Supplementary Material — LOHO: Latent Optimization of Hairstyles via Orthogonalization

Rohit Saha[1,2] Brendan Duke[1,2] Florian Shkurti[1,4] Graham W. Taylor[3,4] Parham Aarabi[1,2]

University of Toronto[1] ModiFace[2] University of Guelph[3] Vector Institute[4]

1. Images and Masks

For each selected tuple (I_1, I_2, I_3), we extract hair and face masks using Graphonomy [2]. We separately dilate and erode M_2^h, the hair mask of I_2, to produce the dilated version, $M_2^{h,d}$, and the eroded version, $M_2^{h,e}$. Using $M_2^{h,d}$ and $M_2^{h,e}$, we compute the ignore region $M_2^{h,ir}$. We exclude the ignore region from the background and let StyleGANv2 inpaint relevant features. We want to optimize for reconstruction of I_1's face, reconstruction of I_2's hair shape and structure, transfer of I_3's hair appearance and style, and inpainting of the ignore region. Given a tuple, Figure 1 shows the images and relevant masks used during optimization.

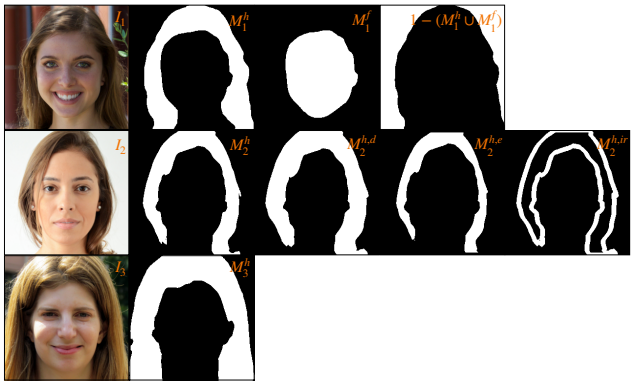

Figure 1: Tuple (I_1, I_2, I_3) and relevant masks used in LOHO.

2. Alignment Metrics

To categorize each selected tuple (I_1, I_2, I_3), we calculate the Intersection over Union (IoU) and pose distance (PD) between face masks, and 68 2D facial landmarks. We extract the masks using Graphonomy [2], and estimate landmarks using 2D-FAN [1].

IoU and PD quantify to what degree two faces align. Given the two binary face masks, M_1^f and M_2^f, we compute IoU as

$$\text{IoU} = \frac{M_1^f \cap M_2^f}{M_1^f \cup M_2^f}. \tag{1}$$

The pose distance (PD), on the other hand, is defined in terms of facial landmarks. Given the two 68 2D facial landmarks, $K_1^f \in \mathbb{R}^{68\times 2}$ and $K_2^f \in \mathbb{R}^{68\times 2}$, corresponding to I_1 and I_2, PD is calculated by averaging the L_2 distances computed between each landmark

$$\text{PD} = \frac{1}{68}\sum_{k=1}^{68} \|K_{1,k}^f - K_{2,k}^f\|_2 \tag{2}$$

where k indexes the 2D landmarks. Therefore, a tuple where I_1 and I_2 are the same person (Figure 2) would have an IoU of 1.0 and PD of 0.0.

3. StyleGANv2 Architecture

StyleGANv2 [3] can synthesize novel photorealistic images at different resolutions including 128^2, 256^2, 512^2 and 1024^2. The number of layers in the architecture therefore depends on the resolution of images being synthesized. Additionally, the size of the extended latent space $\mathcal{W}^+$ and the noise space $\mathcal{N}$ also depend on the resolution. Embeddings sampled from $\mathcal{W}^+$ are concatenations of 512-dimensional vectors w, where $w \in \mathcal{W}^+$. As our experiments synthesize images of resolution 512^2, the latent space is a vector subspace of $\mathbb{R}^{15\times 512}$, i.e., $\mathcal{W}^+ \subset \mathbb{R}^{15\times 512}$. Additionally, noise maps sampled from $\mathcal{N}$ are tensors of dimension $\mathbb{R}^{1\times 1\times h\times w}$, where h and w match the spatial resolution of feature maps at every layer of the StyleGANv2 generator.

4. Effect of Regularizing Noise Maps

To understand the effect of noise map regularization, we visualize noise maps at different resolutions post optimization. When the regularization term is set to zero, we normalize the noise maps to be zero mean and unit variance. This causes the optimization to inject actual signal into the noise maps, thereby causing overfitting. Figure 3 shows that the noise maps encode structural information of the facial re-

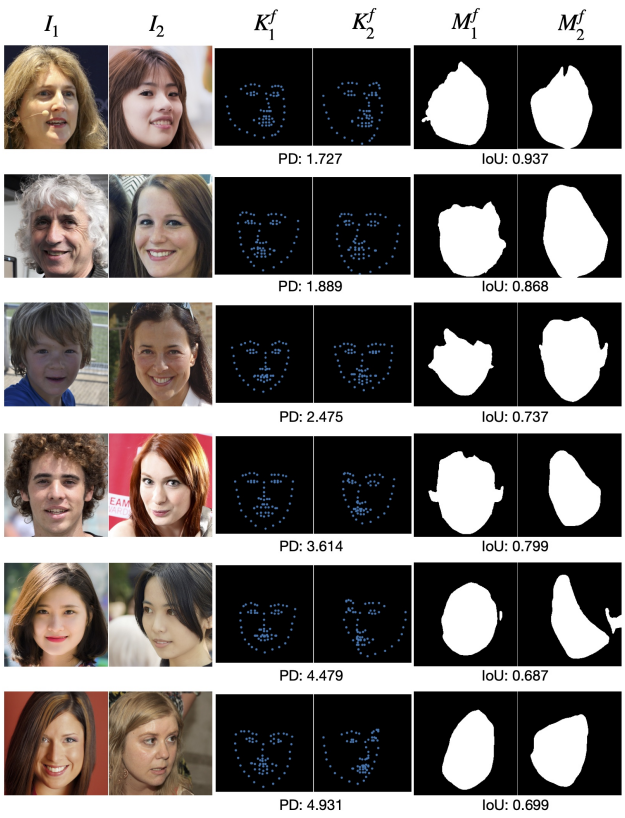

Figure 2: IoU and PD for tuples in each category. **Rows 1-2**: *Easy* tuples. **Rows 3-4**: *Medium* tuples. **Rows 5-6**: *Difficult* tuples.

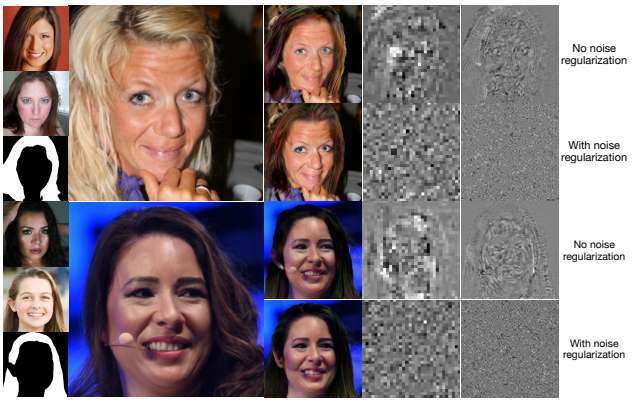

Figure 3: **Effect of regularizing noise maps**. **Col 1 (narrow)**: Reference images. **Col 2**: Identity person. **Col 3**: Synthesized images. **Cols 4&5**: Noise maps at different resolutions.

gion, which is not desirable, and cause the synthesized images to have artifacts in the face and hair regions. Enabling noise regularization prevents this.

Figure 4: **Effect of Gradient Orthogonalization (GO). Rows 1&3**: Reference images (from left to right): Identity, target hair appearance and style, target hair structure and shape. **Rows 2&4**: Pairs (a) and (b), and (c) and (d) are synthesized images and their corresponding hair masks for no-GO and GO methods, respectively. The same holds for pairs (e) and (f), and (g) and (h).

5. Additional Examples of Gradient Orthogonalization

Gradient Orthogonalization (GO) allows LOHO to retain the target hair shape and structure during stage 2 of optimization. Figure 4 shows that no-GO fails to maintain the perceptual structure. On the other hand, GO is able to maintain the target perceptual structure while transferring the target hair appearance and style. As a result, the IoU calculated between M_2^h and M_G^h increases from 0.547 (no-GO, Figure 4 (b)) to 0.603 (GO, Figure 4 (d)). In the same way, the IoU increases from 0.834 (no-GO, Figure 4 (f)) to 0.857 (GO, Figure 4 (h)).

6. Additional comparisons with MichiGAN

We provide additional evidence to show that LOHO addresses blending and misalignment better than MichiGAN [4]. The ignore region $M_2^{h,ir}$ (Figure 1), in addition to StyleGANv2's powerful learned representations, is able to inpaint relevant hair and face pixels. This infilling causes the synthesized image to look more photorealistic as compared with MichiGAN. In terms of style transfer, LOHO achieves similar performance as MichiGAN (Figure 5).

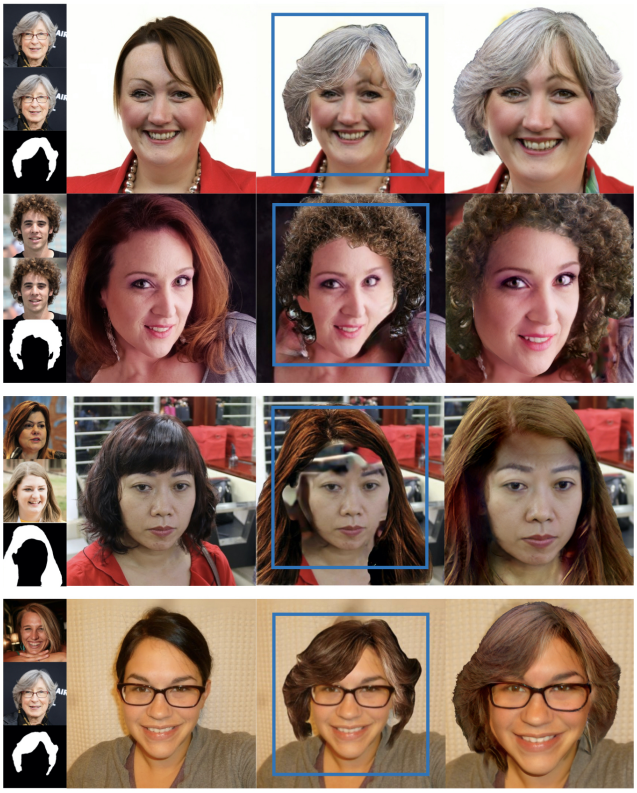

Figure 5: Qualitative comparison of MichiGAN and LOHO. **Col 1 (narrow)**: Reference images. **Col 2**: Identity person **Col 3**: MichiGAN output. **Col 4**: LOHO output (zoomed in for better visual comparison). **Rows 1-2**: MichiGAN "copy-pastes" the target hair attributes while LOHO blends the attributes, thereby synthesizing more realistic images. **Row 3**: LOHO handles misaligned examples better than MichiGAN. **Row 4**: LOHO transfers the right style information.

7. Additional Results of LOHO

We present results to show that LOHO is able to edit individual hair attributes, such as appearance and style (Figure 6), and shape (Figure 7), while keeping other attributes unchanged. LOHO is also able to manipulate multiple hair attributes jointly (Figure 8,9,10).

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[2] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, 2019.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020.

[4] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. Michigan: Multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics (TOG)*, 39(4):1–13, 2020.

Figure 6: **Transfer of appearance and style**. Given an identity image, and reference image, LOHO transfers the target hair appearance and style while preserving the hair structure and shape. **Row 1**: Identity images. **Rows 2-6**: Hair appearance and style references (Cols: $1, 3, 5$), and synthesized images (Cols: $2, 4, 6$).

Figure 7: **Transfer of shape**. Given an identity image, and reference image, LOHO transfers the target hair shape while preserving the hair appearance and style. **Row 1**: Identity images. **Rows 2-6**: Hair shape references (Cols: $1, 3, 5$), and synthesized images (Cols: $2, 4, 6$).

Figure 8: **Multiple attributes editing**. Given an identity image, and reference images, LOHO transfers the target hair attributes.

Figure 9: **Multiple attributes editing**. Given an identity image, and reference images, LOHO transfers the target hair attributes.

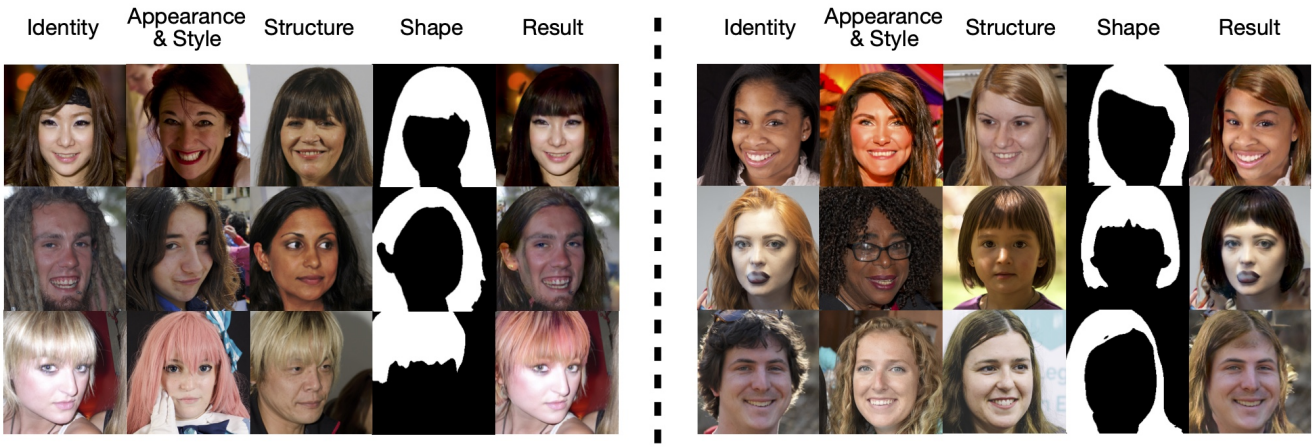

Figure 10: **Multiple attributes editing**. Given an identity image, and reference images, LOHO transfers the target hair attributes.