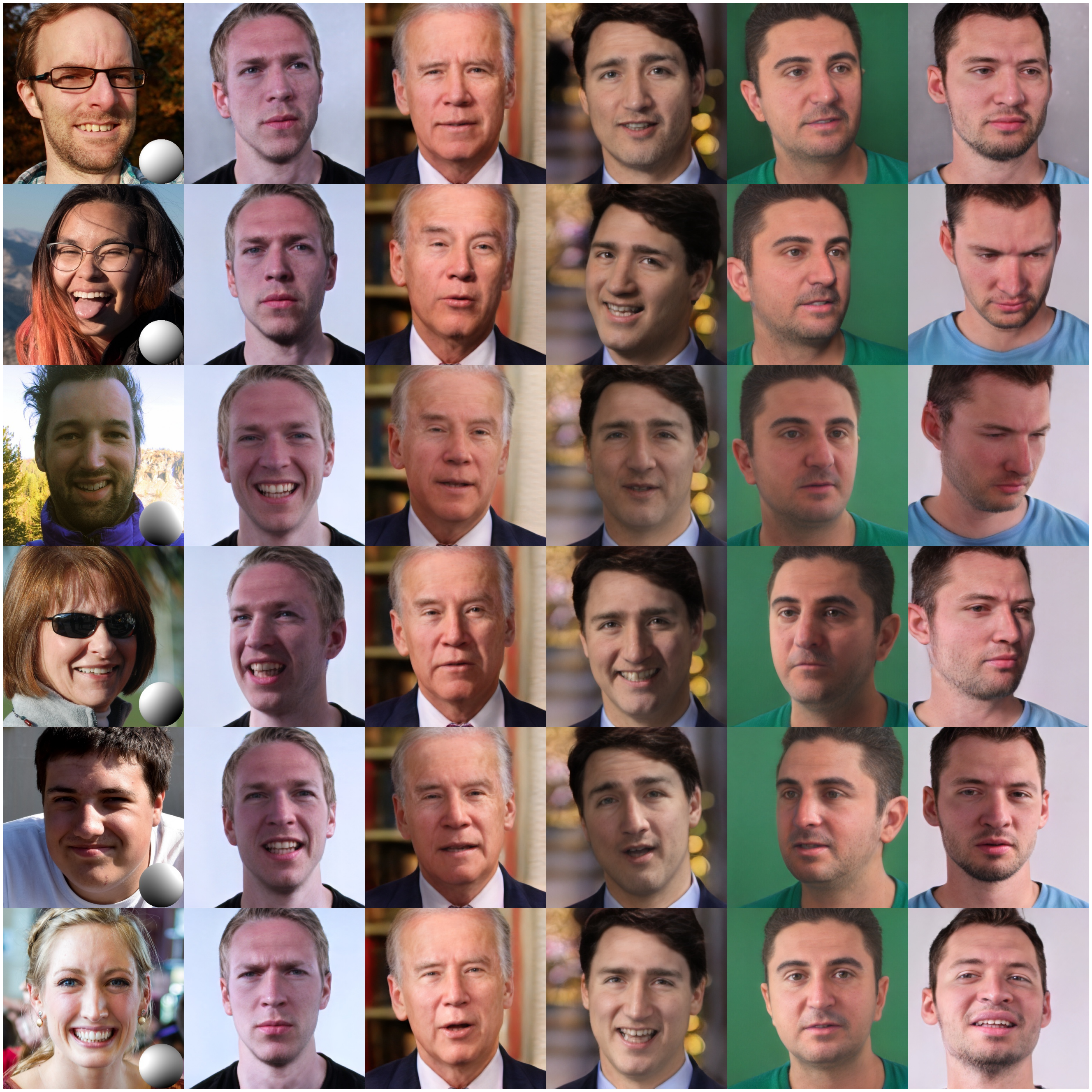

Figure B. Additional qualitative results obtained by applying our relighting technique to an input video through a lighting transfer task. The reference image is displayed in the leftmost column, serving as a visual anchor, while the subsequent columns exhibit the corresponding frames under distinct lighting conditions.

ment an exponential smoothing technique on the detected five facial landmarks to mitigate errors introduced during the keypoint detection process. This smoothing is applied before image cropping and camera pose calculation.

Data Augmentation. In training the tri-plane dual-encoders, we employ camera augmentation techniques similar to those outlined in [9]. To train the temporal consistency network, we randomly select two camera poses and perform interpolation between them to simulate consecutive frames.

Training. In the initial training stage, we adopt settings from [9], exclusively activating the albedo branch. This means that the model learns the intertwined representation within the albedo tri-planes. In the subsequent stage, we follow a fine-tuning approach inspired by [4]. Here, we activate the shading branch to distill shading information

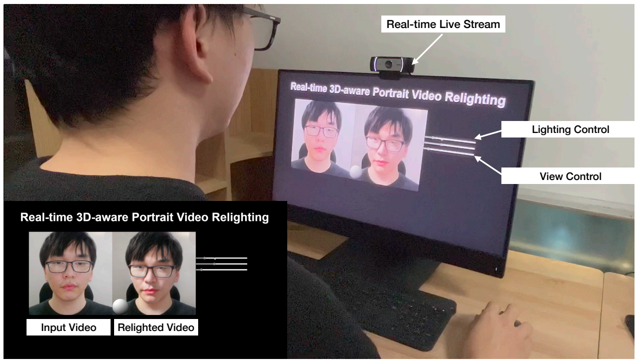
Real-time Live Stream
Real-time 3D-aware Portrait Video Relighting
Lighting Control
View Control
Real-time 3D-aware Portrait Video Relighting
Input Video
Relighted Video

Figure C. Our method is able to lift in-the-wild live stream to relightable 3D faces. Captured portrait at each frame is reconstructed, and rendered under a novel view and with a custom lighting condition for demonstration.

from the entangled albedo tri-planes. This refinement ensures that the albedo tri-planes exclusively contain albedo information while the shading tri-plane encoder acquires valuable insights into shading decomposition. To train the albedo encoder, we freeze the shading tri-plane and replace the predicted shading tri-planes with the corresponding ground truth. After 32K training iterations, we freeze the albedo tri-plane encoder, substituting the predicted albedo with ground truth to exclusively train the shading encoder for an additional 32K iterations. In the final stage, we jointly train the albedo and shading tri-planes for 1.5M iterations. Furthermore, we unfreeze the albedo and shading decoders, along with the super-resolution module, with the intention of improving image quality in the overall system. After the model converges, we freeze the dual-encoder and train the temporal consistency network for 32K iterations. This multi-stage training process allows for a nuanced and comprehensive refinement of the model's capabilities.

Inference. We leverage the benefits of mixed precision and torch.compile across all network components during inference, excluding the patch embedding layer of the Vision Transformer (ViT), tri-plane decoders, and volume rendering. We sample 96 depth points per ray following EG3D [1]. Our model exhibits efficient resource utilization, consuming less than 4GB of GPU memory.

C. Evaluation Details

C.1. Baselines

For all the baselines, we use official codes and pre-trained checkpoints.

To construct our baselines using PTI [8], we modify the official code release (https://github.com/danielroich/PTI) to suit EG3D. We optimize one generator for each video clip, which involves iterating over the latent code in $\mathcal{W}+$ space of each frame for 500 iterations, followed by fine-tuning the generator for a duration equivalent to 10 times the number of frames. For PTI on NeRFFaceLighting, we use the official code release and a pre-trained encoder model (https://github.com/IGLICT/NeRFFaceLighting), which initiates the optimization process using the output latent code in $\mathcal{W}$ space from the pre-trained encoder as a starting point. The latent code is optimized for 500 iterations, and the generator is fine-tuned for another 500 iterations. Furthermore, the spherical harmonic coefficients are optimized for an additional 100 iterations.

For DPR [13], we use the official code release and a pre-trained model (https://github.com/zhhoper/DPR).

For SMFR [3], we use the official code release and a pre-trained model (https://github.com/andrewhou1/Shadow-Mask-Face-Relighting).

For ReliTalk [7], we use the official code release (https://github.com/arthur-qiu/ReliTalk) to preprocess the dataset and subsequently conduct training. To ensure consistency, we tailor the training epochs according to the video length, aligning them with the example provided by the author, *i.e.*, more epochs for shorter videos. This meticulous adjustment results in an identical number of training iterations.

For comparison with SIPR-W [10], TR [6], NVPR [12], and Lumos [11], we apply our method to the input provided by the authors of Lumos and then compare our output images with those respectively provided by the authors of SIPR-W, TR, NVPR, and Lumos. The comparison is also demonstrated in the accompanying video for clear evaluation. We observe a misalignment in the provided environment maps due to different coordinate conventions. To address this, we rotate environment maps by 90 degrees (counter-clockwise when viewed from the positive Z-axis) for alignment. However, a slight misalignment persists, as our coordinate system is constructed on the front of the human face, whereas others use a world coordinate system. To rectify this, we further adjust the environment map by considering the yaw angle of the human face, ensuring correct lighting direction alignment. Additionally, we re-normalize the extracted spherical harmonic (SH) coefficients to maintain consistency across comparisons. To ensure alignment, we recrop their outputs and utilize the background masking technique from [5] on our results.

D. Discussion

Ethical Considerations. While our method provides innovative capabilities in manipulating the viewpoint and lighting conditions of a portrait video clip, it is essential to acknowledge the potential for misuse. To counteract this, using advanced image analysis tools, like fake image detectors and image watermarking, can help detect and prevent deceptive practices.

E. Application

Live-stream Video Relighting System As shown in Figure C, we introduce a real-time system to lift live-stream video into relightable 3D faces. Users are allowed to freely adjust the camera parameters and lighting conditions. A live demonstration is recorded and shown in the accompanying video. We run our system on two NVIDIA GeForce RTX 3090 GPUs, and achieve 20 fps for rendering one view due to the preprocessing and transmission overhead. This performance ensures that users can seamlessly and interactively relight and render their faces under novel views.

References

[1] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 12, 14

[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. 12

[3] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 14

[4] Kaiwen Jiang, Shu-Yu Chen, Hongbo Fu, and Lin Gao. NeRFFaceLighting: Implicit and disentangled face lighting representation leveraging generative prior in neural radiance fields. *ACM Transactions on Graphics*, 42(3), 2023. 13

[5] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. MODNet: Real-time trimap-free portrait matting via objective decomposition. In *Association for the Advancement of Artificial Intelligence*, 2022. 14

[6] 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*, 40(4):1–21, 2021. 14

[7] Haonan Qiu, Zhaoxi Chen, Yuming Jiang, Hang Zhou, Xiangyu Fan, Lei Yang, Wayne Wu, and Ziwei Liu. ReliTalk: Relightable talking portrait generation from a single video. In *International Journal of Computer Vision*, 2024. 14

[8] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1):1–13, 2022. 14

[9] Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. In *ACM Transactions on Graphics*, 2023. 12, 13

[10] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics*, 39(6):1–13, 2020. 14

[11] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics*, 2022. 14

[12] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *International Conference on Computer Vision*, pages 802–812, 2021. 14

[13] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 12, 14