HairCLIP: Design Your Hair by Text and Reference Image Supplementary Material

Tianyi Wei[1], Dongdong Chen[2,†], Wenbo Zhou[1], Jing Liao[3],
Zhentao Tan[1], Lu Yuan[2], Weiming Zhang[1], Nenghai Yu[1]
[1]University of Science and Technology of China [2]Microsoft Cloud AI [3]City University of Hong Kong
{bestwty@mail., welbeckz@, tzt@mail., zhangwm@, ynh@}ustc.edu.cn
cddlyf@gmail.com, jingliao@cityu.edu.hk, luyuan@microsoft.com

	Afro Hairstyle				Bobcut Hairstyle				Bowlcut Hairstyle				Mohawk Hairstyle				Purple Hair			
Methods	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM
Ours	40.9	0.84	27.8	0.92	15.2	0.85	28.1	0.92	44.4	0.78	26.8	0.90	39.7	0.82	26.4	0.89	62.2	0.90	30.8	0.95
StyleCLIP	73.1	0.79	24.5	0.87	29.2	0.77	24.7	0.88	48.2	0.74	24.2	0.88	38.5	0.70	20.6	0.81	83.4	0.87	24.3	0.90
TediGAN	29.0	0.17	24.2	0.79	26.4	0.15	23.9	0.79	29.4	0.14	23.8	0.79	27.4	0.17	24.2	0.79	26.7	0.18	24.5	0.80
	Green Hair				Blond Hair				Braid Brown Hair				Crewcut Yellow Hair				Perm Gray Hair			
Methods	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM	FID	IDS	PSNR	SSIM
Ours	62.1	0.90	30.2	0.95	23.1	0.87	30.0	0.95	34.2	0.76	26.2	0.89	69.5	0.78	25.8	0.89	93.4	0.82	26.3	0.91
StyleCLIP	81.2	0.85	22.8	0.87	53.5	0.81	26.3	0.91	22.3	0.81	23.3	0.89	84.0	0.78	20.1	0.83	76.6	0.80	20.8	0.85
TediGAN	34.7	0.17	24.3	0.79	30.3	0.18	24.4	0.80	29.9	0.16	23.9	0.79	32.0	0.18	24.3	0.80	33.2	0.15	23.9	0.79

Table 1. Quantitative comparisons regarding FID [1] and the preservation of irrelevant attributes. FID measures the distributions of images before and after editing. IDS denotes identity similarity before and after editing. PSNR and SSIM are calculated in the region of intersection of non-hair regions before and after editing. Our approach exhibits the best irrelevant attributes preservation ability.

1. Quantitative Results

We give the detailed quantitative results about FID [1] in Table 1 for each hair description. Although TediGAN [6] performs the best in terms of FID, it hardly performs any hair editing task very well (as demonstrated by the qualitative results). We argue that FID may not be a suitable metric for evaluating the manipulation ability. Similar conclusion was also drawn in e4e [5].

2. Hyper-Parameters Sensitivity Analysis

In our experiments, we manually check the absolute value scale of each loss term and empirically set the loss weights to make them at the same scale to achieve the balance. And all the hyper-parameters are fixed in the experiments.

To analyze the hyper-parameter sensitivity of Eq.14, we modify the hyper-parameters from 2:1:1 (default setting) to 1:1:1, 1:2:1, and 1:1:2 respectively to retrain the network. Then user study with 30 participants and 100 groups of results (50 text-based and 50 reference-based) is conducted. The users are asked to select the best editing result while keeping other attributes unchanged. The detailed preference rates are shown in below Table 2:

$\lambda_t : \lambda_i : \lambda_{ap}$	2:1:1	1:1:1	1:1:2	1:2:1
Preference rate	29.87%	28.73%	26.63%	14.77%

Table 2. Hyper-parameter analysis for Eq.14

It shows that, when the ratio is set to 1:2:1 (larger image manipulation loss than other losses), it will result in poor performance because of worse text-based manipulation. And the remaining three settings achieve comparable results.

We further perform sensitivity analysis for hyper-parameters of Eq.13 in Table 3, by adjusting $\lambda_{id} \in \{0.3^*, 0.1, 0.5\}$, $\lambda_{s\_mc} \in \{0.02^*, 0.01, 0.03\}$,$\lambda_{bg} \in \{1^*, 0.5, 1.5\}$, $\lambda_{norm} \in \{0.8^*, 0.6, 1.0\}$, where * mean the default values. It shows that our model is relatively sensitive to λ_{id} (because of large absolute loss scale) and stable for other hyper-parameters.

† Dongdong Chen is the corresponding author. Our code is available at https://github.com/wty-ustc/HairCLIP

	IDS ↑	PSNR ↑	SSIM ↑	ACD ↓
Ours (default*)	0.85	27.0	0.91	0.02
$\lambda_{id}(0.1/0.5)$	0.61/0.90	24.8/29.0	0.85/0.94	0.03/0.02
$\lambda_{s\_mc}(0.01/0.03)$	0.85/0.86	27.0/27.6	0.91/0.92	0.03/0.02
$\lambda_{bg}(0.5/1.5)$	0.83/0.86	25.6/28.3	0.89/0.93	0.02/0.02
$\lambda_{norm}(0.6/1.0)$	0.86/0.86	27.1/27.9	0.91/0.92	0.02/0.02

Table 3. Hyper-parameter analysis for Eq.13

3. User Study for Network Structure Ablation Analysis

To quantitatively evaluate the network structure design, we ask 30 volunteers to conduct user study with 100 randomly picked result groups, and ask them to select the hair editing results that best match the text description. The preference rates are shown in Table 4, and demonstrate the superiority of our new network designs.

Ours	variant (a)	variant (b)	variant (c)
71.07%	11.27%	5.23%	12.43%

Table 4. Quantitative results to demonstrate the superiority of our new network designs. (a) concatenate conditional inputs with the latent code. (b) replace the conditional inputs of coarse and medium sub hair mappers with hair color embedding, and fine sub hair mapper with hairstyle embedding. (c) replace the conditional input of medium sub hair mapper with the hair color embedding and leave the rest unchanged.

4. More Qualitative Results

In Figures 1, 2, and 3 we give more visual comparison results with other state-of-the-art methods and results for the cross-modal conditional inputs.

References

[1] 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. In *NIPS*, 2017. 1

[2] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[3] Rohit Saha, Brendan Duke, Florian Shkurti, Graham W. Taylor, and Parham Aarabi. Loho: Latent optimization of hairstyles via orthogonalization. In *CVPR*, 2021. 4

[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):95–1, 2020. 4

[5] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40:1 – 14, 2021. 1

[6] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2256–2265, 2021. 1, 3

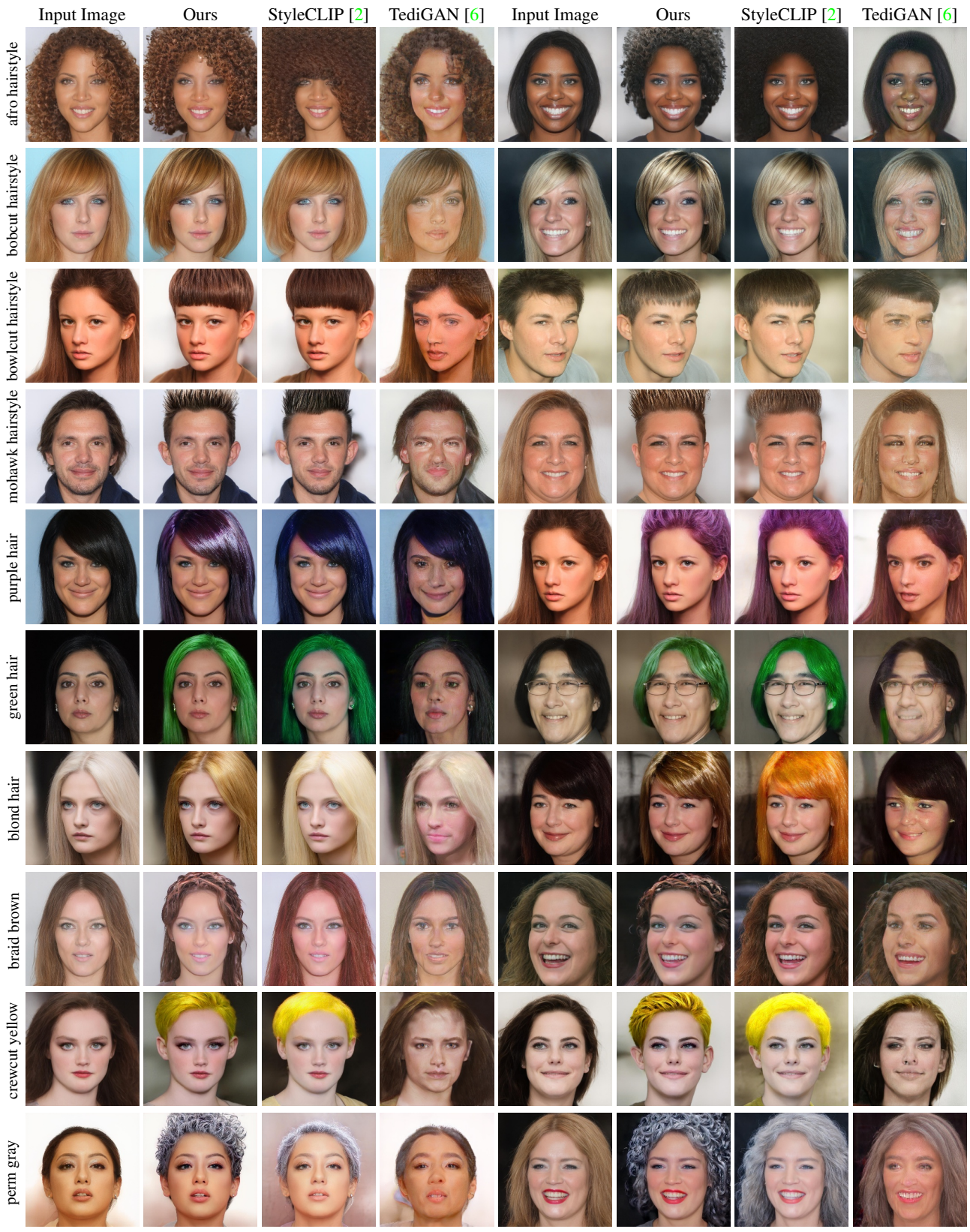

Figure 1. Visual comparison with StyleCLIP [2] and TediGAN [6]. The corresponding simplified text descriptions (editing hairstyle, hair color, or both of them) are listed on the leftmost side of each row, and all input images are the inversions of the real images. Our approach demonstrates better visual photorealism and irrelevant attributes preservation ability while completing the specified hair editing.

Figure 2. Comparison of our approach with LOHO [3] and MichiGAN [4] on hair transfer. Even for extreme examples like the third row in this figure, our method can yield plausible hairstyle transfer results.

Figure 3. Our single framework supports hairstyle and hair color editing individually or jointly, and conditional inputs can come from either image or text domain. "Style" in the text description is the abbreviation for hairstyle.