Supplementary Materials of Recoverable Facial Identity Protection via Adaptive Makeup Transfer Adversarial Attacks

Xiyao Liu[1], Junxing Ma[1], Xinda Wang[2], Qianyu Lin[1], Jian Zhang[1 *], Gerald Schaefer[3], Cagatay Turkay[4], Hui Fang[3 *]

[1]School of Computer Science and Engineering, Central South University, China
[2]School of Software and Microelectronics, Peking University, China
[3]Department of Computer Science, Loughborough University, U.K.
[4]Centre for Interdisciplinary Methodologies, University of Warwick, U.K.
{lxyzoewx, mjx2021}@csu.edu.cn, 2401210713@stu.pku.edu.cn, {8212220928, jianzhang}@csu.edu.cn, gerald.schaefer@ieee.org, cagatay.turkay@warwick.ac.uk, h.fang@lboro.ac.uk

This supplementary document provides implementation details for method reproducibility, additional experimental results and the limitations of our method which are not included in our main manuscript.

Implementation Details

Network Architecture

In this section, we describe the architectures of the networks in RMT-GAN. We construct the architecture of generator G, discriminator D and recovery module R in RMT-GAN.

Makeup-transfer and Adversarial Attack Generator The makeup-transfer and adversarial attack generator G consists of two sub-networks and an attentive makeup morphing module(AMM): (i) G_{MD} is used for makeup distilling, the detailed architecture of it is the encoder-bottleneck architecture used in (Choi et al. 2018) without decoder part. It extracts the makeup style from the neighbour target. It disentangles the makeup related features, e.g., lip gloss, and eye shadows, from the intrinsic facial features, e.g., facial shape, and the size of eyes. (ii) G_{MA} applies the makeup to the source image. It utilises a similar encoder-bottleneck-decoder architecture as (Choi et al. 2018). (iii) AMM morphs the makeup metrics extracted by G_{MD} and fuses the makeup style with the source image. Note that the encoder part of G_{MA} shares the same architecture with G_{MD}, but they do not share parameters. In the encoder section, we utilise instance normalisations without affine parameters, transforming the feature map into a normal distribution.

We enforce the FR loss to ensure the generated image being recognised as its neighbour target, which encourages G to constrain adversarial attacks to the makeup regions.

Discriminator The architecture of discriminator D follows PatchGANs (Li and Wand 2016). The discriminator is trained to compete with the generator G and the recovery module R to ensure the realistic synthesis and recovery.

*Corresponding Authors.

Algorithm 1 Training procedure.

Input: Source image set S , their adaptive target ts , generator G , recovery module R , discriminator D , local white-box FR models F , optimiser $Adam$
Parameter: Iterations T, hyper-parameters Λ of loss weights
Output: Parameters w_G, w_D and w_R for trained G, R and D for RMT-GAN

Initialise w_G, w_D, w_R;
for $i = 0$ to $T - 1$ **do**
  Randomly select source $s \in S$ and its corresponding adaptive target t as input of generator G;
  Optimise D to minimise L_D with fixed G and R ;
  Calculate L_D;
  $w_D \leftarrow Adam(w_D, L_D)$;
  Optimise G to minimise L_G with fixed D and R ;
  Calculate L_G;
  $w_G \leftarrow Adam(w_G, L_G)$;
  Optimise R to minimise L_R with fixed G and D ;
  Calculate L_R;
  $w_R \leftarrow Adam(w_R, L_R)$;
end forreturn w_G, w_D, w_R.

Recovery module We employ an encoder-decoder structure for the Recovery module R, which utilises Residual-in-Residual Dense Blocks (RRDB) as its key blocks. RRDBs have been widely used in image super-resolution tasks. These blocks are capable of eliminating adversarial perturbations. With the help of the module, we can effectively recover images with high-quality details and reduce distortions.

Training procedure

To train the generator G, discriminator D, and the recovery module R, we optimise them sequentially as illustrated in algorithm 1. The images in the training dataset MT(Makeup Transfer) dataset (Li et al. 2018; Chen et al. 2019) are with a resolution of 361×361, for all experiments, we resize them to 256×256. In our implementation, we train RMT-GAN on a 24GB RTX 3090 GPU, and the batch-size is set to 1. We

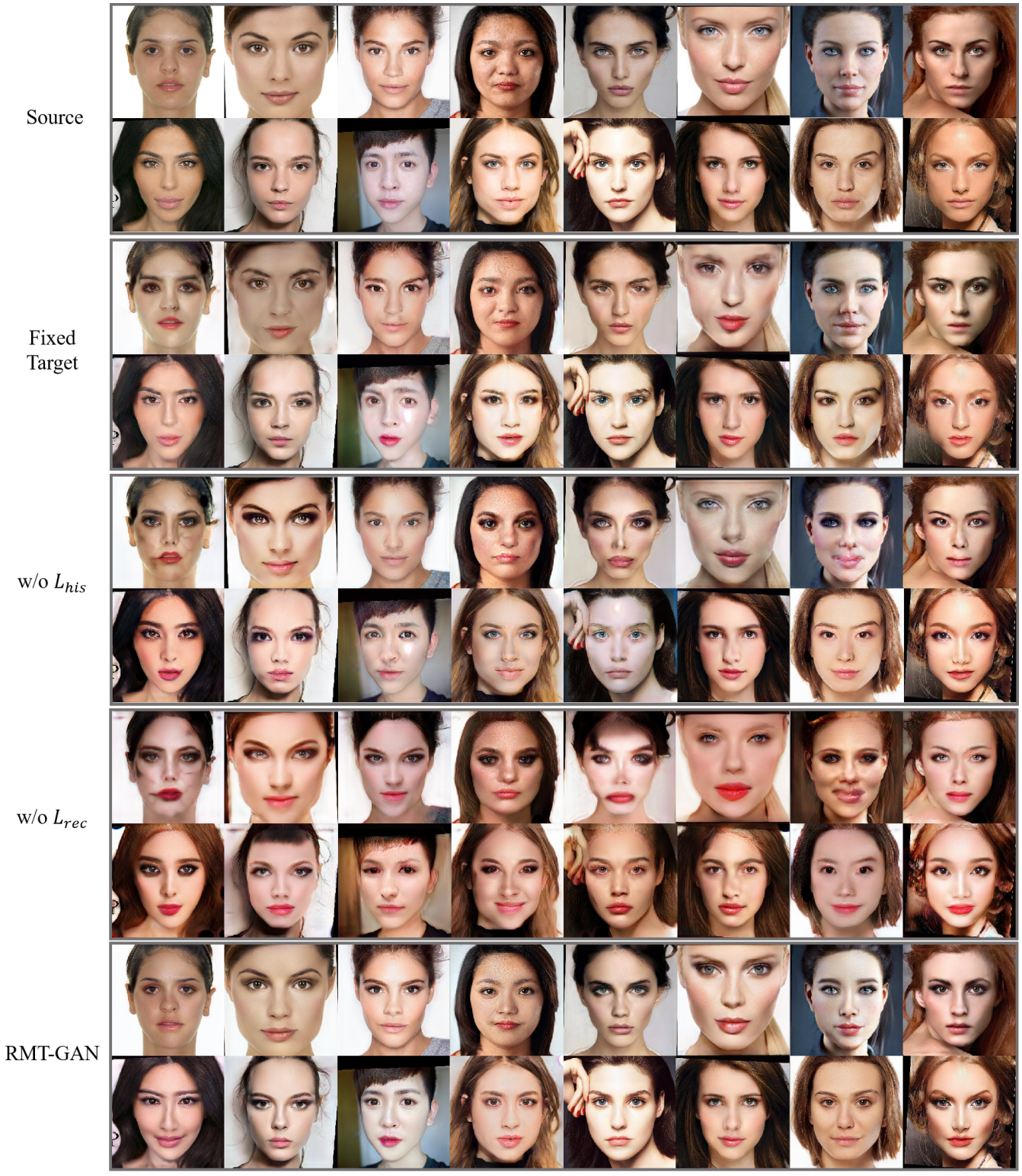

Figure 1: Generated examples of different designs in the ablation study.

use the Adam optimiser (Kinga, Adam et al. 2015) with an initial learning rate of 0.0002 and exponential decay rates of $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The loss weights λ_{adv}, λ_{cyc}, λ_{make}, λ_{sr}, and λ_{rec} are set (heuristically) to 0.6, 1, 0.4, 0.5, and 1, respectively.

	MobileFace	FaceNet	IR152	IRSE50
CelebA-HQ				
source	19.4	1.6	4.0	9.0
w/o L_G^{his}	89.2	26.7	34.6	83.2
w/o $L_{G,R}^{sr}$	93.2	30.5	47.8	90.8
RMT-GAN	93.4	31.0	51.8	91.4
LADN dataset				
source	24.6	6.6	5.7	10.8
w/o L_G^{his}	97.6	68.8	60.7	97.0
w/o $L_{G,R}^{sr}$	99.1	72.4	71.8	98.2
RMT-GAN	99.4	76.9	80.8	98.5

Table 1: Ablation results for L_G^{his} and $L_{G,R}^{sr}$ in terms of ASR. ASR results for all models when training on three FR models and testing on the hold-out model(e.g. the column of MobileFace means that the model is trained with FaceNet, IR152 and IRSE50 and tested on MobileFace).

Ablation Study

	FID(↓)	PSNR(↑)	SSIM(↑)
CelebA-HQ			
w/o L_G^{his}	23.908	20.837	0.803
w/o $L_{G,R}^{sr}$	79.436	13.108	0.622
RMT-GAN	21.254	21.171	0.811
LADN dataset			
w/o L_G^{his}	34.998	20.775	0.803
w/o $L_{G,R}^{sr}$	90.603	12.998	0.574
RMT-GAN	29.777	21.654	0.810

Table 2: Ablation results for L_G^{his} and $L_{G,R}^{sr}$ in terms of image quality.

We conduct an ablation study to evaluate the impacts of the auxiliary losses (the histogram matching loss L_G^{his} and the self-reconstruction loss $L_{G,R}^{sr}$) as shown in Tables 1 and 2 in terms of ASR and image quality of adversarial images, respectively.

From Tables 1 and 2, we can see that when omitting the histogram matching loss term, the ASRs decrease for all four black-box test models, while image quality is also slightly degraded. On the other hand, omitting the self-reconstruction loss term leads to a significant drop in image quality and a slight ASR decrease. Thus, both losses are shown to make important contributions to our model.

In the following, the generated examples for ablation study are illustrated in Figure 1. It further confirms the importance of the loss terms as well as the better synthesising effect when compared to the fixed target strategy.

Recovery Examples

We have demonstrated the effectivenss of our designed identity recovery module in terms of ASR in our main manuscript. In Figure 2, we show some recovery examples that further demonstrate that the recovered images are visually identical to their source images after makeup removal. This result also indicates that our designed identity recovery module is effective.

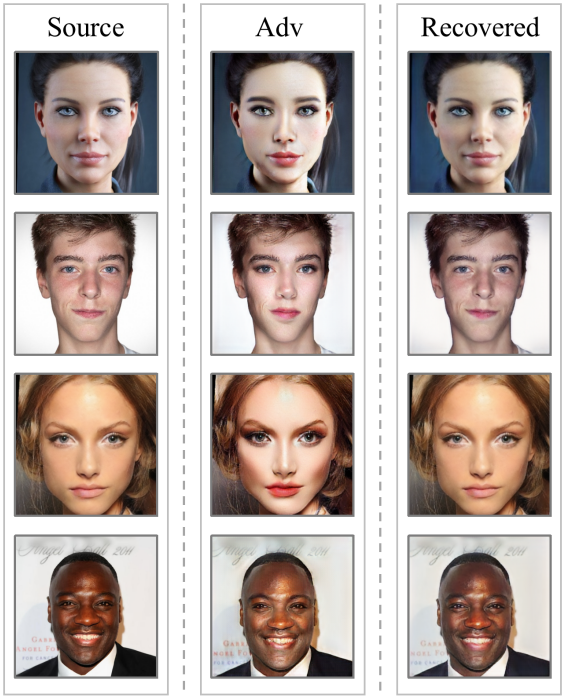

Figure 2: Identity recovery examples.

Limitations

Although the designed method is recoverable with superior performances compared to other makeup methods, our study can be further improved. Firstly, our method can only perform makeup transfers on holistic face. One of our future focuses is to have more flexibility on local makeup transfers, e.g., the eyes or lips, with high visual quality and attack success rate. Secondly, our method is mainly tested by applying to two FR systems in the real-world applications due to the free trial permission limitation and we will further test the effectiveness of our method on various unauthorised FR systems in the future.

References

Chen, H.-J.; Hui, K.-M.; Wang, S.-Y.; Tsao, L.-W.; Shuai, H.-H.; and Cheng, W.-H. 2019. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10042–10050.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Kinga, D.; Adam, J. B.; et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, 6. San Diego, California;.

Li, C.; and Wand, M. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks.

In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 702–716. Springer.

Li, T.; Qian, R.; Dong, C.; Liu, S.; Yan, Q.; Zhu, W.; and Lin, L. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, 645–653.