Supplementary Material for CoLa-SDF: Controllable Latent StyleSDF for Disentangled 3D Face Generation

Rahul Dey[1,2] Bernhard Egger [3] Vishnu Naresh Boddeti[2] Ye Wang[1]
Tim K. Marks[1]

[1]Mitsubishi Electric Research Laboratories (MERL)
[2]Michigan State University (MSU)
[3]Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg

1. Overview

We organize the supplementary as follows. We first present ablation studies involving our model compared against possible variants in Sec. 2. Next, we highlight further directional illumination editing in Sec. 3.1, simultaneous pose-attribute control in Sec. 3.2, and additional physical attribute transfer experiments in Sec. 3.3. Finally, we discuss further implementation details in Sec. 4.

2. Ablation Studies

The development of CoLa-SDF involved a number of important design choices. To show the effects of some of these choices, these ablation studies demonstrate how various omissions from or additions to CoLa-SDF detract from its overall performance.

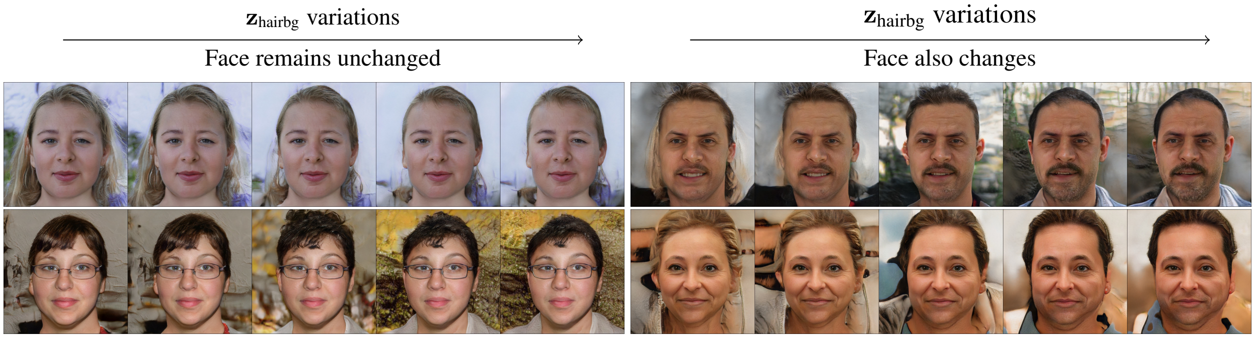
2.1. Without face consistency loss

CoLa-SDF-NoFaceLoss: As described in Sec. 4.3.1 in the main paper, we enforce hair/background disentanglement through a combination of the hair/background consistency loss $\mathcal{L}_{\text{hairbg}}$ and the face consistency loss $\mathcal{L}_{\text{face}}$. The hair/background consistency loss, $\mathcal{L}_{\text{hairbg}}$, ensures that when we keep the hair/background code $\mathbf{z}_{\text{hairbg}}$ the same but change the other latent codes, the hair/background regions in the image will change as little as possible. Similarly, $\mathcal{L}_{\text{face}}$ ensures that when we change $\mathbf{z}_{\text{hairbg}}$ but keep the other latent codes the same, the face region will change as little as possible.

To study the importance of the face consistency loss, we train a model without $\mathcal{L}_{\text{face}}$ loss and evaluate it in terms of its hair/background disentanglement. We call this variant CoLa-SDF-NoFaceLoss. Specifically, we perform interpolation between two hair/background codes while keeping all the other latent codes the same. If the model has well disentangled hair/background from the face region, changing $\mathbf{z}_{\text{hairbg}}$ should not affect the face. We show the comparison between our model and CoLa-SDF-NoFaceLoss in Fig. 1. Note that, with CoLa-SDF-NoFaceLoss, changing $\mathbf{z}_{\text{hairbg}}$ changes facial-hair in the first row, and causes shape changes in the second row. On the other hand, with our model, changing $\mathbf{z}_{\text{hairbg}}$ does not any cause noticeable changes in the face regions.

2.2. Independent mapping of each attribute:

CoLa-SDF-SeparateMappers: In CoLa-SDF, the five latent codes $\mathbf{z}_\alpha$, $\mathbf{z}_\tau$, $\mathbf{z}_\gamma$, $\mathbf{z}_{\text{hairbg}}$, and $\mathbf{z}_{\text{rest}}$ all feed into the same Renderer Mapping Network, which outputs a combined style code $\mathbf{w}$, as shown in the top left of Fig. 2 in the main paper. For this ablation study, we replace the single volume renderer mapping network with five separate renderer mapping networks, one for each of shape $\boldsymbol{\alpha}$, albedo $\boldsymbol{\tau}$, illumination $\boldsymbol{\gamma}$, hair/background $\mathbf{z}_{\text{hairbg}}$, and $\mathbf{z}_{\text{rest}}$. We sample the shape parameters from $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, where $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ are the data mean and covariance obtained from MOST-GAN encodings of the FFHQ dataset [3] images. We used the same method to sample $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}$. For sampling $\mathbf{z}_{\text{hairbg}}$ and $\mathbf{z}_{\text{rest}}$, we use the standard normal distribution $\mathcal{N}(0, 1)$. The individual mappers have similar architecture as the original combined renderer mapping network, but with different input and output dimensions. The input dimensions for shape, albedo, illumination, hairbg, and rest are 150, 200, 27, 64, 64, respectively. Their output dimensions are 37, 64, 27, 64, 64, respectively, to match the latent code factorization of CoLa-SDF. We concatenate the shape, albedo, illumination, hairbg, and rest style-codes obtained from these independent mappers to form the combined style code, $\mathbf{w}$, which is then passed through the rest of the algorithm exactly as in CoLa-SDF. Note that we adopt separate mappers only during the volume renderer phase; the gener-


$\mathbf{z}_{\text{hairbg}}$ variations
Face remains unchanged
$\mathbf{z}_{\text{hairbg}}$ variations
Face also changes

(a) CoLa-SDF (With face consistency loss)

(b) CoLa-SDF-NoFaceLoss (Without face consistency loss $\mathcal{L}_{\text{face}}$)

Figure 1. Effect of Face Consistency Loss on Hair/Background Disentanglement. We vary only $\mathbf{z}_{\text{hairbg}}$, with all the other latents same and show the changes in the images generated by two models: (a) Our full model CoLa-SDF (on the left), and (b) Our model without face consistency loss CoLa-SDF-NoFaceLoss (on the right). Without face consistency loss $\mathcal{L}_{\text{face}}$, the hair/background latent code does not get fully disentangled from the face. This leads to changes in the face region as the hair/background code $\mathbf{z}_{\text{hairbg}}$ is varied, as can be seen in (b).

	Ablation Variants			
	Separate Mappers	With Perceptual Consistency	With Photometric Consistency	Ours
FID (↓)	25.85	23.04	21.38	**19.4**

Table 1. FID evaluations at 256x256 resolution. CoLa-SDF with Separate Mappers performs the worst, while enforcing photometric or perceptual consistency losses also harm the FID scores. Our proposed method scores the best FID scores while maintaining latent space disentanglement.

ator mapping network remains a single network as in CoLa-SDF. This variant, though, results in a loss of image quality and diversity, as evaluated in terms of FID [2] (see Tab. 1).

2.3. Add low-res to high-res consistency loss

CoLa-SDF-Photometric: In this variant, we adopt a photometric consistency loss $\mathcal{L}_{\text{photocons}}$ to enforce consistency between the high-resolution image obtained from the styled generator (after downsampling it to the low-resolution scale), and the low-resolution image obtained from the volume renderer:

$$\mathcal{L}_{\text{photocons}} = ||\text{down}(\mathbf{I}_{gen},\ \text{size}(\mathbf{I}_{vol}) - \mathbf{I}_{vol}||_2^2, \quad (1)$$

where $\text{down}(\mathbf{x}, (h, w))$ downsamples image $\mathbf{x}$ to height h and width w using bilinear interpolation. This acts as an additional loss to ensure that the disentanglement of physical attributes in the volume renderer reflects well in the styled-generator too.

CoLa-SDF-Perceptual: This variant is similar to CoLa-SDF-Photometric, except that we replace the photometric consistency loss with a perceptual consistency loss [7]:

$$\mathcal{L}_{\text{vggcons}} = ||\phi(\text{down}(\mathbf{I}_{gen},\ \text{size}(\mathbf{I}_{vol})) - \phi(\mathbf{I}_{vol})||_2^2, \quad (2)$$

where ϕ is VGGFace [5] model.

We found that, both these variants lead to higher FID metrics as reported in Tab. 1, which is an indicator of low image quality and diversity.

3. Additional Qualitative Results

3.1. Directional Illumination Editing

To further demonstrate the correspondence between CoLa-SDF's illumination latent code and the spherical harmonics coefficients [6], we show controlled illumination manipulation in Fig. 2. Starting from an initial illumination setting (shown in the left column), we project it to the spherical harmonics space using Eq. 1 in the main paper and rotate the lighting around the camera axis in increments of $\pi/5$ radians $(36°)$. The results demonstrate that CoLa-SDF can perform any desired illumination editing.

3.2. Simultaneous Pose and Attribute Control

We show simultaneous pose control and attribute transfer in Fig. 3. These results show the fully disentangled face generation capability of CoLa-SDF in terms of pose, shape, albedo, illumination and hair/background.

3.3. Additional Attribute Transfer

We show additional source-to-target attribute transfer results, including shape transfer in Fig. 4, texture transfer (transfer of both albedo and illumination) in Fig. 5, and hair/background transfer in Fig. 6. In Fig. 6, we again observe that while the hair geometry and style is mainly controlled by the hair/background code, its appearance is partly controlled by the albedo and illumination codes. These

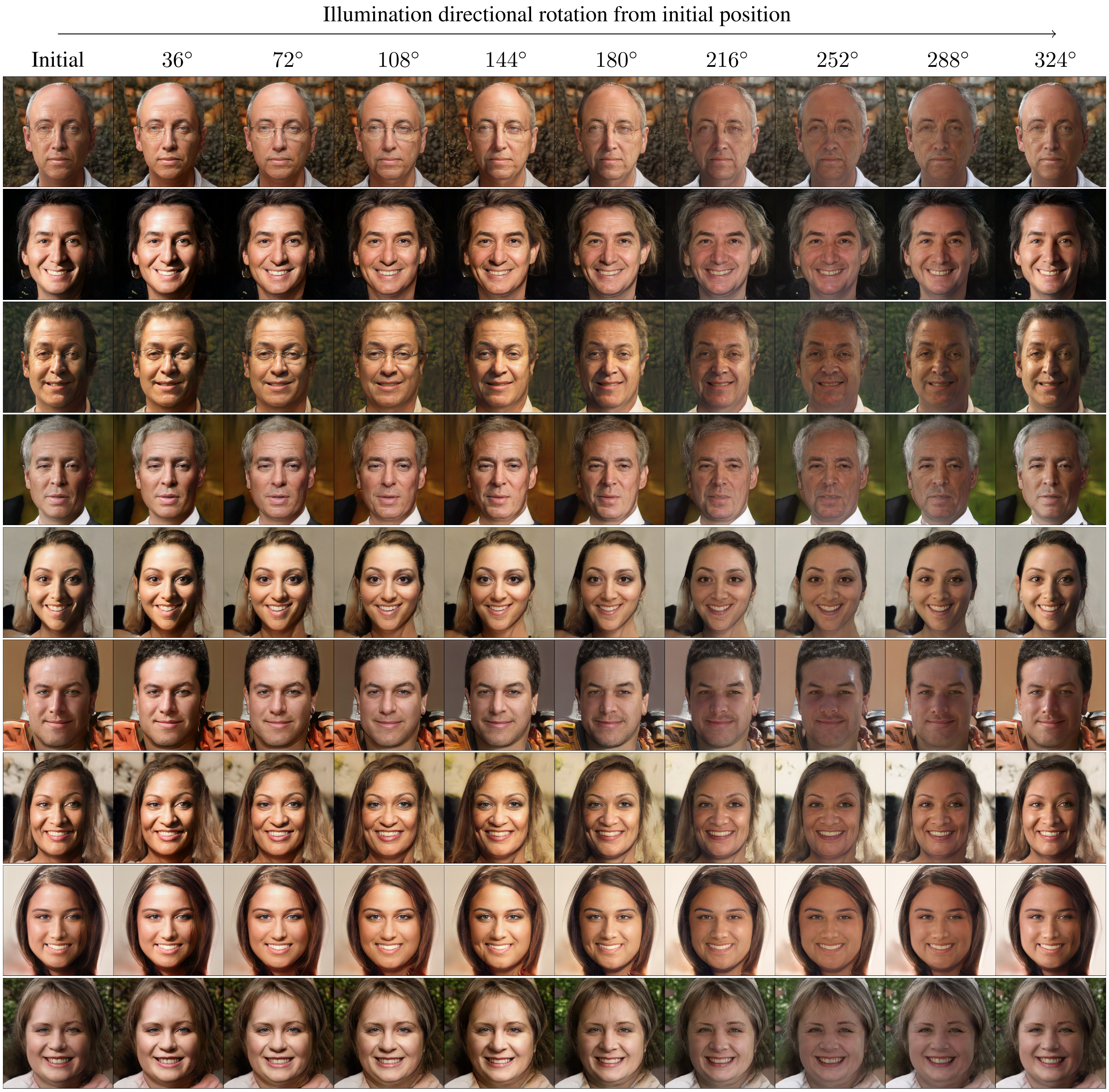
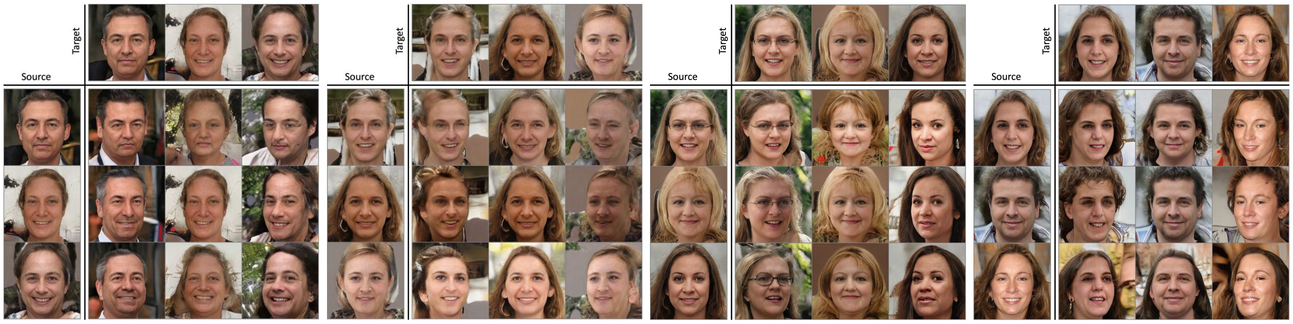


Figure 2. Directional rotation of illumination.

results show CoLa-SDF's ability to transfer one attribute while keeping the rest intact and demonstrate the attribute disentangled latent space learned by CoLa-SDF.

4. Implementation Details

Training Details: We trained CoLa-SDF's volume renderer and styled generator separately for 400,000 and 200,000 iterations, respectively. We trained the volume renderer with a batch-size of 20 and ray-sampling frequency (samples per ray) of 24, on a machine with Intel Xeon Gold 6326 processor with 64 cores and 10 Nvidia A40 GPUs. While training the styled-generator, we freeze the volume renderer and the renderer mapping network and increase the ray-sampling frequency to 64. We trained the styled-generator with a batch-size of 40 on the same machine. Training the volume renderer takes 3 days and the styled-generator takes 4 days on this machine.



(a) CoLa-SDF Shape transfer. (b) CoLa-SDF Albedo transfer. (c) CoLa-SDF Illumination transfer. (d) CoLa-SDF Hair/background transfer.

Figure 3. Simultaneous Pose and Attribute Control

StyleSDF Losses: In the main paper (Eqs. 2, 3), we defined the volume renderer loss $\mathcal{L}_{\text{vol}}$ and the styled-generator loss $\mathcal{L}_{\text{gen}}$ used in StyleSDF [4]. We now define the components $\mathcal{L}_{\text{view}}$, $\mathcal{L}_{\text{eik}}$, and $\mathcal{L}_{\text{surf}}$.

The pose alignment loss $\mathcal{L}_{\text{view}}$ is the smoothed L1 loss between the pose (ϕ, θ) used by the volume renderer to generate images, and the pose $(\hat{\phi}, \hat{\theta})$ predicted by the low-resolution discriminator:

$$\mathcal{L}_{\text{view}} = \begin{cases} (\hat{\theta} - \theta)^2 & \text{if } |\hat{\theta} - \theta| \leq 1 \\ |\hat{\theta} - \theta| & otherwise \end{cases} \tag{3}$$

The eikonal loss enforces physical validity of the signed distance field [1]:

$$\mathcal{L}_{\text{eik}} = \mathbb{E}_{\mathbf{x}} \left(||\nabla d(\mathbf{x})||_2 - 1 \right)^2 . \tag{4}$$

The minimal surface loss penalizes the SDF values that are close to zero to avoid spurious zero-crossings and non-visible surfaces from being formed:

$$\mathcal{L}_{\text{surf}} = \mathbb{E}_{\mathbf{x}} \left(\exp(-100 |d(\mathbf{x})|) \right) . \tag{5}$$

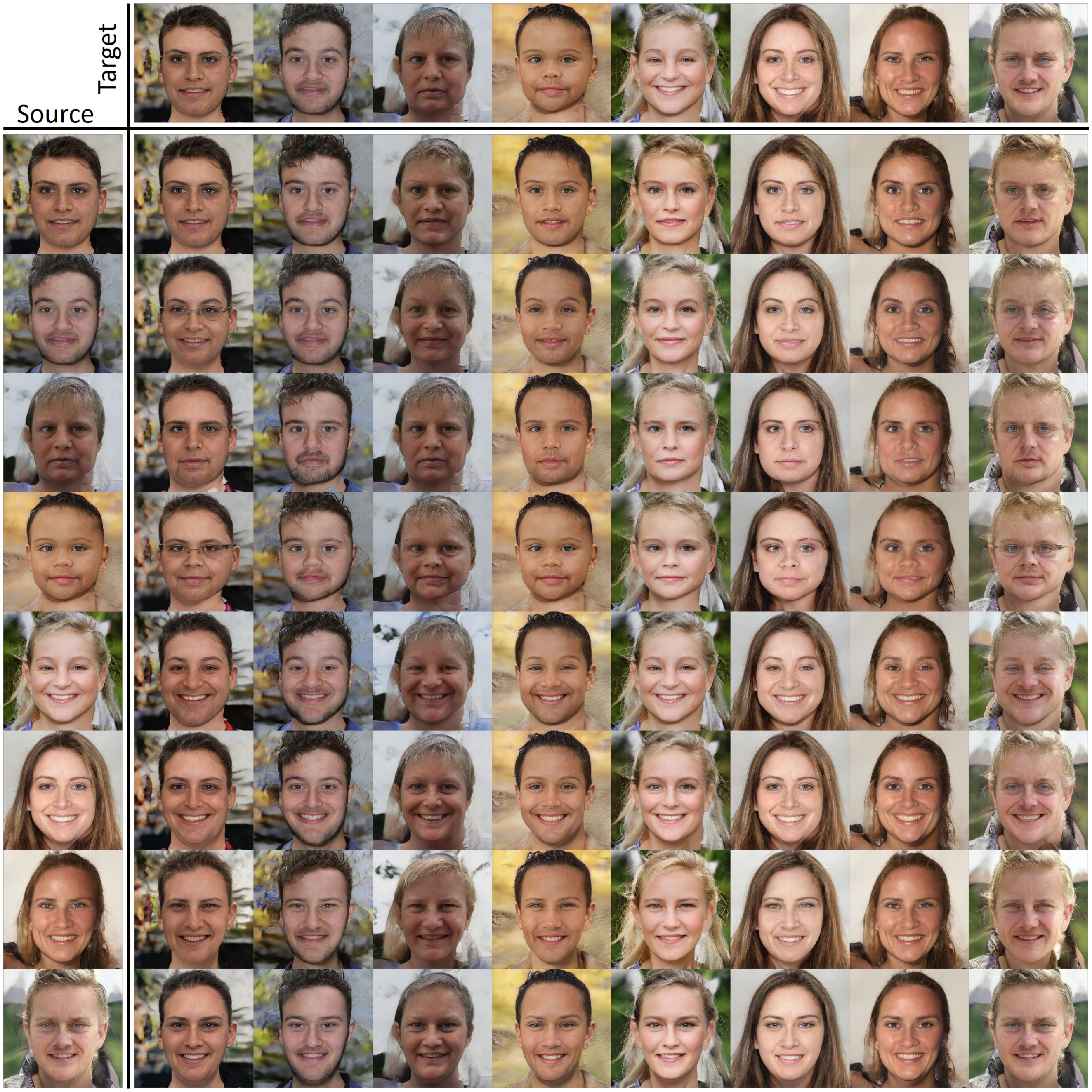


Figure 4. Shape Transfer

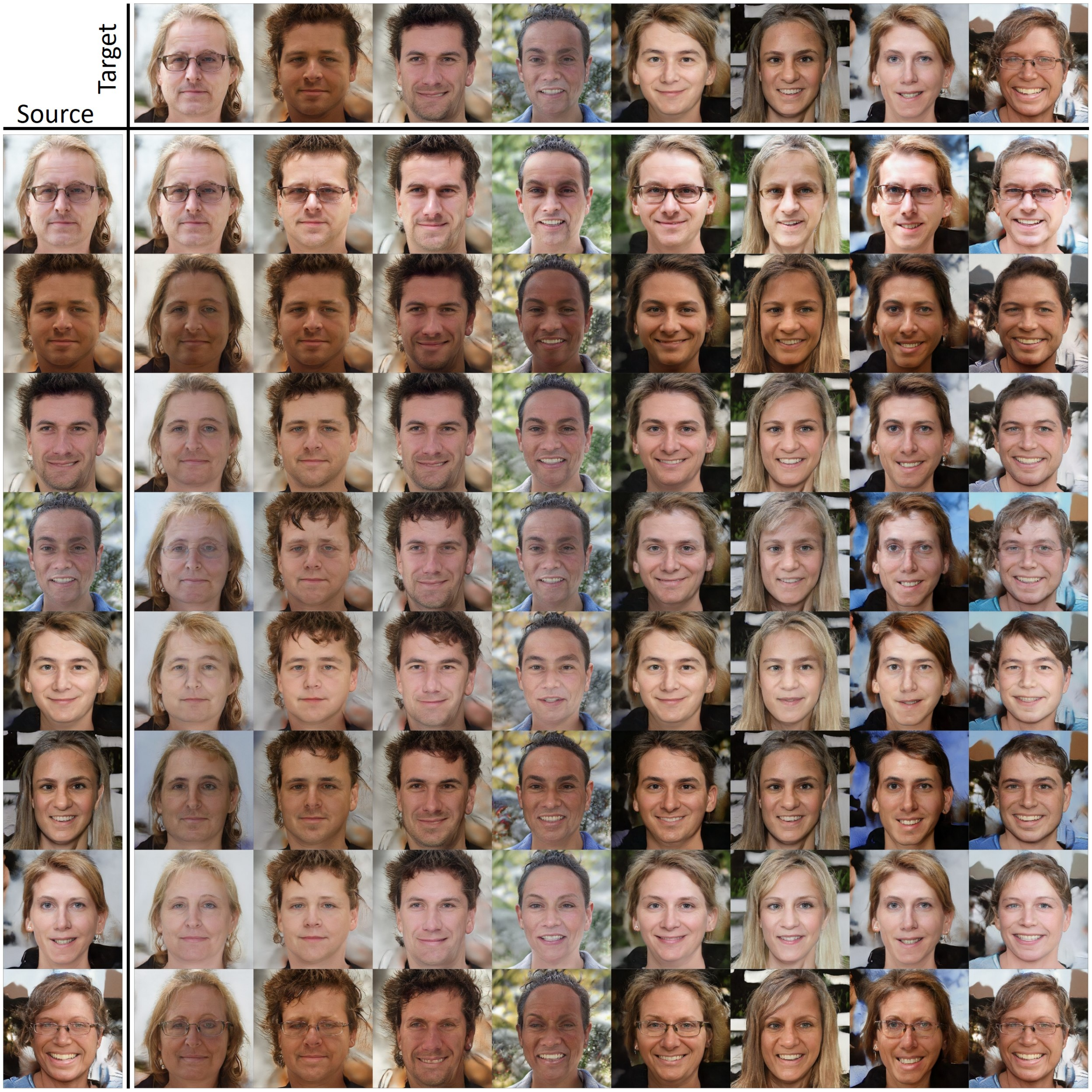


Figure 5. Texture (Albedo + Illumination) Transfer

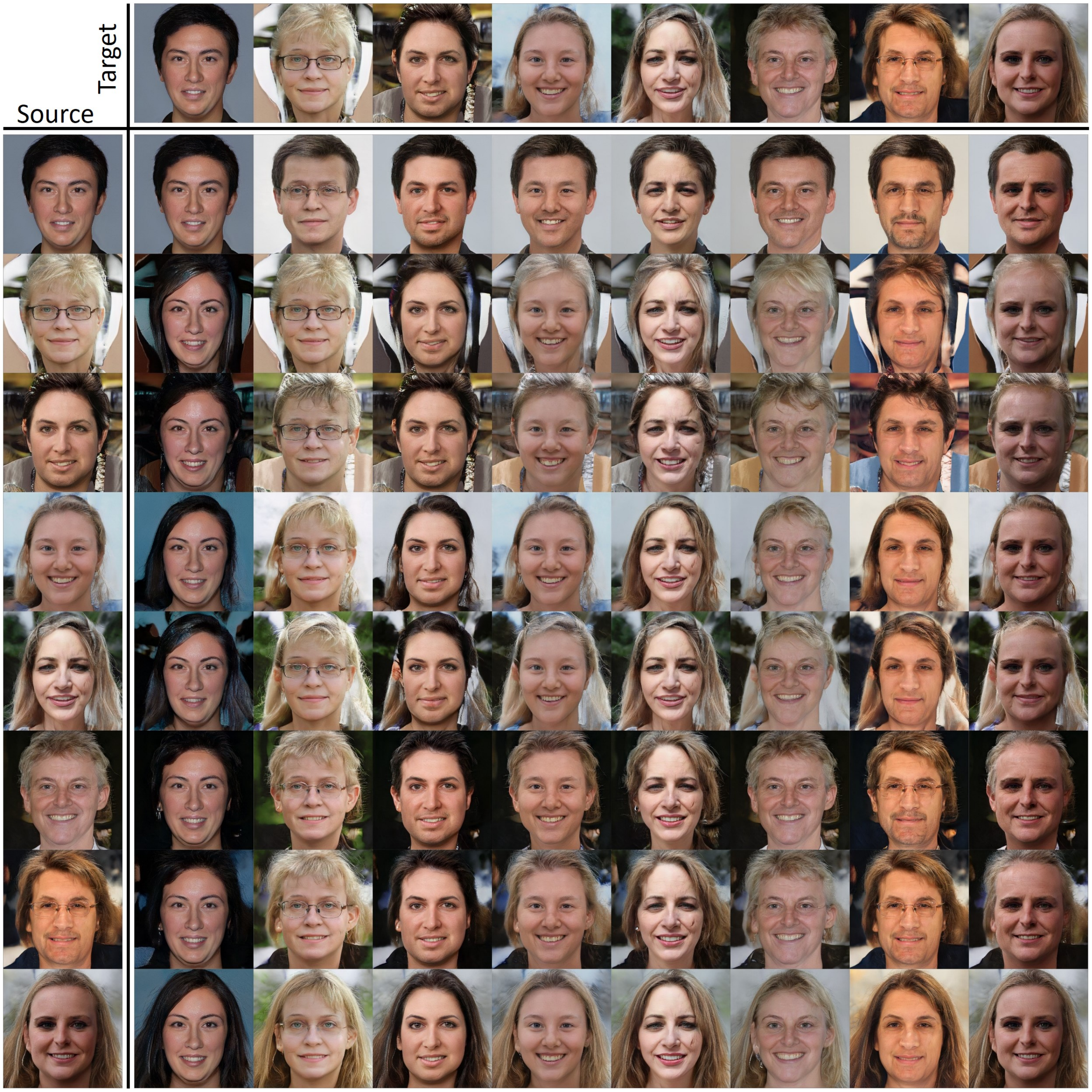


Figure 6. Hair/Background Transfer

References

[1] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 4

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 2

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[4] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 4

[5] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 2

[6] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500, 2001. 2

[7] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2