Appendix For Paper: Protecting Facial Privacy: Generating Adversarial Identity Masks via Style-robust Makeup Transfer

1. Contents

- Sec. 2: Detailed experiment settings for the proposed AMT-GAN, competitors and evaluation metrics. And the urls of resource codes for reproducing our method and evaluations.
- Fig. 1: Some visual results of AMT-GAN.
- Fig. 2: Some visual results of ablation study about regularization module.

2. Experiment Settings

2.1. Implementation Details.

In this section, we will introduce the detailed experiment settings of our proposed AMT-GAN.

We first implement our framework by Pytorch. The generator and the discriminators in our method are based on the official code for PSGAN[1], which is an efficient framework for makeup transfer. But please note that the AMT-GAN can be built on most existing makeup transfer frameworks or new ones as long as the chosen GAN has cycle consistency constraints. The regularization module we used in our method is based on the *residual-in-residual dense block* (RRDB). We implement this block following their official codes[2].

2.2. Competitors

In our experiments, We use PGD, MI-FGSM, TI-DIM, Adv-Makeup and TIP-IM to serve as the competitors for comparison. Here, we introduce their detailed settings.

For the gradient-based methods (*i.e.*PGD, MI-FGSM, and TI-DIM), we use these characters as the notations of specific parameters:

parameter	notation
maximum perturbation	ϵ
perturbation size per-step	α
number of steps	t
momentum decay factor	μ
the probability of applying input diversity	p
resize factor used in input diversity	r

PGD. We re-implement the PGD attack (as well as MI-FGSM and TI-DIM) based on *Torchattacks*[3], which is a PyTorch library that contains adversarial attacks to generate adversarial examples and to verify the robustness of deep learning models. Specifically, the settings of PGD are:

ϵ	α	t
16 / 255	0.8 / 255	10

MI-FGSM. The settings of MI-FGSM are:

ϵ	α	t	μ
16 / 255	0.8 / 255	10	1.0

TI-DIM. The settings of TI-DIM are:

ϵ	α	t	μ	p	r
16 / 255	0.8 / 255	10	1.0	0.7	0.9

In addition, We use Gaussian kernel and set kernel size to 15×15 following their official paper.

Adv-Makeup. We implement Adv-Makeup in our experiments by their official codes[4]. In addition, to make the Adv-Makeup be trained on different datasets rather than the specific dataset attached with the official codes, we bulid an auxiliary Pytorch program to process different datasets based on Face++ landmark[5].

TIP-IM. We use the official codes[6] of TIP-IM and set $\gamma = 5 \times 10^{-4}$ with other settings as default. Please note that in the official codes of TIP-IM, the γ is not the same γ (in the aspect of magnitude) in their paper.

[1] https://github.com/wtjiang98/PSGAN

[2] https://github.com/xinntao/ESRGAN

[3] https://github.com/Harry24k/adversarial-attacks-pytorch

[4] https://github.com/TencentYoutuResearch/Adv-Makeup

[5] https://www.faceplusplus.com.cn/face-detection/

[6] https://github.com/ShawnXYang/TIP-IM

2.3. Evaluation Metrics.

ASR. In our experiments, we use *attack success rate* (ASR) to evaluate the attack ability of different methods, which is formulated as:

$$ASR = \frac{\sum_{n=1}^{N} \Gamma(\cos[M(z), M(x_n)])}{N} \times 100\%, \quad (1)$$

where M is the target model, z stands for the face image with target identity, and $\{x_n\}_{n=1,\cdots,N}$ represents the adversarial images. The $\Gamma(\cdot)$ is an indicator function:

$$\Gamma(x) = \begin{cases} 1, x > \tau, \\ 0, x \leq \tau, \end{cases} \quad (2)$$

the τ in this function is the threshold parameter which differs in different *false acceptance rate* (FAR) of target FR models. In our experiments, we calculate ASR with FAR@0.01, so the τ of each target model will be set to 0.241, 0.167, 0.409 and 0.302 for IRSE50, IR152, Facenet and Mobileface respectively following Adv-Makeup.

PSNR and SSIM. *Peak Signal-to-Noise Ratio* (PSNR)[7] and *Structural Similarity* (SSIM)[8] are popular methods for image quality assessment. In our paper, we use them to evaluate the image quality of the outputs from different methods and discuss the gap between human eyes and these quantitative metrics in perturbation-based adversarial examples and makeup transferred images. Specifically, we use scikit-image[9] for fast implementation.

FID. *Fréchet Inception Distance* (FID) is a crucial method in the field of GANs, which measures the similarity between two data distributions. In practice, researchers often use FID to evaluate the similarity between generated images and nature images, investigate the ability of a generator in terms of generating natural images. In our paper, we extend FID to adversarial machine learning and use them to measure the naturalness of adversarial examples. We benefit from the well-developed package *clean-fid*[10] for fast implementation.

[7] https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

[8] https://en.wikipedia.org/wiki/Structural_similarity

[9] https://scikit-image.org/

[10] https://github.com/GaParmar/clean-fid

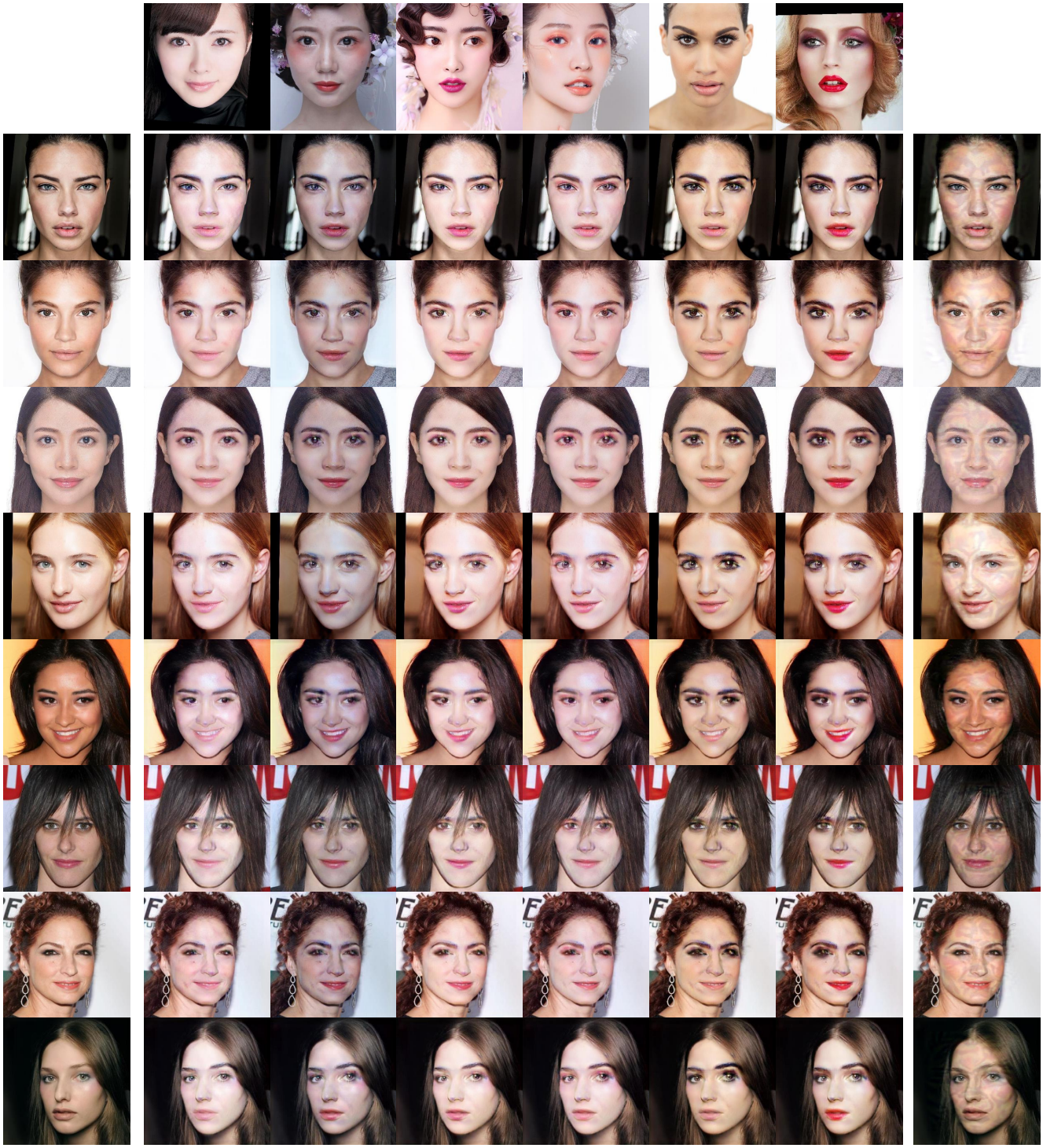

Figure 1. Results from AMT-GAN and TIP-IM. Here we show the images from the same outputs dataset which we use for evaluations in our paper. The images on the top row are the cosmetic references for makeup transfer. Please zoom in for a better view.

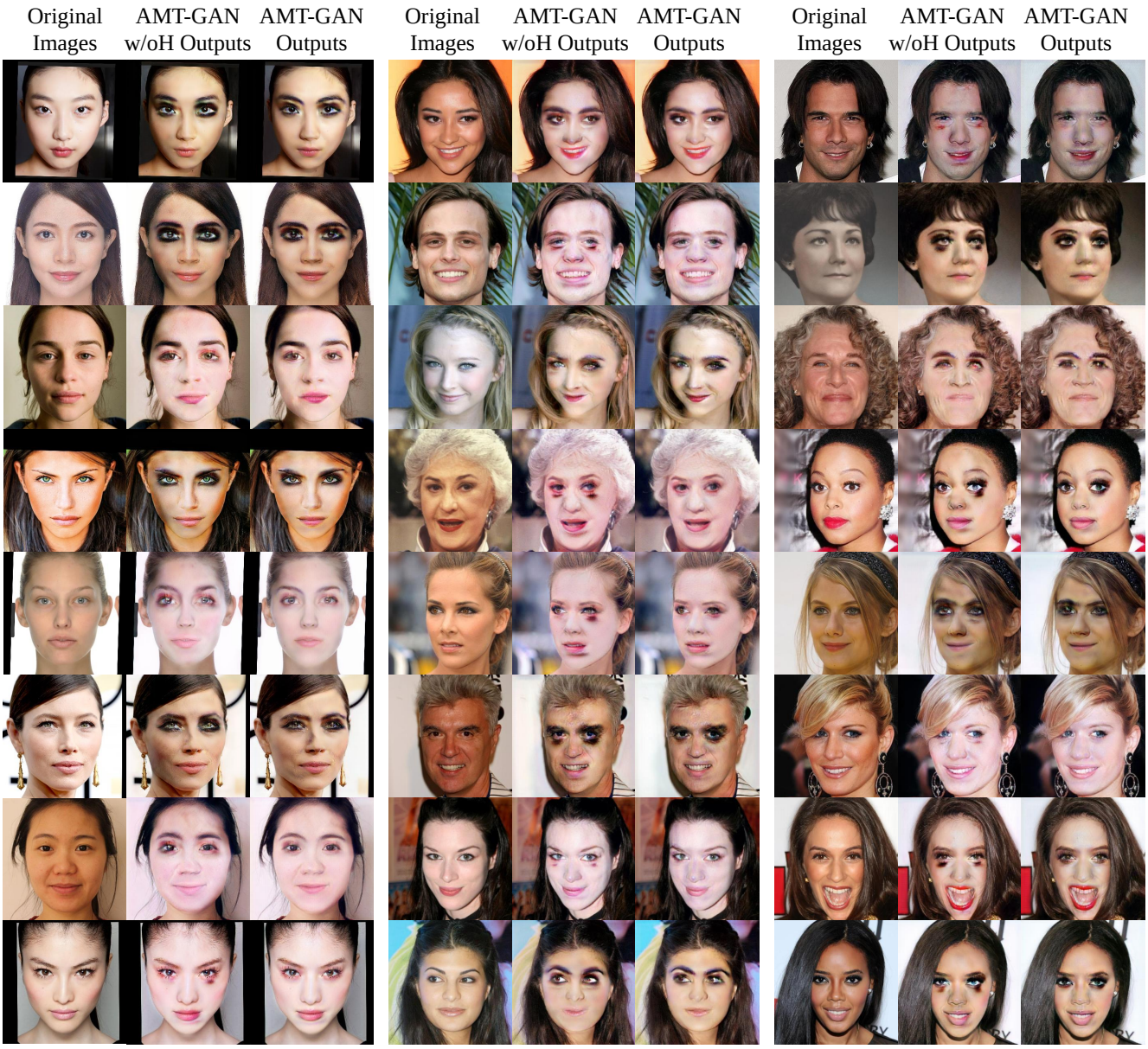

Figure 2. Visual results of ablation study about regularization module. The images from the generator trained without the regularization module suffer from fake shadows, distortion of structure information, unaligned makeup position, etc., which are typical indications of weak domain mappings, caused by damaged cycle consistency loop by adversarial toxicity in the training phase. As the visual results here and the quantitative results in our paper have shown, the regularization module can eliminate or alleviate this phenomenon. Notably, as makeup transfer is still in development and may have some little issues, a small minority of images (no matter with or without regularization module) may have asymmetrical eye-shadow, which is beyond the scope of our investigation in this paper. Please zoom in for a better view.