HairNet: Hairstyle Transfer with Pose Changes

Peihao Zhu[1], Rameen Abdal[1], John Femiani[2], and Peter Wonka[1]

[1] KAUST, Saudi Arabia
[2] Miami University, USA

A Additional Comparisons

In this section, we show additional comparisons of images in our test set. A comparison with LOHO [1] and HairCLIP [2] was not included in the main text since they generally performs worse than Barbershop, but a comparison is included here in Fig. S1. Additional comparisons to other methods are shown in Fig. S2 and Fig. S3. The reader is encouraged to zoom in to look at the details.

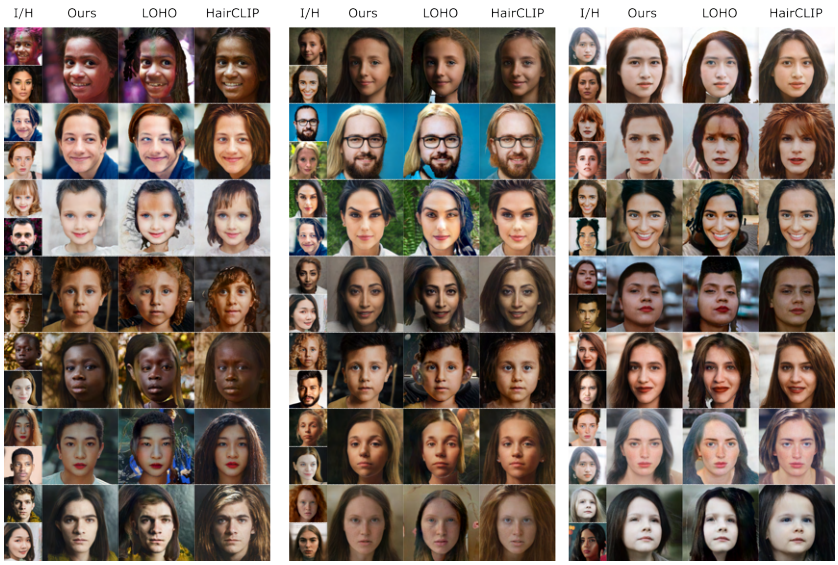

Fig. S1. Comparison with LOHO and HairCLIP. LOHO often produces noticeable artifacts, and HairCLIP cannot preserve the facial identities.

Fig. S2. Additional comparison with prior methods.

Fig. S3. Additional comparison with prior methods.

B Unsupervised Training for HairNet

Our unsupervised training process is described in Section 3.5 of the main paper, however it is sufficiently complex that an overview figure is helpful. In Fig. S4 we highlight various stages of the process used to train hairnet. The architecture of HairNet is shown in detail in Fig. 3 of the main paper, and the same colors are used to show the two stages of HairNet in Fig. S4. Fig. S4(c) and (d) are a visual representation of *both* the mask image for the hair, and an embedding of the face image. The two are multiplied as part of the HairNet architecture.

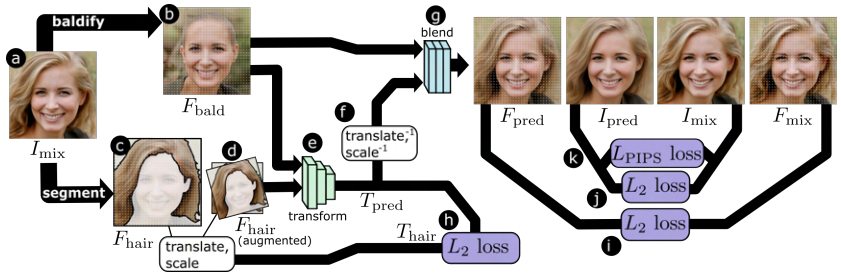

Fig. S4. The unsupervised training pipeline. An image (a) is treated as a perfect *mixed* image, I_{mix}, and a latent-edit is used to baldify the image and create the embedding F_{bald} (b). A segmentation network is used to mask out the hair of F_{mix} in order to produce F_{hair}(c), which is then augmented using translation and scale (d). The hairnet is shown in two parts; the first stage (e) predicts the transformation used to augment the hair image, which is then inverted (f) and input into the blending portion of the network (g). The losses used for training are indicated in violet; (h) guides the net to predict correct spatial transformations, (i) is used to keep the predicted and actual F-codes similar, (j) and (k) guide the training so that the mixed image is reconstructed.

C Ablation of different loss terms

We show the contribution of different loss terms used in HairNet training in Fig. S5.

D Visualizing The Blending of HairNet

We make the claim that our blending network, part of HairNet, makes more intelligent decisions for depth-ordering and dealing with occlusion and disocclusion of the hair. That is, it is capable of deciding if hair should pass in front or behind the ear, face, or shoulder. In addition, it will inpaint regions of hair that were disoccluded by a differently shaped head and neck. These effects are illustrated in Fig. S6. In the column labeled 'w/o Blend' in this figure, hair is

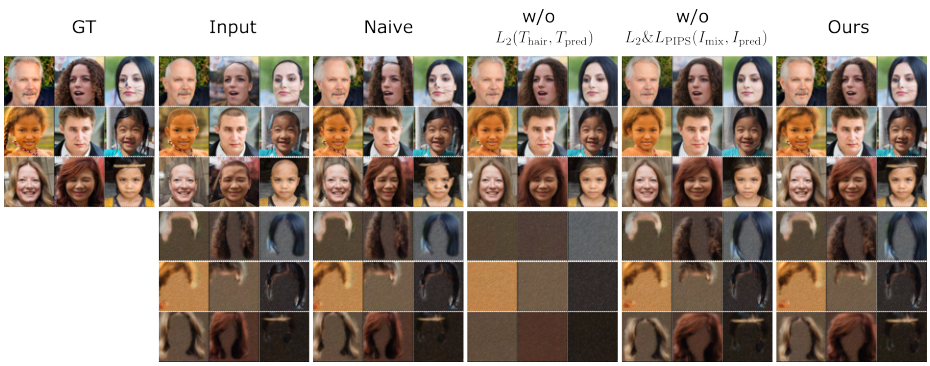

Fig. S5. Baseline hair (col. 3) is misaligned, col. 4 shows that without transform loss, the system is still able to predict the missing hair using only info. "leaked" into the latent code of the bald image, col. 5 shows that detail is missing w/o reconstruction losses.

spatially transformed to align it to the new face, however the F-code of the hair is simply copied and pasted rather than using the blending network. We use yellow ellipses to indicate regions where a decision was made.

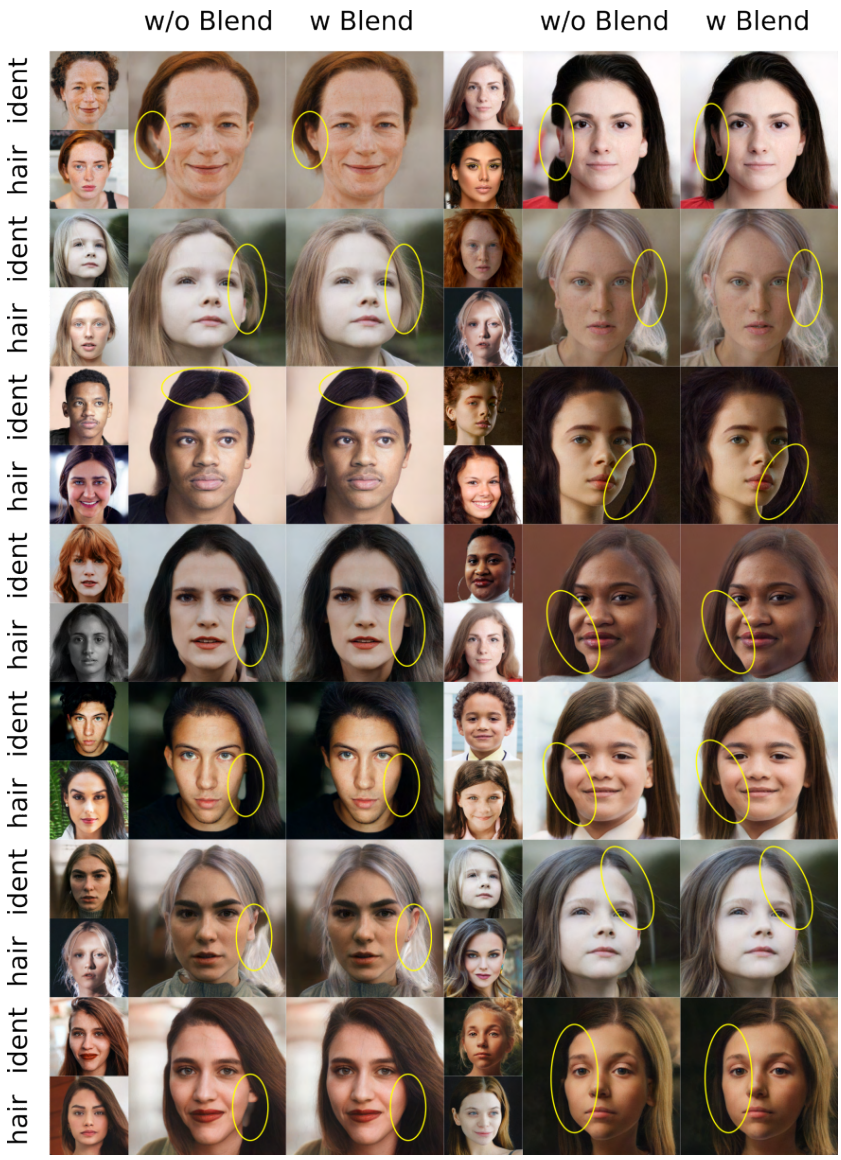

Fig. S6. The blending module of our HairNet can make an intelligent decision for ordering occlusion.

E Digging Deeper into the User Study

The setup of our user original study is shown in Fig. S7(top). Each pair of images ('ours' vs 'theirs') is shown twice, once with our results on the left, and again with ours on the right. There are a total of 760 tasks given to mechanical turkers, that is, 380 image pairs.

We were surprised at the similar scores for face reconstruction, since our own inspection of the results showed we had superior results in a majority of the images. We expected that the user study would show that our method is preferred in 70 - 90% of the cases. To understand this better, we did a second user study shown in Fig. S7(bottom) asking only about face reconstruction, and showing a much larger version of the reference face image. This time, our face reconstruction was preferred by 57% of the 760 subjects (vs 51% in the original study). This is still lower than expected, though.

As shown in Table 1 of the main paper, a user study found that our results are better at hair reconstruction than other methods, including StyleFusion, but are similar to StyleFusion for overall image quality. This very much supports our contribution, as our approach is about high quality hair transfer. We show a few examples where StyleFusion results were selected as having better image quality than ours in Fig. S8. We note that in some cases (e.g. rows 3 and 4) there are indeed some artifacts in some images using our approach. However, these artifacts are part of a tradeoff for better reconstruction quality - which we consider a clear priority for this application.

It is important to emphasize that the quality question in the user study specifically called-out image artifacts and did *not* ask the users which method reconstructed the hair more accurately; that was a separate question, in which we outperformed StyleFusion. We did not do a second user study focusing on image quality, as we did for face reconstruction quality, due to time constraints.

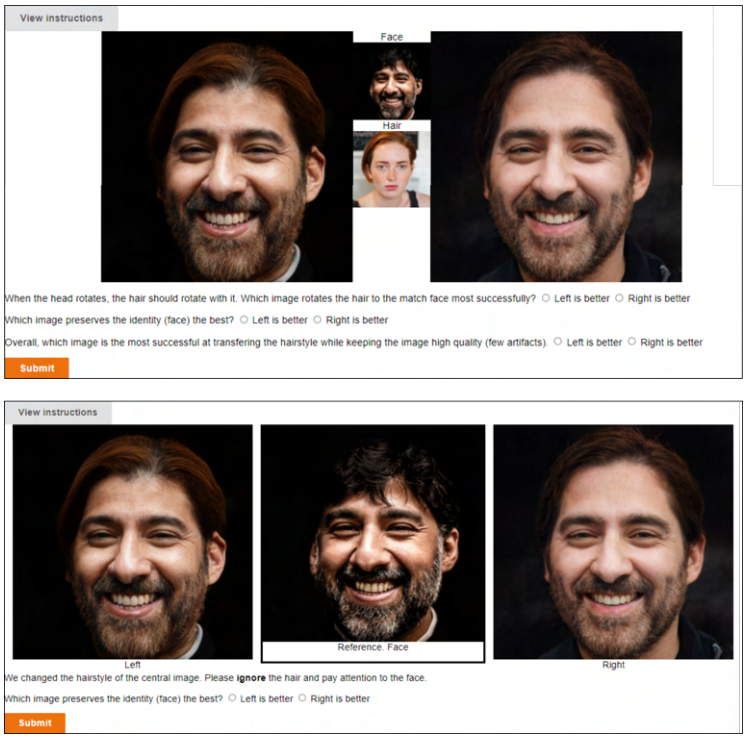

Fig. S7. Setups for the original user study included in the paper (top) and an additional study discussed here aimed at getting a more accurate answer regarding face reconstruction (bottom). In the bottom user study, the identity (face) image is shown much larger and the turkers are only asked one question. Each pair is shown with ours on the left, and again with ours on the right side.

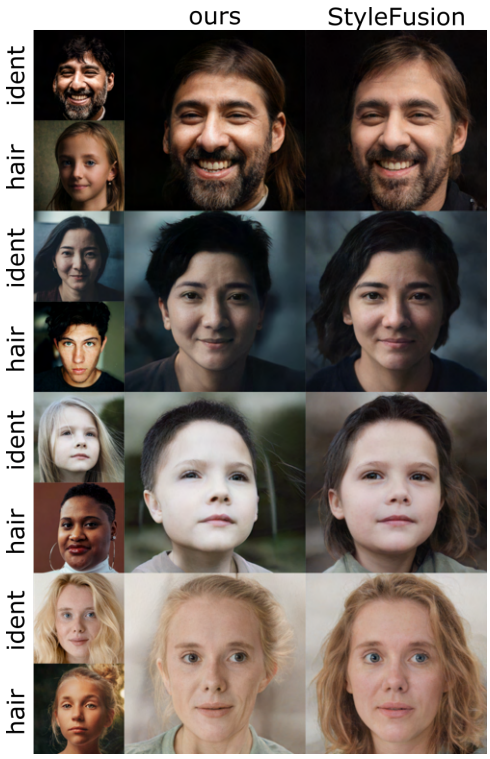

Fig. S8. Selected examples where users preferred 'StyleFusion' rather than our method in our user study. Overall both methods have similar quality, occasionally ours has issues (bottom two rows) that could be considered lower quality. However, our results are more successful at the actual editing task, and when neither method has artifacts (top rows) the more attractive, better-lit, or youthful image is often selected by users rather than the more accurate one. L_{mask} is discussed in section 3.6 of the main paper, and comes from the Barbershop paper.

F Live Editing

We have created a video demonstrating a user-interface for hairstyle transfer. Please find the video included in these supplemental materials. In the software, images are imported and some initial processing is done (embedding), and then the embedded images are mixed using our approach. Users can simultaneously change the hairstyle, pose, and other image attributes.

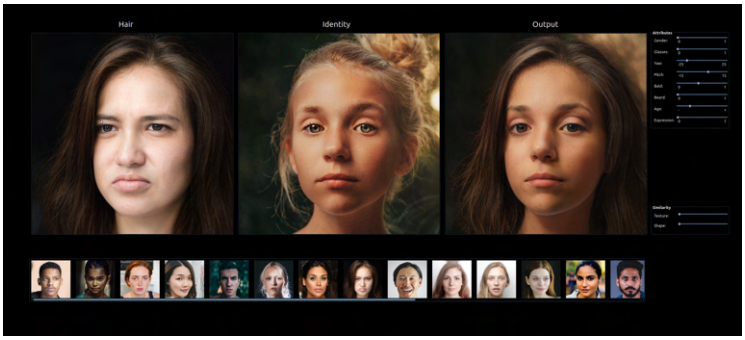

Fig. S9. A video of the live-editing of hairstyles is included with these supplementary materials.

References

1. Saha, R., Duke, B., Shkurti, F., Taylor, G.W., Aarabi, P.: Loho: Latent optimization of hairstyles via orthogonalization (2021)
2. Wei, T., Chen, D., Zhou, W., Liao, J., Tan, Z., Yuan, L., Zhang, W., Yu, N.: Hairclip: Design your hair by text and reference image (2021)