Style Your Hair: Latent Optimization for Pose-Invariant Hairstyle Transfer via Local-Style-Aware Hair Alignment

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.

Abstract. Editing hairstyle is unique and challenging due to the complexity and delicacy of hairstyle. Although recent approaches significantly improved the hair details, these models often produce undesirable outputs when a pose of a source image is considerably different from that of a target hair image, limiting their real-world applications. HairFIT, a pose-invariant hairstyle transfer model, alleviates this limitation yet still shows unsatisfactory quality in preserving delicate hair textures. To solve these limitations, we propose a high-performing pose-invariant hairstyle transfer model equipped with latent optimization and a newly presented local-style-matching loss. In the StyleGAN2 latent space, we first explore a pose-aligned latent code of a target hair with the detailed textures preserved based on local style matching. Then, our model inpaints the occlusions of the source considering the aligned target hair and blends both images to produce a final output. The experimental results demonstrate that our model has strengths in transferring a hairstyle under larger pose differences and preserving local hairstyle textures. The codes are available at https://github.com/Taeu/Style-Your-Hair.

Keywords: Hairstyle transfer; Latent optimization; Conditional image generation.

1 Introduction

With the advance of conditional generative adversarial networks (GANs) [9,19,13], editing facial attributes has drawn great attention and shows a promising result on editing multiple attributes. Despite the success, modifying strongly correlated facial attributes is still challenging, often beyond the capacity of existing editing models. In this paper, we focus on hairstyle editing, which aims at transferring a target hairstyle to a source image, proposing high-performance neural networks to solve the problem. Hairstyle editing is similar to that of a facial attribute, but it has unique, challenging aspects: (1) Due to the hairstyle's complexity and delicacy, preserving its strands given an arbitrary hairstyle is highly demanding. (2) Transferred hairstyle requires to be exactly fitted to a given source image.

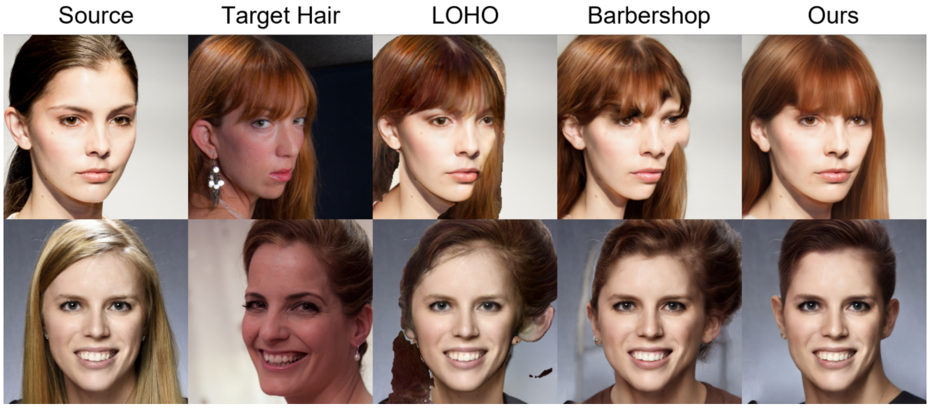


Fig. 1. Our model produces more realistic results compared to LOHO [21] and Barbershop [33] even with a large pose difference between a source and a target hair.

These challenges make the previous approaches for editing the specified facial attributes less suitable for this problem.

Recent solutions for hairstyle transfer address the problem with the power of a pre-trained image generator. For example, LOHO [21] and Barbershop [33] largely enhance the visual quality of the generated images via latent optimization based on StyleGAN2 [15]. However, these approaches produce undesirable outputs (See Fig. 1) when handling a target and source image pair with a significant pose difference.

To the best of our knowledge, HairFIT [6] is the only work to address the pose difference issue between a source and target image. HairFIT presents a pose-invariant hairstyle transfer model where a target hairstyle is aligned to a source image pose using a flow-based warping module trained on multi-view datasets such as VoxCeleb [18] and K-hairstyle dataset [16]. Although its attempt, HairFIT requires a high-quality multi-view hairstyle dataset during training, and it falls behind state-of-the-art models [21,33] in light of hair preserving capacity.

In response to these limitations, we present a framework that performs a *high-quality* pose-invariant hairstyle transfer based on latent optimization *without multi-view dataset*. Specifically, given a source and a target hair image, our model generates a hair-transferred output through embedding, hair pose alignment, inpainting, and blending step. We first take advantage of GAN-inversion algorithms [1,3,34], feeding a source and a target hair image, for the purpose of obtaining latent codes residing in the StyleGAN2 space [15], respectively. Next, we navigate the StyleGAN2 space to optimize the latent code of the target hair image to follow a source image pose. During the pose alignment, we utilize a newly-presented local-style-matching loss to penalize visually degraded hair textures by locally comparing the original target hair with the aligned one. In the inpainting, we first obtain a segmentation mask to guide the latent code of the source to fill the occluded regions by its hair. We optimize the latent code of the source image to follow the obtained segmentation mask. Lastly, we blend the aligned target hairstyle and the inpainted source image with a final optimization step. In this manner, our model is able to transfer a target hairstyle to a source image overcoming the difference in poses as well as successfully preserving the fine details of the target hair. Experiments demonstrate the superiority of our model in a quantitative and qualitative manner. Our contributions are summarized as follows:

- We propose a framework that achieves a *high-quality* pose-invariant hairstyle transfer based on latent optimization *without multi-view dataset.*
- We present local-style-matching loss to maintain the fine details of a hairstyle during pose optimization.
- Our model achieves state-of-the-art performance in quantitative and qualitative evaluations with various datasets.

2 Related Work

2.1 Latent space manipulation

With the understanding of the latent space in GANs, recent approaches based on latent space manipulation [10,22,25] have shown promising results in image editing. For example, GANSpace [10] and InterfaceGAN [22] modify facial attributes via manipulation in the latent space of StyleGAN [14]. While the former takes advantage of principal component analysis, the latter utilizes semantic scores to identify disentangled directions related to the target attributes. In a similar manner, Viazovetskyi et al. [25] and Zhuang et al. [35] attempt to edit the images by shifting latent vectors to semantically meaningful directions in the latent space of StyleGAN2 [15], which can easily be obtained by a pre-trained face classifier or learned transformations.

Recent hairstyle transfer approaches also actively utilize latent space manipulation to synthesize high-quality images. LOHO [21] and Barbershop [33] edit hairstyles by manipulating the extended StyleGAN2 latent space [34] via latent optimization. These methods not only significantly enhance the visual quality of the generated images but also preserve the semantic details of the target images. In particular, Barbershop introduces FS space with a larger capacity than the original StyleGAN2 latent space, where the original hair structure is well-preserved. In this work, we leverage latent optimization to reach the photo-realistic image quality. Our model mainly focuses on the pose alignment of a target hairstyle to a source image without losing its detailed hair texture based on local-style-matching loss to achieve a pose-invariant hairstyle transfer.

2.2 Hairstyle Transfer

GAN-based facial image editing [12,13,17,20,27,28] successfully modifies the target facial attributes such as a facial expression or makeup style while maintaining other features. Common approaches for facial image editing are to utilize hand-drawn sketches [13,20,27,28] or user-edited semantic masks[17] as the conditions to precisely guide the manipulated appearance.

In spite of the remarkable progress in facial image editing, hairstyle transfer is still tricky, considering the diversity and intricacy of hairstyles. In practice, a hairstyle transfer is required to convey a wide range of target hairstyles to a given source image while preserving their subtle hair strands and color. As a prior work, MichiGAN [24] presents a hairstyle transfer framework aiming to preserve the detailed textures of a target hairstyle. Specifically, MichiGAN leverages different conditional generators responsible for decomposed hairstyle attributes (*i.e.*, hair shape, and appearance). Moreover, LOHO [21] achieves visually pleasant image quality through latent optimization and hair-related losses for reflecting a target hairstyle features. Barbershop [33] also proposes a latent optimization approach and further improves the visual quality of the outputs based on the newly presented FS space. Barbershop utilizes F tensor in FS space to enhance the capability of preserving the overall structure of a target image, including delicate hair structure.

However, since the existing approaches have been developed to handle the images, where the head poses of a source and a target are aligned, they show limited generalization capacity for dealing with the inputs having a large pose difference. To tackle this problem, HairFIT [6] introduces a pose-invariant hairstyle transfer with flow-based target hair warping and semantic-region-aware inpainting. HairFIT leverages an optical flow estimation network and a multi-view hairstyle dataset [16] to align the target hair to the source face. Despite the aid of the high-quality multi-view dataset, the model fails to preserve the detailed features of hairstyles comprehensively. In this paper, we propose a novel latent optimization framework for pose-invariant hairstyle transfer to synthesize high-quality images regardless of the pose differences.

3 Method

3.1 Overview

Our framework consists of several optimization steps described in Fig. 2. We first find latent codes $w \in \mathbb{R}^{18\times512}$ of a source image $\mathbf{I}_{src} \in \mathbb{R}^{C\times H\times W}$ and a target hairstyle image $\mathbf{I}_{trg}$ in $W+$ space using the existing GAN-inversion algorithms [33,34]. Then, we optimize the target hair latent codes to have the pose aligned to $\mathbf{I}_{src}$. While aligning the pose, we mainly focus on preserving fine details of the target hair with a newly-presented local-style-matching loss. Local-style-matching loss allows preserving each local texture in the aligned target hair by matching the corresponding region of a similar style from the original target hair. For the next step, we inpaint the source regions occluded by its original hair by optimizing the source latent codes. Lastly, we blend the aligned target hairstyle and the inpainted source image for the final output.

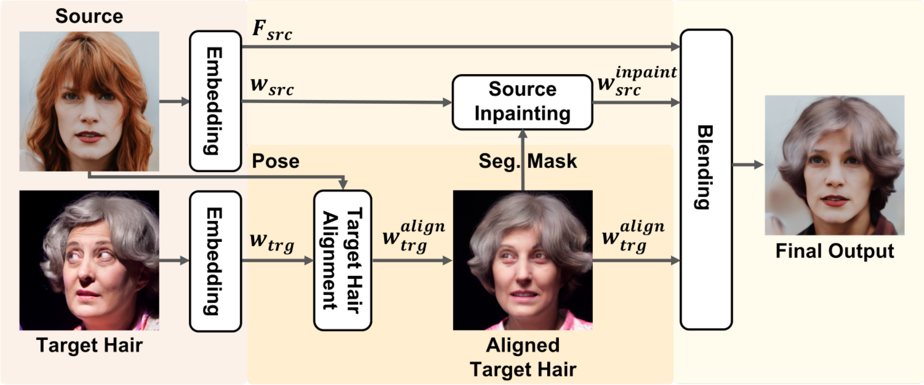


Fig. 2. An overview of our framework. First, we obtain w_{src}, w_{trg}, and F_{src} by embedding a source and a target hair image into $W+$ and FS space. Then, we optimize w_{trg} to follow the source pose, resulting in w_{trg}^{align}. With the segmentation mask of aligned target hair, we find $w_{src}^{inpaint}$, where the source occlusions are inpainted. Finally, we blend F_{src}, $w_{src}^{inpaint}$, and w_{trg}^{align} to generate the final output.

3.2 Embedding

First of all, we obtain the latent codes of each reference image (*i.e.*, source and target images) before pose alignment and blending. Given a source image $\mathbf{I}_{src}$ and a target hair image $\mathbf{I}_{trg}$, we find the source latent codes w_{src} and the target latent codes w_{trg} in an extended latent space of StyleGAN2 denoted as $W+$ space [2]. We employ an improved embedding algorithm [34] to enhance the reconstruction and editing quality. Moreover, we embed $\mathbf{I}_{src}$ to FS space following Barbershop [33] to gain $\mathbf{F}_{src} \in \mathbb{R}^{32\times32\times512}$, which preserves the detailed structure of the source image by encoding the spatial information.

3.3 Target Hair Alignment

To transfer the hairstyle regardless of the pose differences, we align the target hairstyle to the source face via the latent optimization, as presented in Fig. 3. Starting from w_{trg}, we aim to find w_{trg}^{align}, where the head pose and face shape are aligned to $\mathbf{I}_{src}$, while other features, especially the hairstyle, correspond to $\mathbf{I}_{trg}$. We optimize the first m style vectors among 18 style vectors of w_{trg} to optimize coarse style vectors rather than fine style vectors [14]. We set m as 6 in our experiments.

Pose Align Loss. To modify w_{trg} to have a source pose, we propose a novel pose align loss $\mathcal{L}_{pose}$ based on 3D facial keypoints. Since the hairstyle significantly depends on other facial features (*e.g.*, face shape and location of eyes), the head pose alone is insufficient to fully guide the target hair alignment. Thus, we leverage 3D facial keypoints, which effectively represent the overall facial

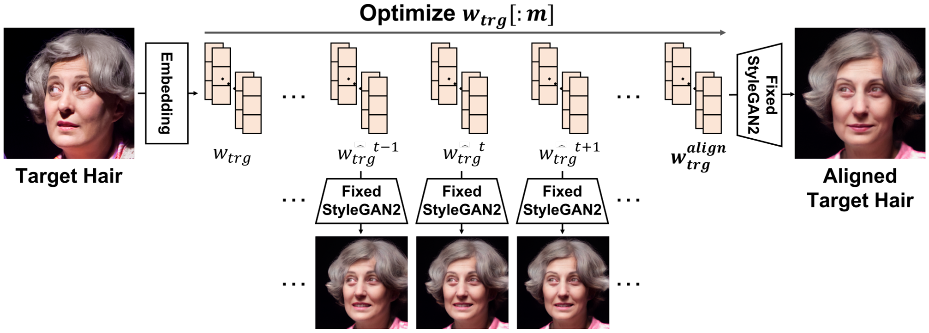


Fig. 3. Target hair alignment. We obtain the aligned target hair latent codes w_{trg}^{align} by optimizing the first m vectors of w_{trg} to have a source pose with its hairstyle preserved.

features as well as the head pose. With the source 3D facial keypoints, we can provide detailed supervision of which shape and pose w_{trg} should pursue.

$\mathcal{L}_{pose}$ computes the L2 distance between the 3D keypoint heatmaps of $\mathbf{I}_{src}$ and the aligned target hair image as:

$$\mathcal{L}_{pose} = \frac{1}{N_H} \|\mathbf{H}_{src} - E(G(\hat{w_{trg}}))\|_2^2. \tag{1}$$

N_H indicates the number of elements in a 3D keypoint heatmap $\mathbf{H} \in \mathbb{R}^{68 \times H \times W}$ and $\mathbf{H}_{src} = E(\mathbf{I}_{src})$, where E is a pre-trained keypoint extractor [5]. G is a pre-trained StyleGAN2 generator and $\hat{w_{trg}}$ indicates the optimized w_{trg} in progress.
Local-Style-Matching Loss. To preserve locally distinct hairstyles, we newly present a local-style-matching loss, which matches similar local styles between the target hair and the aligned target hair. Basically, we utilize a style loss based on the Gram matrix [8], which captures the repeated patterns (*i.e.*, texture) of given features. A style loss $\mathcal{L}_{style}$ measures the L2 distance between the Gram matrix of feature maps extracted by a VGG network [23], formulated as:

$$\mathcal{L}_{style}(\cdot,\cdot) = \frac{1}{V} \sum_{i=1}^{V} \frac{1}{N_{\mathcal{G}^i}} \|\mathcal{G}^i(\mathrm{VGG}^i(\cdot)) - \mathcal{G}^i(\mathrm{VGG}^i(\cdot))\|_2^2, \tag{2}$$

where V indicates the number of VGG layers we use, which are $relu1\_2$, $relu2\_2$, $relu3\_3$, and $relu4\_3$ layer of VGG [6,21,33]. Also, $N_{\mathcal{G}^i}$ represents the number of elements in $\mathcal{G}^i$. Here, $\mathcal{G}^i$ and VGG^i indicate the i-th Gram matrix and i-th layer of VGG, respectively. $\mathcal{G}^i$ is calculated as ${v^i}^{\intercal} v^i$, where $v^i \in \mathbb{R}^{H^i W^i \times N_{C^i}}$ corresponds to the activation of VGG^i.

In local-style-matching loss $\mathcal{L}_{style}^{LSM}$, we first identify *style regions* each of which includes locally different style and apply L_{style} to each style region, respectively. To identify the style regions, we leverage a simple linear iterative clustering (SLIC) [4]. The SLIC is an algorithm that conducts a K-means clustering based on the similarity of color and spatial distance between pixels. Since the SLIC con-

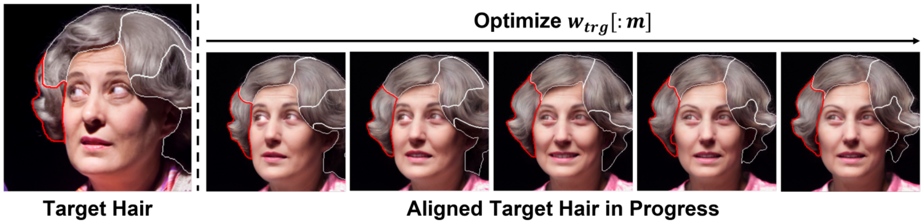


Fig. 4. Local-style-matching loss. During target hair alignment, a local-style-matching loss is applied to style regions in the target hair and those in the aligned target hair. The white boundary regions are segmented style regions, and the red boundary regions describe an example of a consistently tracked style region.

siders both the appearance and location of neighboring pixels, it can successfully segment the target hair into proper style regions.

As presented in the first column of Fig. 4, we first find the style regions in the hair of $\mathbf{I}_{trg}$. Then, during the latent optimization, we detect the style regions of $G(\hat{w_{trg}})$ and match each region to the most similar style region of $\mathbf{I}_{trg}$, as shown in the rest columns of Fig. 4. In each step, we track the regions of similar style by setting the same label to the region of the closest centroid compared to the previous step. Fig. 4 shows that an example style region marked with a red boundary is successfully tracked based on the proposed algorithm. $SLIC_{hair}(\mathbf{I}) \in \{0,1\}^{N_{style} \times H \times W}$ indicates style region masks extracted from a hair region of $\mathbf{I}$ using the SLIC algorithm. Here, N_{style} indicates the number style regions. We set N_{style} as 5 in our experiments. $\mathcal{L}^{LSM}_{style}$ is formulated as:

$$\mathcal{L}^{LSM}_{style} = \sum_{i=1}^{N_{style}} \mathcal{L}_{style}(SLIC^i_{hair}(\mathbf{I}_{trg}) \odot \mathbf{I}_{trg}, SLIC^i_{hair}(G(\hat{w_{trg}})) \odot G(\hat{w_{trg}})). \quad (3)$$

$SLIC^i_{hair}(\cdot)$ is the i-th channel of $SLIC_{hair}(\cdot)$ and $\odot$ indicates element-wise product. Note that a valid region of each channel, where the style region mask corresponds to 1, is cropped before calculating the style loss.

Regularization Loss. We add a step-wise regularization loss to keep the overall features of $\hat{w_{trg}}$, especially hairstyle, similar to the previous step. The regularization loss $\mathcal{L}_{reg}$ encourages a stable optimization via a gradual modification without a noticeable loss of the original hairstyle features. $\mathcal{L}_{reg}$ is formulated as:

$$\mathcal{L}_{reg} = \frac{1}{N_w} \|\Delta \hat{w_{trg}}\|^2_2, \quad (4)$$

where N_w indicates the number of elements in w. $\Delta \hat{w_{trg}}$ at step t is obtained by $\hat{w_{trg}}^t - \hat{w_{trg}}^{t-1}$, where t ranges from 2 to the total number of steps.

Formally, the total objective function in the target hair alignment step is $\mathcal{L}_{pose} + \lambda^{LSM}_{style}\mathcal{L}^{LSM}_{style} + \lambda_{reg}\mathcal{L}_{reg}$, where λ^{LSM}_{style} and λ_{reg} denote the hyper-parameters to control relative importance between different losses.

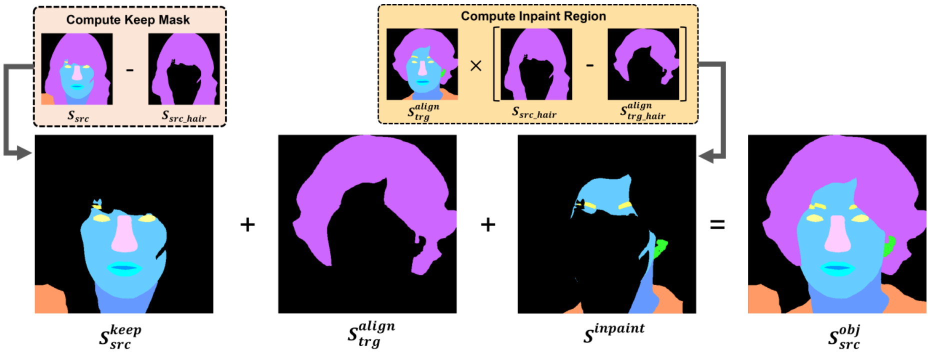


Fig. 5. Generation of an objective label. For source inpainting, we create an objective label $\mathbf{S}_{src}^{obj}$ to guide the occluded regions to be inpainted with proper semantics.

3.4 Source Inpainting

Source inpainting step aims to inpaint the regions occluded by the original source hair. As shown in Fig. 2, if we remove the source hair region from the source image, the occluded region should be filled with the proper semantics (*e.g.*, forehead, face, neck, clothes, and background) to fit the aligned target hair.

To find the inpainted source latent code $w_{src}^{inpaint}$, we generate an objective label $\mathbf{S}_{src}^{obj} \in \mathbb{Z}^{H \times W}$ to guide the occluded regions to be filled with the appropriate semantic regions. $\mathbf{S}_{src}^{obj}$ is generated by the following process, as also described in Fig. 5. First, we compute a keep label $\mathbf{S}_{src}^{keep}$, which indicates the regions that need to be maintained in the source, by removing a source hair region $\mathbf{S}_{src\_hair}$ from a source semantic label $\mathbf{S}_{src}$. Here, $\mathbf{S}_{src}$ is estimated by a pre-trained segmentation network [29]. Next, we calculate a label of regions to be inpainted $\mathbf{S}^{inpaint}$ as described in Fig. 5. Finally, we obtain $\mathbf{S}_{src}^{obj}$ which indicates the inpainting regions of the source image considering the aligned target hair. Now, we optimize $w_{src}^{inpaint}$ to follow the given $\mathbf{S}_{src}^{obj}$. Here, as in the target hair alignment step, we optimize the first m w vectors to newly generate coarse features to fill the occlusions while preserving the fine details or the overall appearance of the source. For optimizing $w_{src}^{inpaint}$, we use a pixel-wise cross-entropy loss between the label of $\mathbf{S}_{src}^{obj}$ and a segmentation probability heatmap of the generated image, which consists of 16 semantic region categories. The heatmap is estimated by the pre-trained segmentation network.

3.5 Blending

The final optimization step aims to find a blending weight w^{weight} that merges the optimized latent codes from the previous steps to generate the final output. First, as presented in Fig. 6(a), w^{blend} is obtained by blending $w_{src}^{inpaint}$ and w_{trg}^{align} with the blending weight w^{weight}. w^{blend} is formulated as $w_{src}^{inpaint} + w^{weight} \odot w_{trg}^{align}$, where w^{weight} implies how much of w_{trg}^{align} needs to be reflected

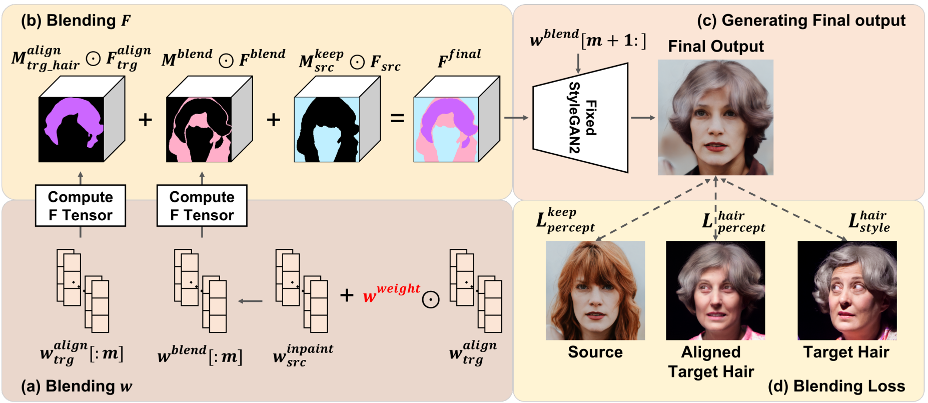


Fig. 6. Blending. (a) w vectors from the previous steps are blended with the optimized blending weight w^{weight} to obtain w^{blend}. (b) Next, we combine $\mathbf{F}^{align}_{trg}$, $\mathbf{F}^{blend}$, and $\mathbf{F}_{src}$ with the corresponding masks to obtain $\mathbf{F}^{final}$. (c) $\mathbf{F}^{final}$ and w^{blend} are fed to the StyleGAN2 generator to synthesize the final output. (d) Blending loss consists of $\mathcal{L}^{keep}_{percept}$, $\lambda^{hair}_{percept}$, and $\mathcal{L}^{hair}_{percept}$.

to synthesize the final output. Then, we prepare $\mathbf{F}$ tensors by feeding the first m w vectors to the pre-trained StyleGAN2 generator, as shown in Fig. 6(b). Here, we leverage $\mathbf{F}$ tensors in FS space to effectively reconstruct the detailed spatial information [33] in the further blending. We blend $\mathbf{F}^{align}_{trg}$, $\mathbf{F}^{blend}$, and $\mathbf{F}_{src}$ to gain $\mathbf{F}^{final}$ which contains detailed spatial information of the final output. $\mathbf{F}^{final}$ is calculated as follows:

$$\mathbf{F}^{final} = \mathbf{M}^{align}_{trg\_hair} \odot \mathbf{F}^{align}_{trg} + \mathbf{M}^{blend} \odot \mathbf{F}^{blend} + \mathbf{M}^{keep}_{src} \odot \mathbf{F}_{src}. \tag{5}$$

$\mathbf{F}^{align}_{trg}$ and $\mathbf{F}^{blend}$ are extracted from w^{align}_{trg} and w^{blend}, respectively, and $\mathbf{F}_{src}$ is from the embedding step. $\mathbf{M}^{align}_{trg\_hair}$ is a binary mask indicating the hair region in the aligned target hair image. $\mathbf{M}^{keep}_{src}$ is also a binary mask denoting the regions which are neither the source hair nor the aligned target hair. $\mathbf{M}^{keep}_{src}$ indicates the area that needs to be preserved in the source image. Lastly, $\mathbf{M}^{blend}$ denotes the remaining regions. The final output $\hat{\mathbf{I}}$ is generated from the pre-trained StyleGAN2 generator given w^{blend} and $\mathbf{F}^{final}$ as inputs. Here, vectors in w^{blend} except the first m vectors are fed to the generator.

Losses. In order to blend the previous optimized latent codes while preserving their structure and styles, we utilize the following losses.

First, to maintain the source face, clothes, background, etc., we apply the perceptual loss [30] on the valid regions in $\mathbf{S}^{keep}_{src}$ (*i.e.*, the regions to be preserved in the source image) as follows:

$$\mathcal{L}^{keep}_{percept} = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{N_{\mathrm{VGG}^i}}\|\mathbf{M}^{keep}_{src} \odot (\mathrm{VGG}^i(\mathbf{I}_{src}) - \mathrm{VGG}^i(\hat{\mathbf{I}}))\|_1, \tag{6}$$

where VGG^i denotes i-th layer of VGG16 network [23] and N_{VGG^i} is the number of elements in the activation of VGG^i.

Also, in order to preserve the aligned target hairstyle from the aligned target latent code w_{trg}^{align}, we use the hair perceptual loss formulated as follows:

$$\mathcal{L}_{percept}^{hair} = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{N_{\mathrm{VGG}^i}}\|\mathbf{M}_{trg\_hair}^{align} \odot (\mathrm{VGG}^i(\mathbf{I}_{trg}^{align}) - \mathrm{VGG}^i(\hat{\mathbf{I}}))\|_1. \tag{7}$$

Lastly, we maintain the texture of the original target hair by utilizing the hairstyle loss $\mathcal{L}_{style}^{hair}$, where the style loss $\mathcal{L}_{style}$ is applied on the hair regions of the target hair image and the final output as $\mathcal{L}_{style}(\mathbf{M}_{trg\_hair} \odot \mathbf{I}_{trg}, \mathbf{M}_{\hat{I}\_hair} \odot \hat{\mathbf{I}})$.

The total blending loss to optimize w^{weight} is $\mathcal{L}_{percept}^{keep} + \lambda_{percept}^{hair}\mathcal{L}_{percept}^{hair} + \lambda_{style}^{hair}\mathcal{L}_{style}^{hair}$, where $\lambda_{percept}^{hair}$ and λ_{style}^{hair} are the hyper-parameters to balance the relative importance between the losses.

4 Experiments

4.1 Experimental Setup

Dataset. We utilize Flickr-Faces-HQ (FFHQ) dataset [14] for hairstyle transfer and K-hairstyle [16] and VoxCeleb2 [7] for reconstruction task. For hairstyle transfer, we sample 6,000 pairs of two different identities (one for source and the other for target hairstyle) from 70,000 1,024×1,024 images in FFHQ.

For the reconstruction, we create 500 test pairs by sampling the images from the K-hairstyle dataset, which includes 500,000 high-resolution multi-view images with more than 6,400 identities. Following HairFIT [6], we filtered the images to remove the ones whose hairstyle is significantly occluded, or whose face is extremely rotated. Additionally, we sample 500 pairs of a source and a target from more than 1 million videos in VoxCeleb2. In the reconstruction task, two images in each pair have the same identity and different poses, and the source image in each pair is considered the ground truth image for the model to reconstruct. Each image is resized to 256×256 in the experiments.

Baseline Models. We conduct a quantitative and qualitative comparison between our model and the following baselines: LOHO [21], Barbershop [33], and HairFIT [6]. Here, we follow the official implementation code of LOHO and Barbershop. Since LOHO utilizes an external inpainting network, we use a state-of-the-art inpainting network CoModGAN [32]. Also, we implement HairFIT with the codes and guidelines provided by the authors of HairFIT.

4.2 Comparison to Baselines

Quantitative evaluations. First, we compare the fréchet inception distance (FID) score [11] of LOHO, Barbershop, and our model on hairstyle transfer task. The FID score measures how similar the distributions of the synthesized images and the real images are, where the lower FID score indicates a higher

Pose difference level	Easy	Medium	Difficult	Total
LOHO [21]	21.70	23.40	28.36	19.63
Barbershop [33]	**20.75**	21.45	26.30	18.07
Ours	20.79	**20.56**	**22.72**	**17.06**

Table 1. Quantitative comparison with baselines. We measure the FID scores with three different levels of pose difference and with total pairs.

Dataset	K-hairstyle		VoxCeleb2	
Metric	SSIM$_\uparrow$	LPIPS$_\downarrow$	SSIM$_\uparrow$	LPIPS$_\downarrow$
HairFIT	0.7242	0.2054	0.7520	**0.2033**
Ours	**0.7424**	**0.1786**	**0.7717**	0.2078

Table 2. Quantitative comparisons with HairFIT using multi-view datasets.

similarity. We compare 6,000 pairs of real and fake images, where each image is resized to 256×256 for the evaluation. As shown in the last column of Table 1, we achieve the lowest FID score compared to the baselines.

For further analysis, we compare the FID scores on three different levels of pose difference as conducted in the previous work [6,21]. We calculate the pose difference, PD, following the protocol presented in HairFIT [6]. In particular, we use 17 facial jaw keypoints extracted by the pre-trained 3D-keypoint extraction model [5]. The pose difference is calculated as PD $= \frac{1}{17}\sum_{i=1}^{17}\|\mathbf{k}^i_{src} - \mathbf{k}^i_{trg}\|_1$, where $\mathbf{k}^i_{src} \in \mathbb{R}^3$ is a 3D coordinates of the i-th source keypoint and $\mathbf{k}^i_{trg} \in \mathbb{R}^3$ is a 3D coordinates of the i-th target keypoint. Then, we divide the 6,000 pairs of a source and a target into three categories of 2,000 pairs: Easy, Medium, and Difficult. As presented in Table 1, our model outperforms the other baselines for Medium and Difficult. Moreover, the margin between the FID scores of our model and other baselines increases as the pose difference increases from Easy to Difficult.

Additionally, we conduct a comparison with HairFIT on the reconstruction task using K-hairstyle and VoxCeleb2. As in HairFIT, we measure the structural similarity (SSIM) [26] and learned perceptual image patch similarity (LPIPS) [31] between generated images and ground truth images. Table 2 presents that our model outperforms HairFIT (except for LPIPS of VoxCeleb2) *even without* learning to reconstruct different views of a source image using a multi-view dataset.

Qualitative evaluations. Fig. 7 and Fig. 8 demonstrate that our model successfully transfers the target hairstyle into the source regardless of the pose differences. Especially, as presented in Fig. 7, our model shows superiority over other baselines on the Difficult level. Furthermore, although the FID score of our model is slightly higher than Barbershop on Easy level, Fig. 8 present that the quality of our model is better to reflect the target hairstyle than the baselines. The results show that our model successfully aligns the target hair to the source

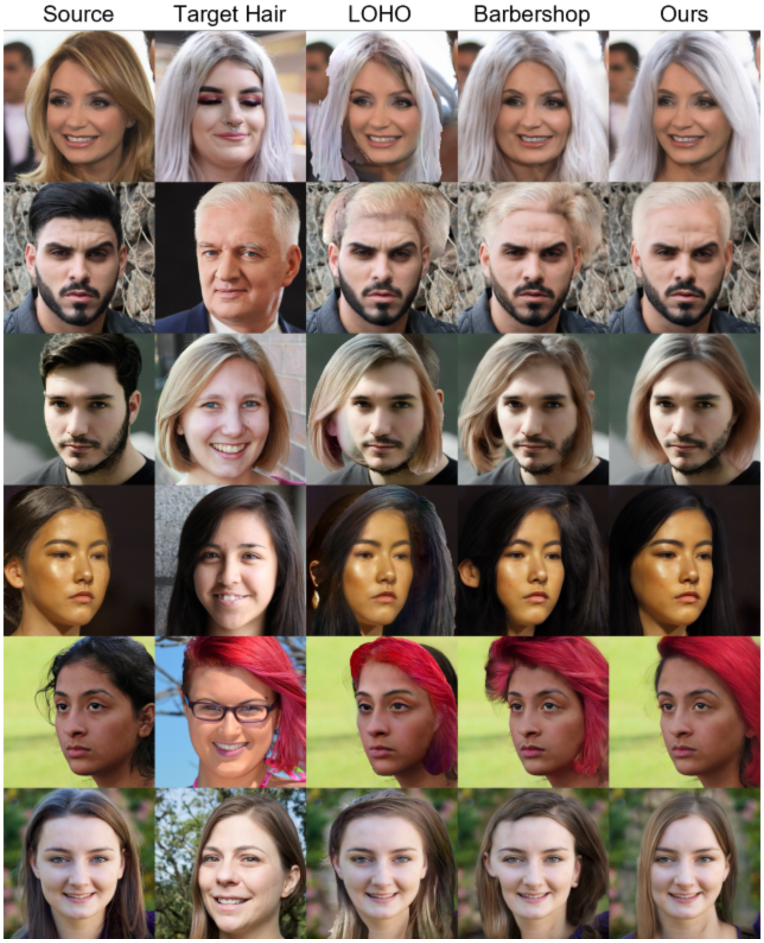
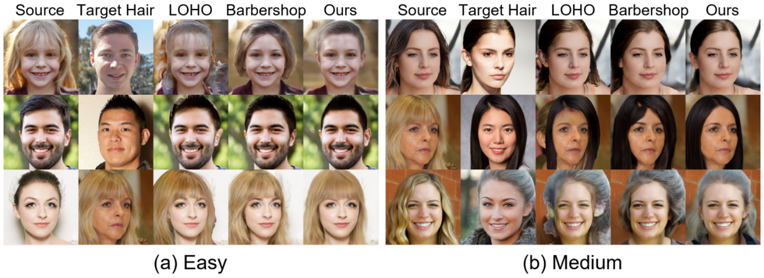


Fig. 7. Qualitative comparison with the baselines on Difficult level of pose difference.



(a) Easy

(b) Medium

Fig. 8. Qualitative comparison on (a) Easy and (b) Medium level of pose difference.

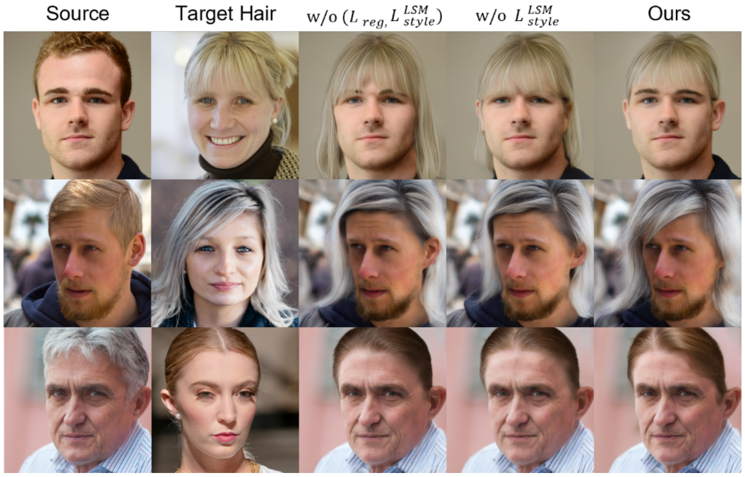


Fig. 9. Qualitative ablation study on the losses in target hair alignment step.

Configurations	w/o $(\mathcal{L}_{reg}, \mathcal{L}^{LSM}_{style})$	w/o $\mathcal{L}^{LSM}_{style}$	Ours
SSIM$_{\uparrow}$	0.7667	0.7716	**0.7717**
LPIPS$_{\downarrow}$	0.2125	0.2082	**0.2078**

Table 3. Quantitative ablation study on the losses in target hair alignment step.

image, producing high-quality images of hairstyle transfer. More results of the qualitative comparison are presented in the supplementary materials.

4.3 Ablation Study

In the ablation study, we demonstrate the effectiveness of a local-style-matching loss and regularization loss in our target hair alignment step. We conduct a qualitative evaluation on hairstyle transfer using the FFHQ dataset and quantitative evaluation on the reconstruction task with VoxCeleb2. In Fig. 9 and Table 3, *w/o* $(\mathcal{L}_{reg}, \mathcal{L}^{LSM}_{style})$ denotes our framework without $\mathcal{L}_{reg}$ and $\mathcal{L}^{LSM}_{style}$). Also, *w/o* $\mathcal{L}^{LSM}_{style}$ indicates our framework without $\mathcal{L}^{LSM}_{style}$ and *Ours* is our full framework.

The first row of Fig. 9 indicates that the generated target hair is longer than the original target hair due to the absence of the $\mathcal{L}_{reg}$. In the second row, the direction of the front hair of the outputs without $\mathcal{L}^{LSM}_{hair}$ are different from the original target hairstyle. Moreover, in the third row, the “part” of the target hair is better reflected in the output of ours. The results present that our proposed losses effectively reflect the local style of the target hair while preserving its overall style. Additionally, as seen in Table 3, our full model outperforms other configurations with a gradual performance increase. Although the difference between *Ours* and *w/o* $\mathcal{L}^{LSM}_{style}$ is marginal, the qualitative results presented above

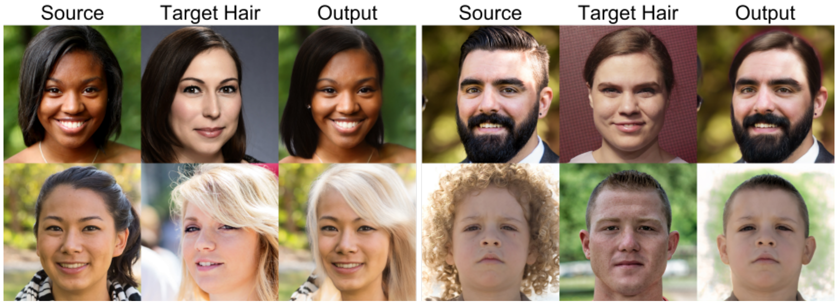


Fig. 10. Limitations of our proposed method.

clearly illustrate the high visual quality of our full model in terms of preserving delicate hair features.

5 Discussions

Although our model achieves a state-of-the-art performance compared to the baselines, several challenges still remain. First, since we transfer hairstyles via online latent optimization, it takes a few minutes on average for each image pair. Also, our framework cannot newly generate the occluded part of the target hair due to the extremely turned head pose. For example, the first three columns of Fig. 10 show that where the hair on the side is extremely occluded so that the final output barely has side hair. Finally, the output might contain undesired background when the hair segmentation mask is inaccurately predicted. The last three columns of Fig. 10 present undesirable background leaking.

6 Conclusions

This paper proposes a latent optimization framework for high-quality pose-invariant hairstyle transfer via local-style-aware hair alignment. By leveraging latent optimization, we align the target hair without a multi-view dataset, while maintaining fine details of the hairstyle. In addition, during the hair alignment, our newly-presented local-style-matching loss encourages our model to preserve the distinct structure and color of each local hair region in detail. Finally, we perform occlusion inpainting and blending via latent optimization. In this way, our model produces high-quality final output without noticeable artifacts.

Acknowledgments. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST) and the Ministry of Culture, Sports and Tourism and Korea Creative Content Agency (Project Number: R2021040097, Contribution Rate: 50).

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proc. of the IEEE international conference on computer vision (ICCV). pp. 4432–4441 (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proc. of the IEEE international conference on computer vision (ICCV) (2019)
3. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 8296–8305 (2020)
4. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(11), 2274–2282 (2012)
5. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proc. of the IEEE international conference on computer vision (ICCV) (2017)
6. 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)
7. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: Conference of the International Speech Communication Association (INTERSPEECH) (2018)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2016)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. the Advances in Neural Information Processing Systems (NeurIPS) (2014)
10. Harkonen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. In: Proc. the Advances in Neural Information Processing Systems (NeurIPS) (2020)
11. 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)
12. Jiang, W., Liu, S., Gao, C., Cao, J., He, R., Feng, J., Yan, S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2020)
13. Jo, Y., Park, J.: Sc-fegan: Face editing generative adversarial network with user's sketch and color. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2019)
14. 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)
15. 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)
16. 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)

17. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 5549–5558 (2020)
18. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612 (2017)
19. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: Proc. the International Conference on Learning Representations (ICLR) (2017)
20. Portenier, T., Hu, Q., Szabo, A., Bigdeli, S.A., Favaro, P., Zwicker, M.: Faceshop: Deep sketch-based face image editing. arXiv preprint arXiv:1804.08972 (2018)
21. Saha, R., Duke, B., Shkurti, F., Taylor, G., Aarabi, P.: Loho: Latent optimization of hairstyles via orthogonalization. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2021)
22. 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)
23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proc. the International Conference on Learning Representations (ICLR) (2015)
24. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Yuan, L., Tulyakov, S., Yu, N.: Michigan: Multi-input-conditioned hair image generation for portrait editing. ACM Transactions on Graphics (TOG) **39**(4), 1–13 (2020)
25. Viazovetskyi, Y., Ivashkin, V., Kashin, E.: Stylegan2 distillation for feed-forward image manipulation. In: Proc. of the European Conference on Computer Vision (ECCV) (2020)
26. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing (TIP) **13**(4), 600–612 (2004)
27. Xiao, C., Yu, D., Han, X., Zheng, Y., Fu, H.: Sketchhairsalon: Deep sketch-based hair image synthesis (2021)
28. Yang, S., Wang, Z., Liu, J., Guo, Z.: Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In: Proc. of the European Conference on Computer Vision (ECCV). pp. 601–617. Springer (2020)
29. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proc. of the European Conference on Computer Vision (ECCV) (2018)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2018)
31. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the IEEE conference on computer vision and pattern recognition (CVPR) (2018)
32. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. Proc. the International Conference on Learning Representations (ICLR) (2021)
33. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks (2021)
34. Zhu, P., Abdal, R., Qin, Y., Femiani, J., Wonka, P.: Improved stylegan embedding: Where are the good latents? arXiv preprint arXiv:2012.09036 (2020)

35. Zhuang, P., Koyejo, O., Schwing, A.G.: Enjoy your editing: Controllable gans for image editing via latent space navigation. In: Proc. the International Conference on Learning Representations (ICLR) (2021)