Supplementary Material

Taewoo Kim*, Chaeyeon Chung*, Yoonseo Kim*,
Sunghyun Park, Kangyeol Kim, and Jaegul Choo

Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{specia1ktu, cy_chung, grandchasevs
psh01087, kangyeolk, jchoo}@kaist.ac.kr
* indicates equal contributions.

In supplementary material, we conduct additional experiments, including a comparison to additional baselines, an ablation study, and analysis of individual modules. Also, we include implementation details and additional qualitative results to help further understand of our framework.

1 Qualitative Comparison with Additional Baselines

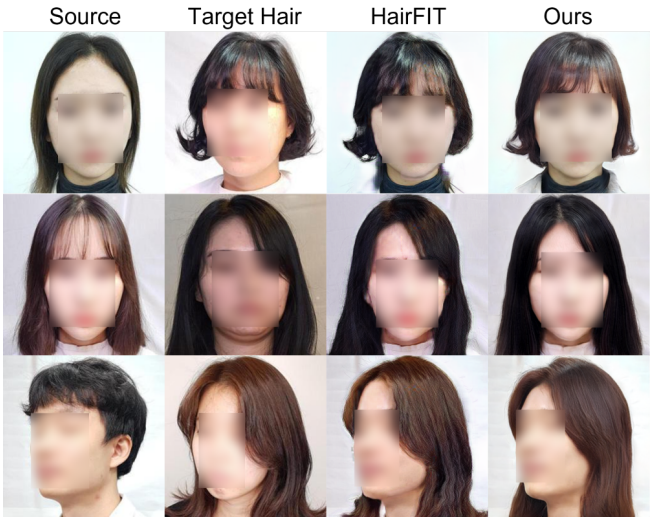

Fig. 1. Qualitative comparison with HairFIT when a source and a target hair have similar poses. Note that we blur the face of the images from the K-hairstyle dataset due to the privacy issue.

As stated in our main paper, HairFIT [1] proposes a pose-invariant hairstyle transfer model via flow-based hair warping and high-quality multi-view datasets. Also, StyleFusion [3] is a recently-proposed generative model which is capable of editing local features of an image (e.g., hairstyle in a facial image) by learning disentanglement of semantic regions in the StyleGAN [4] latent space. We

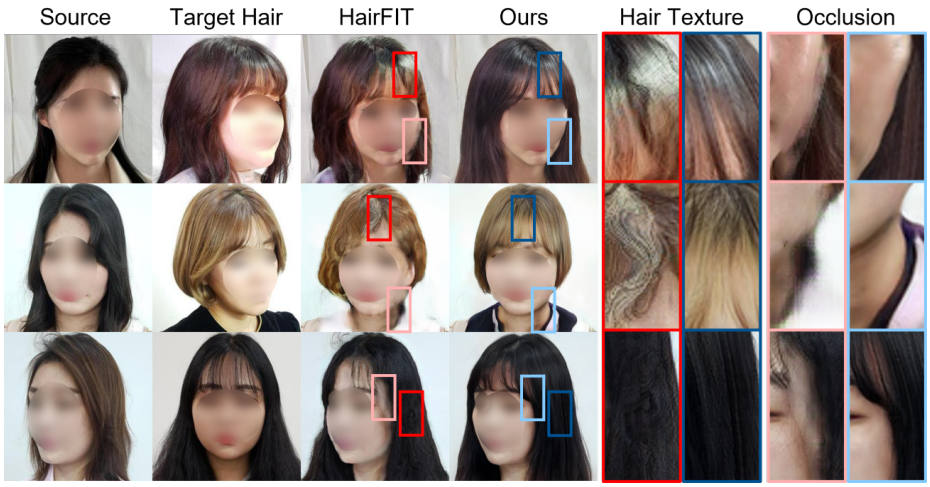

Fig. 2. Qualitative comparison with HairFIT when a source and a target hair have different poses. The last two columns present zoomed-in regions of interest, each corresponding area indicated in the third and fourth columns. (Best viewed in color.) The second last column and the last column contain regions of hair texture and occluded regions in the source, respectively. Note that we blur the face of the images from the K-hairstyle dataset due to the privacy issue.

conducted additional qualitative evaluation to demonstrate our superiority over HairFIT and StyleFusion.

First, we compare our model with HairFIT. We trained HairFIT in the same way described in the original paper and utilized the K-hairstyle dataset [6]. The implementation codes and the dataset are provided by the authors of HairFIT. K-hairstyle [6] includes 500,000 high-resolution multi-view hairstyle images with more than 6,400 identities. Following HairFIT, we filtered the images to remove the ones whose hairstyle is significantly occluded, or whose face is extremely rotated. The training set consists of 37,602 images with 4,291 identities, and the test set contains 4,309 images with 498 identities. We cropped each image based on its hair and face segmentation mask and resized the images into 256×256 for a fair comparison. For the embedding step of our framework, we trained StyleGAN2 [5] with the same dataset before the inference.

Fig. 1 and Fig. 2 present the results with similar poses and with large pose differences, respectively. Fig. 1 illustrates that HairFIT achieves comparable performance to ours when a target hair is well aligned with a source image. However, according to Fig. 2, HairFIT produces unrealistic outputs where a source and a target hair have different poses. To be specific, HairFIT could not preserve the texture of straight strands of the target hairstyles, as shown in the fifth column of Fig. 2. Moreover, HairFIT inpaints occluded regions, such as cheeks next to hair, neck, and shoulders, of a source image with undesirable blurry artifacts, as in the last column of Fig. 2.

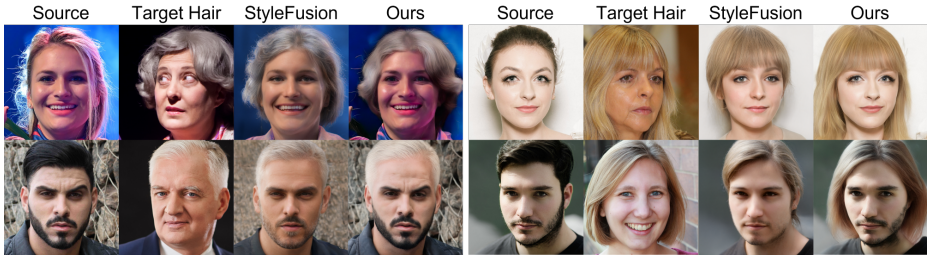

Fig. 3. Qualitative comparison with StyleFusion.

Ablated Version	Target Hair Alignment	Semantic Label of Occlusion in $\mathbf{S}_{src}^{obj}$	$w_{src}^{inpaint}$ Optimization	FID$_{\downarrow}$
(a)	✗	✗	in Blending	43.37
(b)	✓	✗	in Blending	39.68
(c)	✓	✓	in Blending	32.69
Ours	✓	✓	in Source Inpainting	**18.02**

Table 1. Quantitative comparison with the ablated versions of our framework. (a), (b), and (c) indicate each ablated version, respectively.

Additionally, we conduct a qualitative comparison with StyleFusion. We implemented the model with the official codes and utilized FFHQ dataset [4] for the comparison. Following the approach proposed in StyleFusion, we edit the 'hair' attribute in the StyleGAN2 latent space to perform hairstyle transfer. As in Fig. 3, StyleFusion is not shown to properly preserve the detailed textures as well as shapes of the target hairstyle. We speculate that the entangled attributes in the latent space (*i.e.*, hair and inner face) prevent the model from producing fine details of the hair.

2 Additional Ablation Study

To present the advantage of each step in our framework, we conduct additional quantitative and qualitative ablation studies using the FFHQ dataset. For the quantitative evaluation, we measure the fréchet inception distance (FID) score [2].

Starting from the embedding and blending step only, we gradually add each step to compare the corresponding results. In Table 1 and Fig. 4, we perform only the embedding and blending step in (a), append the target hair alignment step in (b), and add a semantic label of occluded regions to $\mathbf{S}_{src}^{obj}$ as a guide for the source inpainting in (c). Lastly, in the last row, we include the source inpainting step, an independent optimization step for inpainting, which indicates our full framework. Note that the source inpainting of (a), (b), and (c) is performed in the blending step, not in the independent source inpainting step.

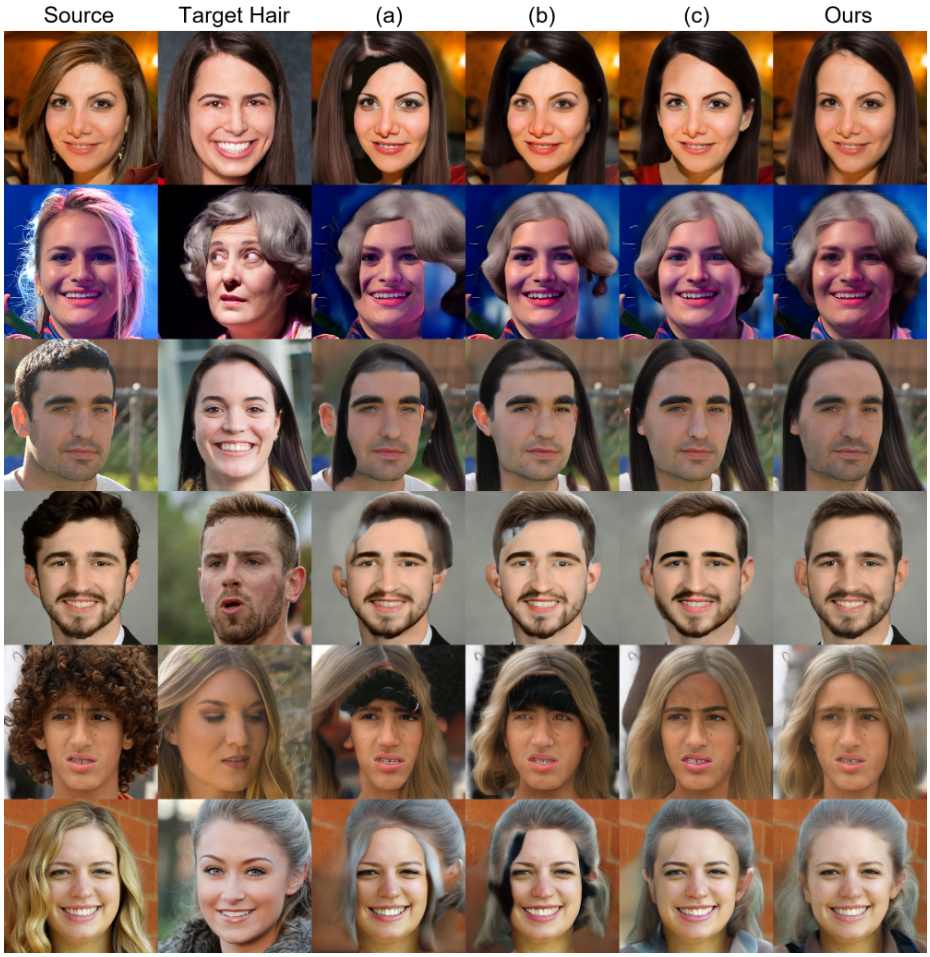

Fig. 4. Qualitative comparison with the ablated versions of our framework using the FFHQ dataset. (a), (b), and (c) indicate each ablated version, respectively.

According to Table 1, the FID score gradually decreases as we add each step of our framework. Since (a) does not have the target hair alignment step and a proper guide for the source inpainting, the corresponding outputs show dissatisfying quality. The third column of Fig. 4 illustrates the results with misaligned hair and unrealistic occlusion inpainting. Although (b) achieves the improved FID score with the aid of the target hair alignment step, source occlusions of the outputs are filled with unnatural textures, as presented in the fourth column of Fig. 4. Since a lack of semantic label of occluded regions in $\mathbf{S}_{src}^{obj}$ cannot provide a proper guide for the source inpainting, (b) allows the occluded regions to be inpainted with random undesirable textures. On the other hand, (c) produces the results with advanced quality, especially in the regions of source occlusion,

Fig. 5. (a) Visual comparison with InterFaceGAN on the alignment module and (b) qualitative comparison with baselines equipped with our alignment module.

with an appropriate assist of $\mathbf{S}_{src}^{obj}$. However, the fifth column of Fig. 4 indicates that the final outputs include regions inpainted with undesirable colors or textures. This is because the source inpainting, *i.e.*, the optimization of $w_{src}^{inpaint}$, is conducted simultaneously with the optimization of a blending weight w^{weight} in blending step.

To address this issue, we added an independent $w_{src}^{inpaint}$ optimization step only for source inpainting in our final framework, which achieves superior performance both quantitatively and qualitatively. The last column of Fig. 4 presents the results of our full framework with a superior quality of target hair alignment and occlusion inpainting compared to other configurations.

3 Analysis of Individual Modules

In this section, we further analyze each module of our framework. First, we conduct an additional evaluation on the target hair alignment module (HA). As a baseline, we adopt InterFaceGAN [7] which has the capability of aligning the target pose similar to a source via latent vector interpolation. Fig. 5(a) demonstrates the qualitative result on alignment, where the target hair is manipulated to show the same objective pose. The objective pose is selected by randomly interpolating the target hair along the pose boundary of InterFaceGAN. InterFaceGAN shows high performance on pose alignment but inappropriately alters the target hairstyle. In contrast, our model properly produces a pose-aligned image while maintaining the target hair details.

Next, we evaluate the rest of our modules except for HA, by combining HA with the baselines: LOHO and Barbershop. To this end, we perform a user study with 20 graduate students, to compare 20 images generated by three different configurations: HA + LOHO, HA + Barbershop, and ours. For each pair, a participant is asked to select a top-1 sample with two criterion: (1) preservation of delicate features of a target hair and (2) inpainting quality against source occlusion. The study results show that 71% and 63% of our method results were selected as the top-1 sample for each criteria, respectively. Fig. 5(b) shows qualitative examples appeared in the user study. As can be seen, our model visually outperforms the baselines by successfully preserving the texture and shape of a target hair.

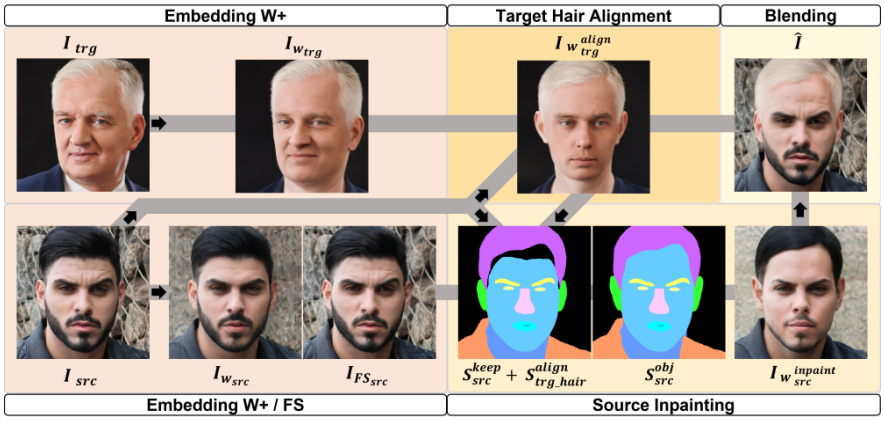

Fig. 6. Visualization of results from each step in our framework.

4 Implementation Details

The optimization step size of embedding $W+$, embedding FS, target hair alignment, source inpainting, and blending are 1,100, 250, 100, 140, and 400, respectively. In the target hair alignment and blending step, we set all lambdas of the losses as 1.

The optimization is conducted on a single GeForce RTX 3090 GPU and it requires 10GB GPU memory. For the inference time, the embedding step takes less than 2 minutes per image, and all the other steps take 78 seconds on average in total.

5 Additional Qualitative Results

First, to further understand our framework, we visualize an example qualitative result with its intermediate outputs from each step in Fig. 6. To be specific, given the source image I_{src} and the target hair image I_{trg}, $I_{w_{trg}}$, $I_{w_{src}}$, and $I_{FS_{src}}$ are the images reconstructed from the embedded latent codes w_{trg}, w_{src}, and FS_{src} obtained in the embedding step. Also, $I_{w_{trg}^{align}}$ is the aligned target hair image generated from w_{trg}^{align} obtained in the target hair alignment step. Then, in the source inpainting step, we first create an objective label S_{src}^{obj} for source inpainting based on S_{src}^{keep} from I_{src} and $S_{trg_hair}^{align}$ from $I_{w_{trg}^{align}}$. By optimizing w_{src} to follow S_{src}^{obj}, we obtain inpainted source latent code $w_{src}^{inpaint}$, which is visualized in $I_{w_{src}^{inpaint}}$. Finally, the final output $\hat{I}$ is generated via the blending step, where we blend w_{trg}^{align} and other features in w_{src} and $w_{src}^{inpaint}$.

Additionally, Fig. 7 presents additional qualitative results with FFHQ dataset. We transfer various target hairstyles on the first row of Fig. 7 to each of the source images in the first column of Fig. 7.

Fig. 7. Additional qualitative results of our framework with the FFHQ dataset.

References

1. Chung, C., Kim, T., Nam, H., Choi, S., Gu, G., Park, S., Choo, J.: Hairfit: Pose-invariant hairstyle transfer via flow-based hair alignment and semantic-region-aware inpainting. In: Proc. of the British Machine Vision Conference (BMVC). British Machine Vision Association (2021)
2. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proc. the Advances in Neural Information Processing Systems (NeurIPS) (2017)
3. Kafri, O., Patashnik, O., Alaluf, Y., Cohen-Or, D.: Stylefusion: A generative model for disentangling spatial segments. arXiv preprint arXiv:2107.07437 (2021)
4. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2019)
5. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2020)
6. Kim, T., Chung, C., Park, S., Gu, G., Nam, K., Choe, W., Lee, J., Choo, J.: K-hairstyle: A large-scale korean hairstyle dataset for virtual hair editing and hairstyle classification. In: Proc. of the IEEE International Conference on Image Processing (ICIP). pp. 1299–1303. IEEE (2021)
7. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2020)