Supplementary Material – CLIP2Protect: Protecting Facial Privacy using Text-Guided Makeup via Adversarial Latent Search

Fahad Shamshad Muzammal Naseer Karthik Nandakumar
Mohamed Bin Zayed University of AI, UAE
{fahad.shamshad, muzammal.naseer, karthik.nandakumar}@mbzuai.ac.ae

In this document, we provide details of the makeup text prompts (Sec. 1), datasets description along with preprocessing steps (Sec. 2), and additional experiments under face verification and identification tasks (Sec. 3 and 4 respectively). We also provide a comparison of quantitative results in terms of PSNR and SSIM with baseline methods (Sec. 5), followed by the effectiveness of our approach against commercial FR API of Tencent (Sec. 6).

1. Makeup Text Prompts

We have collected 40 makeup text styles from online resources to guide the adversarial optimization in the proposed approach. Details about these makeup text styles are provided in Tab. 1.

2. Datasets Description and Preprocessing

In this section, we provide a detailed description of the datasets used in the experiments along with the processing steps. We use CelebA-HQ [6] and LADN [3] datasets for impersonation attack under the face verification tasks. For other settings, we use CelebA-HQ and LFW [5] datasets. The datasets demonstrate the generalization of our methods on both high-quality (CelebA-HQ) and low-quality (LFW) face images, as the generative models we used are trained on high-quality images.

CelebA-HQ [6]. It is a high-resolution version of CelebA dataset [7] and consists of 30,000 images having resolution of 1024×1024. We use 1000 images corresponding to different identities as provided by Hu *et al.* [4].

LADN [3]. It is a makeup-based dataset consisting of 333 non-makeup images and 302 makeup images. We use it for impersonation attack under the face verification task only. Similar to [4], we use 332 images from the non-makeup images. We split these images into four groups, where images in each group aim to impersonate the same target identity. For experimentation, we use the four target identities provided by Hu *et al.* [4].

LFW [5]. LFW is a widely used face identification dataset consisting of 13,233 images and 5,749 identities. We use it for face verification (dodging) and face identification (impersonation and dodging) tasks. For experiments, we select 500 pairs, where each pair belongs to the same identity. For identification, we assign one image in the pair to the gallery set and the other to the probe set. Both impersonation and dodging attacks are performed on the probe set.

Figure 1. Target identities used by [4] for impersonation attack under face verification task. Top row represents images used during training, and bottom row shows images used for evaluation. It mimics a realistic scenario as target images used in the optimization phase differ from those during evaluation.

Preprocessing. Consistent with the previous works, we use MTCNN [13] to detect, crop and align the face image before giving it as input to FR models. For all datasets, we also do preprocessing following the official paper [8] for the *latent code initialization* stage.

3. Dodging Attack under Face Verification

In this section, we provide results of dodging attack under the face verification task for CelebA-HQ and LFW datasets. The result of the impersonation attack under the verification task is in the main paper. For experiments, we select 500 subjects at random, and each subject has a pair of faces. Quantitative results in terms of Protection Success Rate (PSR) under a black-box setting are shown in Tab. 4. As Adv-Makeup [12] and AMT-GAN [4] are trained for the impersonation attack, these are not included in the comparison.

Table 1. Makeup text styles used in our experiments.

	Makeup Text
1	*Tanned makeup*
2	*Pale makeup*
3	*Makeup*
4	*Heavy makeup*
5	*Heavy makeup with red lipstick*
6	*Makeup with purple lipstick*
7	*Funky makeup*
8	*Celebrity makeup*
9	*Dewy makeup*
10	*Matte makeup*
11	*Light makeup with pink eyeshadows*
12	*Soft glam makeup*
13	*Retro makeup*
14	*Ultra glamm makeup*
15	*Vintage makeup*
16	*Shimmer powder makeup*
17	*HD makeup*
18	*Editorial makeup*
19	*Avant Garde Makeup*
20	*Drag Queen Makeup*
21	*Smokey makeup*
22	*No makeup*
23	*Pink eyeshadows*
24	*Clown makeup*
25	*Tanned Makeup with black lipstick*
26	*Vintage makeup*
27	*Big eyebrows with pink eyeshadows*
28	*Tanned makeup with purple lipstick*
29	*Red lipstick with purple eyeshadows*
30	*Pale makeup with red lipstick*
31	*Black eyeshadows with purple lipstick*
32	*Rosy cheeks makeup*
33	*Tanned Makeup with red lipstick*
34	*Purple cheeks makeup with pink lipstick*
35	*Big eyebrows*
36	*Bridal makeup*
37	*Anti-Aging makeup*
38	*Clown makeup with purple lipstick*
39	*Gothic makeup*
40	*Big eyelashes with pink eyeshadows*

4. Results on CelebA for Face Identification

In this section, we provide results of targeted (impersonation) and untargeted (dodging) attacks on CelebA-HQ dataset under the task of face identification. For the experiment, we randomly select 500 subjects, each with a pair of faces. We assign one image in the pair to the gallery set and the other to the probe set. Both impersonation and dodging attacks are performed on the probe set. Quantitative results in terms of Protection Success Rate (PSR) under a black-box setting are shown in Tab. 3. For impersonation, we insert four target identities provided by Hu *et al.* [4] into the gallery set. The results on LFW dataset under the same settings are provided in the main paper.

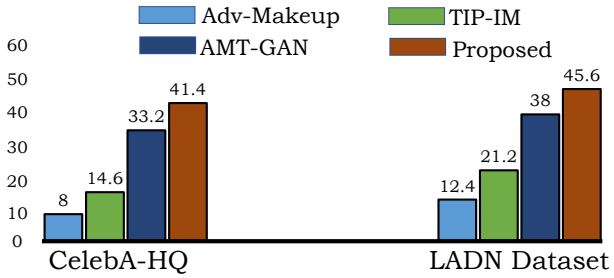

Figure 2. Average confidence score (higher is better) returned by a real-world face verification API, Tencent, for impersonation attack. Our approach has a higher confidence score than state-of-the-art makeup and noise-based facial privacy protection methods.

5. PSNR and SSIM

In this section, we provide quantitative results in terms of PSNR and SSIM [9]. Our method has inferior performance compared to TIP-IM and comparable performance to AMT-GAN for PSNR and SSIM. However, as shown in Tab. 4 of the main manuscript, the proposed approach has a lower FID score, indicating that the outputs generated via our method have a more natural appearance (see Fig. 3). The drop in PSNR and SSIM compared to AMT-GAN can be due to the error between the original image and the inverted image during the GAN inversion (*latent code initialization*) stage. We believe that the progress in the GAN inversion field [10] can help reduce this error.

6. Tencent Face Comparison API

Tencent face comparison API returns confidence scores between 0 to 100 to measure whether two images are similar or not, where a high confidence score indicates high similarity. As the training data and model parameters of these propriety `FR` models are unknown, it effectively mimics a real-world scenario. We protect 100 faces that are randomly selected from CelebA-HQ and LADN datasets using the baselines and the proposed method. In Fig. 2, we show the average confidence score returned by Tencent face comparison API against these images. The results indicate that our method has a high Protection Success Rate compared to baselines.

Figure 3. Qualitative results generated by TIP-IM [11], AMT-GAN [4] and our approach for *black-box* impersonation attack under the face verification task. The first two columns are the original images and the target identity. From top to bottom, the text makeup styles used in our method are *"purple lipstick"*, *"red lipstick"*, *"pink lipstick with big eyebrows"*, *"tanned makeup"*, *"pink lipstick"*, *"pale makeup with pink eyeshadows"*, and *"pale makeup with pink lipstick"*. Best viewed in zoom in.

Table 2. Protection success rate (PSR %) of *black-box* dodging attack under the face verification task. For each column, the other three FR systems are used as surrogates to generate the protected faces.

Method	CelebA-HQ				LFW				Average
	IRSE50	IR152	FaceNet	MobileFace	IRSE50	IR152	FaceNet	MobileFace	
TIP-IM$_{(ICCV'21)}$ [11]	71.2	69.4	88.2	59.0	71.8	76.1	80.6	62.9	72.4
Ours	**83.4**	**83.6**	**93.5**	**62.8**	**79.6**	**80.2**	**86.5**	**73.3**	**80.4**

Table 3. Protection success rate (PSR %) of *black-box* dodging (top) and impersonation (bottom) attacks under the face identification task for CelebA-HQ dataset [5]. For each column, the other three FR systems are used as surrogates to generate the protected faces. R1-U: Rank-1-Untargeted, R5-U: Rank-5-Untargeted, R1-T: Rank-1-Targeted, R5-T: Rank-5-Targeted.

Method	IRSE50		IR152		FaceNet		MobileFace		Average	
	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U
TIP-IM$_{(ICCV'21)}$ [11]	79.6	61.2	62.9	42.8	46.2	27.8	81.9	76.7	67.6	52.1
Ours	**88.5**	**72.3**	**69.0**	**46.2**	**58.5**	**31.7**	**94.7**	**82.6**	**77.7**	**58.2**
	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
TIP-IM$_{(ICCV'21)}$ [11]	16.2	51.4	21.2	56.0	8.1	35.8	9.6	24.0	13.8	41.8
Ours	**24.5**	**64.7**	**24.2**	**65.2**	**12.5**	**38.7**	**11.8**	**28.2**	**18.2**	**49.2**

Table 4. Protection success rate (PSR %) of *black-box* dodging attack under the face verification task. For each column, the other three FR systems are used as surrogates to generate the protected faces.

Method	PSNR	SSIM
TIP-IM$_{(ICCV'21)}$ [11]	33.21	0.92
AMT-GAN$_{(CVPR'22)}$ [4]	19.50	0.79
Ours	19.31	0.75

7. Limitations and Future Directions

Our approach takes around 70 seconds to protect a single high-resolution image of size 1024×1024 on A100 GPU with 40 GB memory. The *latent code initialization stage* takes around 50 seconds, and *the text-guided adversarial optimization stage* takes about 20 seconds. On the other hand, although it takes less than a second for AMT-GAN to protect a high-resolution image, it requires re-training of around 13 hours every time for a new target identity. As our approach is generative, therefore it can be quickly adapted to different target identities at test time without computationally expensive model re-training.

In the future, we aim to replace the iterative latent code initialization stage with a single forward pass following the recent works regarding trainable mapper-based generator fine-tuning [1, 2]. This can considerably reduce the execution time of the proposed approach.

References

[1] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 4

[2] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11389–11398, 2022. 4

[3] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10481–10490, 2019. 1

[4] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15014–15023, 2022. 1, 2, 3, 4

[5] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 1, 4

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[8] 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(4):1–14, 2021. 1

[9] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 2

[10] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[11] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV'21)*, pages 3897–3907, 2021. 3, 4

[12] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 1

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016. 1