Realistic Dynamic Facial Textures from a Single Image using GANs - Supplementary Material

Kyle Olszewski*[1,3,4], Zimo Li†[1], Chao Yang ‡[1], Yi Zhou§[1], Ronald Yu¶[1,3], Zeng Huang‖[1], Sitao Xiang**[1], Shunsuke Saito††[1,3], Pushmeet Kohli‡‡[2], and Hao Li§§[1,3,4]

[1]University of Southern California
[2]DeepMind
[3]Pinscreen
[4]USC Institute for Creative Technologies

1. Additional Results and Evaluations

More results and evaluations on video sequences of various test subjects can be seen in the supplemental video included with this submission. We show examples of the facial albedo textures that are inferred given the target identity and source expression sequence (with the estimated environmental lighting factored out to allow for relighting the textured faces under different illumination conditions).

The retargeting and compositing results seen in the video demonstrate that, in addition to synthesizing the mouth interior, our system is able to generate subtle wrinkles and deformations in the face texture that are too small to be represented in the mesh used for fitting the face model to the subject, but do indeed enhance the realism of the synthesized sequence of expressions generated for the target image. (We show these sequences slowed down to allow for better visualization of the transient details created for each expression.) Furthermore, we note that the wrinkles generated by this system do not correspond directly to those of the source expressions in the video, but rather vary depending on factors such as the appearance and age of the person depicted in the target image (see, for example, the retargeting to images of Brad Pitt and Angelina Jolie in the section "Retargeting Results and Comparison with Static Texture" in the supplemental video).

Our approach to compositing the final rendered image of the target subject into the source sequence requires that the faces be front-facing. However, we note that our network architecture can synthesize dynamic textures even for non-frontal viewpoints of the source subject, as seen in Fig. 1. Cases showing a frontal source subject animating a non-frontal target subject can be seen in Fig. 2, although we note that in this case there are artifacts in the occluded regions that may be visible in the final animation. Additional retargeting and compositing results can be found in Fig. 4.

2. Implementation, Training and Performance Details

Our networks are implemented and trained using the Torch7 framework, using an NVIDIA Titan GPU to accelerate the training and inference. Fig. 3 shows the loss on the training and validation dataset for both the generative and discriminative networks.

Below we list the average per-frame execution time for each stage in our texture generation and compositing pipeline.

While our implementation of the 3D face model fitting approach described in Section 4 is implemented on the CPU and does not run in realtime, we note that [Thies et al. 2016] demonstrate that such an approach can be implemented in parallel on the GPU to achieve realtime performance. Furthermore, the mouth interior synthesis approach described in Section 5.4 of the paper is implemented in Matlab using a single thread for processing, and thus could be further optimized using parallel processing. Thus, while the approach used for replacing the faces in the source video sequence with the rendered target faces is not designed to run in realtime, it should be possible to further optimize the other

*olszewski.kyle@gmail.com (equal contribution)
†zimoli@usc.edu (equal contribution)
‡harryyang.hk@gmail.com (equal contribution)
§zhou859@usc.edu
¶ronaldyu@usc.edu
‖zenghuan@usc.edu
**sitaoxia@usc.edu
††shunsuke.saito16@gmail.com
‡‡pushmeet@google.com, project conducted while at MSR
§§hao@hao-li.com

stages of the pipeline to run at interactive framerates.

1. 3D face model fitting (Section 4): 5.6 seconds
2. Texture inference (Section 5): 12 milliseconds.
3. Mouth interior synthesis (Section 5.4): 156 milliseconds.
4. Video face replacement (Section 6): 4.5 seconds.

3. Acknowledgements

First, we would like to thank all participants who were videotaped. This research is supported in part by Adobe, Oculus & Facebook, Huawei, Sony, the Google Faculty Research Award, the Okawa Foundation Research Grant, and the U.S. Army Research Laboratory (ARL) under contract W911NF-14-D-0005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon.

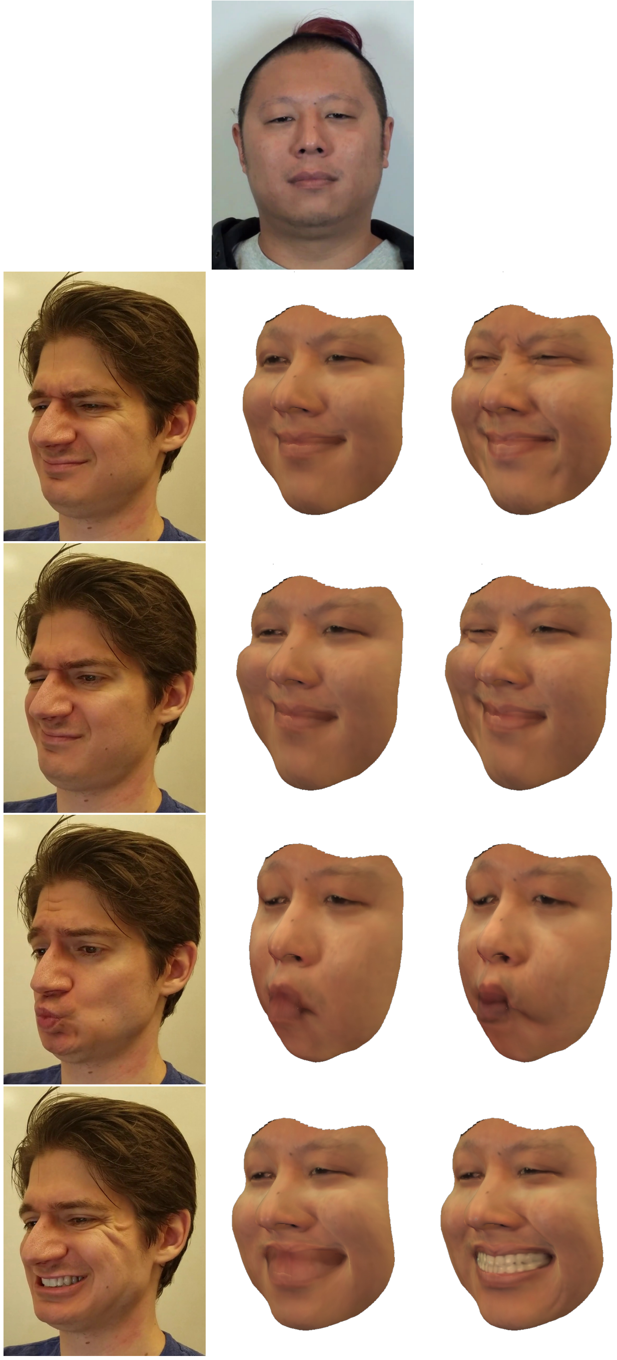


Figure 1. Non-frontal face reenactment. The top row displays the target image. In the remaining rows, from left to right: the source image, the rendered static texture and the rendered dynamic texture. We can see that the dynamic texture contains more subtle details such as wrinkles, resulting in a more expressive and plausible image of the target subject. Also note that the synthesized mouth interior results in much more plausible renderings when the mouth is open.

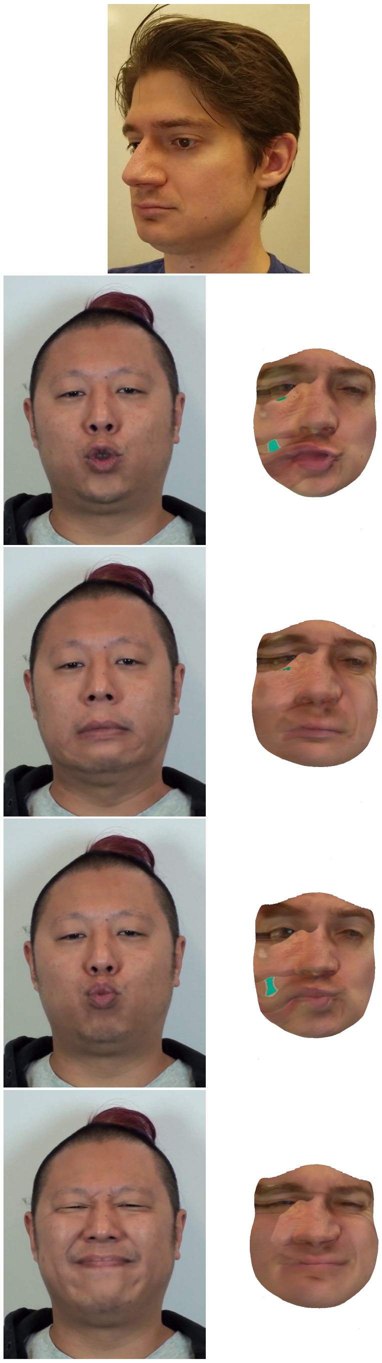
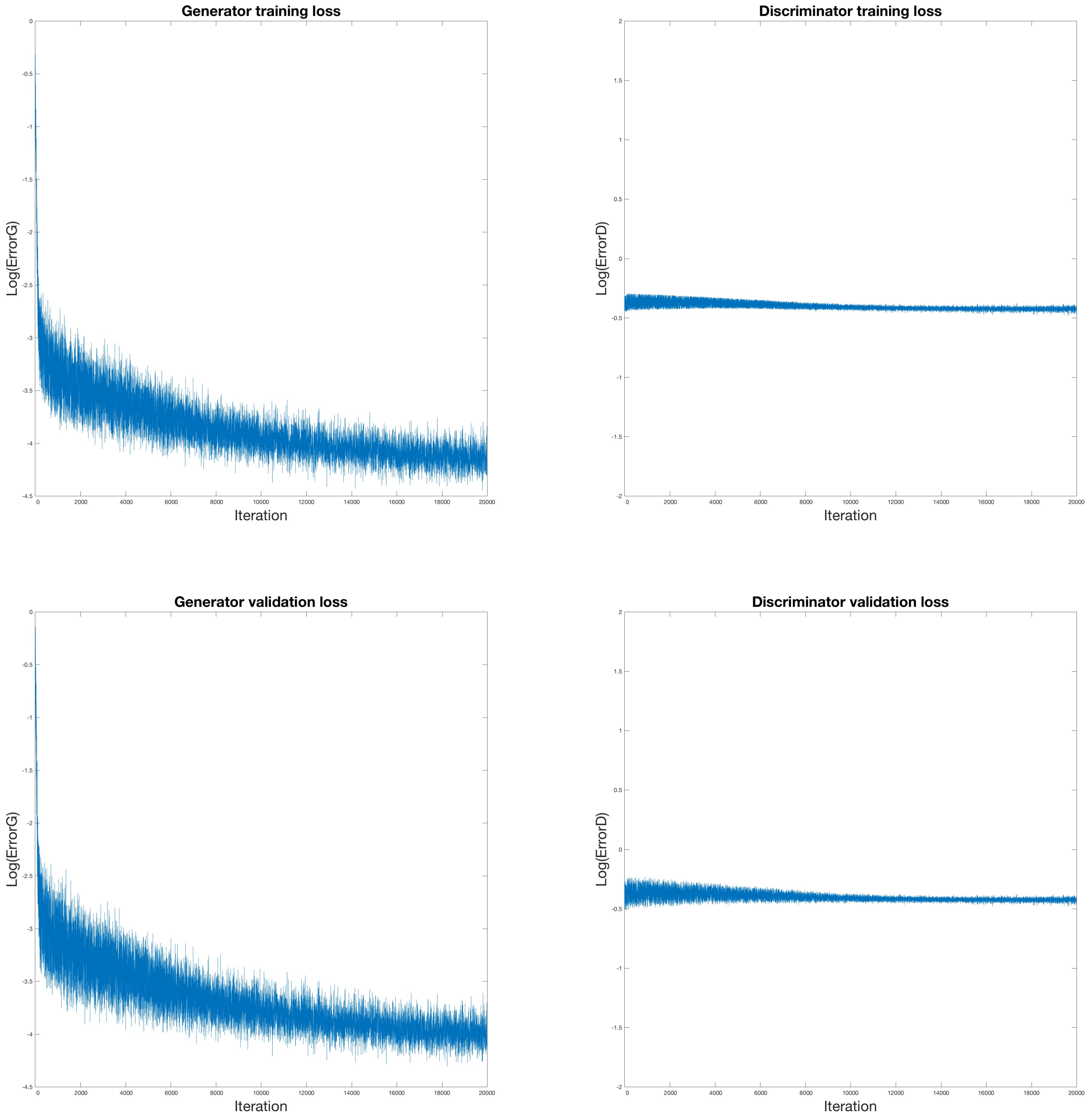


Figure 2. Failure cases induced by an extreme non-frontal target image. The top row displays the target image. The remaining rows display the source expression image (left) and the rendered image with the synthesized dynamic texture (right). While we can synthesize details for the visible region of the target image, the occluded regions contain artifacts that are visible when the image is rendered with these regions visible to the camera.



Generator Error

Discriminator Error

Figure 3. Generative and discriminative training loss (top row) and validation loss (bottom row).

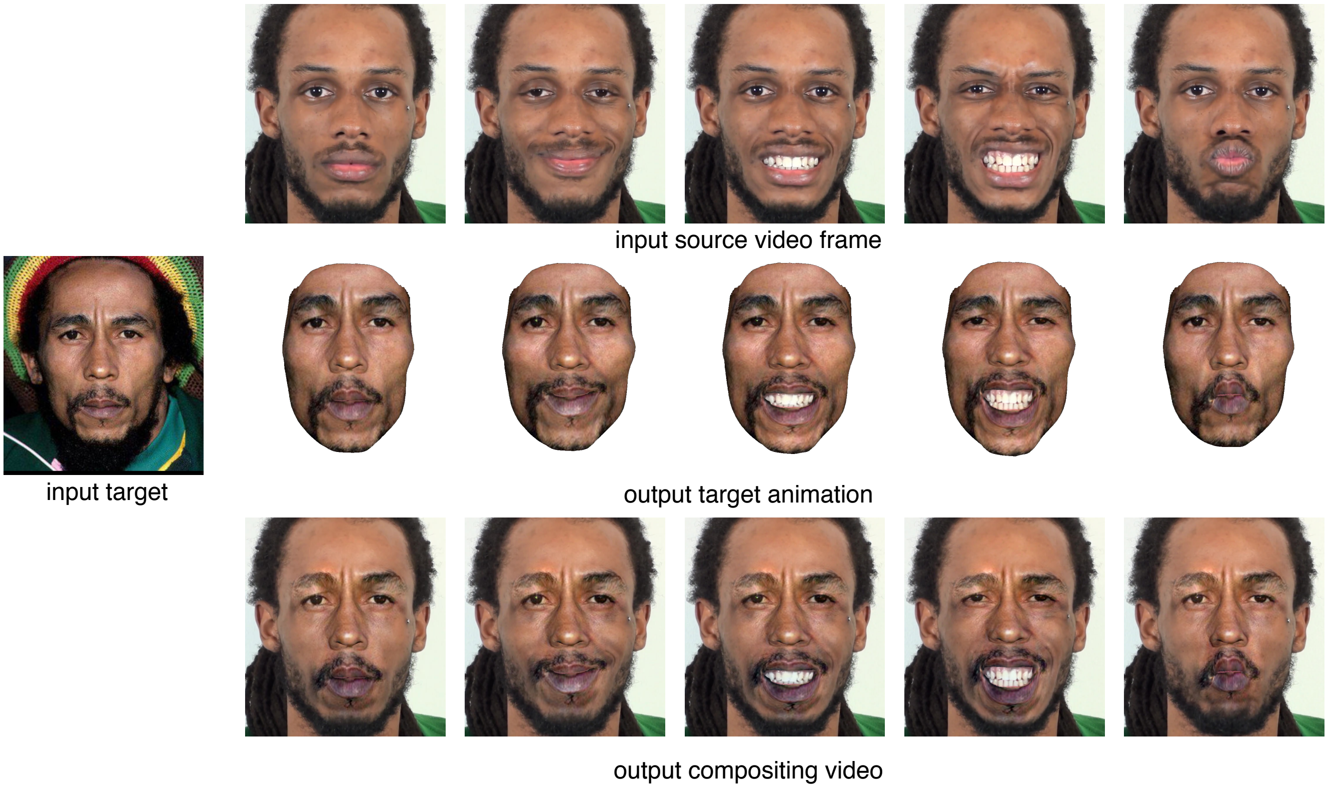


Figure 4. Additional retargeting and compositing results.