StyleRig: Rigging StyleGAN for 3D Control over Portrait Images –Supplementary Material–

Ayush Tewari[1] Mohamed Elgharib[1] Gaurav Bharaj[2] Florian Bernard[1]
Hans-Peter Seidel[1] Patrick Pérez[3] Michael Zollhöfer[4] Christian Theobalt[1]

[1]MPI Informatics, Saarland Informatics Campus [2]Technicolor [3]Valeo.ai [4]Stanford University

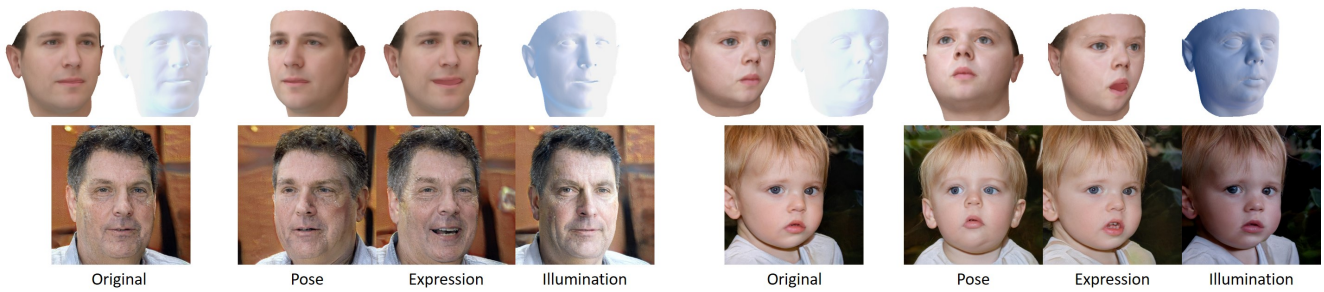

Figure 1: StyleRig allows for face rig-like control over StyleGAN generated portrait images, by translating semantic edits on 3D face meshes to the input space of StyleGAN.

In this supplemental document, we provide further training details and evaluations. We strongly recommend to watch the supplementary video for more editing results.

1. Training Details

We use $\lambda_{land} = 17.5$ for pose editing, $\lambda_{land} = 100.0$ for expression editing and $\lambda_{land} = 7.8$ for illumination editing networks. The same hyperparameters are used for both the editing and consistency losses. When we train networks for simultaneous control, we weight the loss functions for the different parameters differently. Rotation losses are weighted by 1.0, expression by 1000.0 and illumination by 0.001. As before, the weights for both the editing and the consistency losses are equal.

We do not edit the translation of the face. We noticed that the training data for StyleGAN was cropped using facial landmarks, such that there is a strong correlation between the head rotation and the translation parameters. Thus, even when training networks to edit other parameters, we do not try to preserve the translation component, for eg., the face is allowed to translate while rotating.

2. Evaluation of Simultaneous Parameter Edits

As mentioned in the paper, we can also train networks to edit all three sets of parameters (pose, expression and illumination) simultaneously using a single network. As shown in the results section of the main paper as well as the supplemental video, this produces high quality results. To compare the simultaneous editing performance to networks

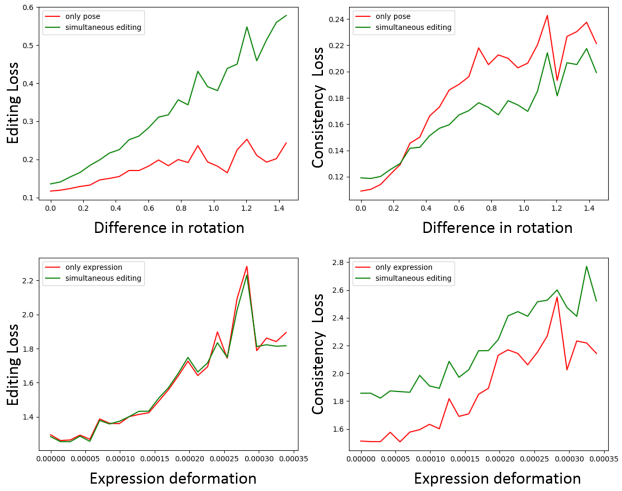

Figure 2: Comparison of models trained to edit individual parameters and the model trained to edit all parameters simultaneously.

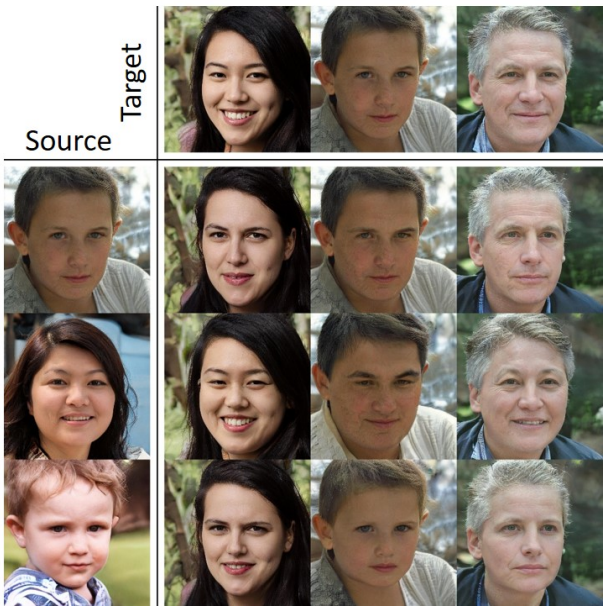

Figure 3: We can also transfer the identity geometry of source images to the target using StyleRig.

that have been trained for editing just a single parameter, we plot the editing and consistency losses with respect to the magnitude of edits in Fig. 2. These numbers are computed for 2500 parameter mixing results on a test set. Rotation difference is measured by the magnitude of the rotation angle between the source and target samples in an axis-angle representation. Expression difference is computed as the ℓ_2 difference between the mesh deformations due to expressions in the source and target samples. All losses are lower when the edits are smaller and increase gradually with larger edits. For the rotation component, the editing loss for the network trained for simultaneous control increases faster. This implies that this network is worse at reproducing the target pose, compared to the network trained only for pose editing. For expressions, while the editing loss remains similar, the consistency losses are higher for the network with simultaneous control. This implies that the network with only expression control is better at preserving other properties (pose, illumination, identity) during editing.

3. Geometry Editing

Similar to rotation, expression and illumination, we can also control the identity geometry of faces using the identity component of the 3DMM. Fig. 3 shows several geometry mixing results, where the source geometry can be transferred to the target images.

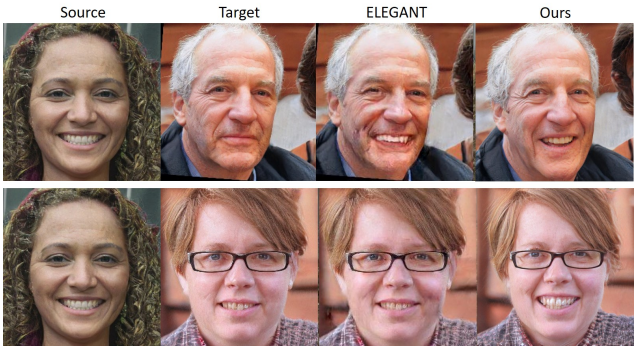

Figure 4: Comparison to ELEGANT [2]. Source expressions are transferred to the target images. We obtain higher quality results, and a better transfer of the source expressions.

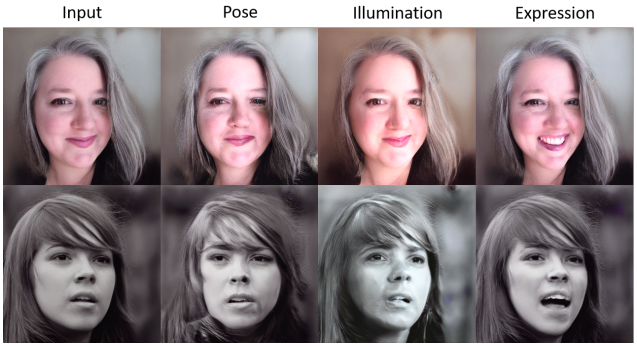

Figure 5: StyleRig can also be used for editing real images. We first optimize the latent embedding of StyleGAN of an input image using Image2StyleGAN [1]. RigNet is then used to edit the result. In some cases such as the bottom row, this leads to artifacts since the optimized latent embedding can be far from the training data.

4. Comparison

We compare our approach to ELEGANT [2], a GAN-based image editing approach. Source expressions are transferred to the target images. We obtain higher-quality results with fewer artifacts. We can also better transfer the source expressions to the target.

5. Editing Real Images

Our method can also be extended for editing real images. We use the recent Image2StyleGAN approach [1] to compute the latent embedding, given an existing real image. RigNet can then be used to compute the edited embedding, thus allowing for editing high-resolution images, see Fig. 5. However, in some cases, such as Fig. 5 (bottom), this approach can lead to artifacts in the edited results, since the embedding optimized using Image2StyleGAN might be outside the training distribution used for training RigNet.

Figure 6: Limitations: Transformations not present in the training data cannot be produced. Thus, our method cannot handle in-plane rotation and asymmetrical expressions.

6. Limitations

We show some failure cases in Fig. 6. As explained in the main paper, in-plane rotations can not be produced by our approach. Expressions other than mouth open/smiling are either ignored or incorrectly mapped. As detailed in the main paper, we attribute these problems to a bias in the training data that has been used for training StyleGAN. In addition, we cannot control high-frequency details in the image, since our employed differentiable face reconstruction network only reconstructs coarse geometry and appearance.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to embed images into the stylegan latent space? In *International Conference on Computer Vision (ICCV)*, 2019.

[2] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.