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

Abstract

Unauthorised face recognition (FR) systems have posed significant threats to digital identity and privacy protection. To alleviate the risk of compromised identities, recent makeup transfer-based attack methods embed adversarial signals in order to confuse unauthorised FR systems. However, their major weakness is that they set up a fixed image unrelated to both the protected and the makeup reference images as the confusion identity, which in turn has a negative impact on both attack success rate and visual quality of transferred photos. In addition, the generated images cannot be recognised by authorised FR systems once attacks are triggered. To address these challenges, in this paper, we propose a Recoverable Makeup Transferred Generative Adversarial Network (RMT-GAN) which has the distinctive feature of improving its image-transfer quality by selecting a suitable transfer reference photo as the target identity. Moreover, our method offers a solution to recover the protected photos to their original counterparts that can be recognised by authorised systems. Experimental results demonstrate that our method provides significantly improved attack success rates while maintaining higher visual quality compared to state-of-the-art makeup transfer-based adversarial attack methods. Our code and supplementary materials are available on Github.

Github — github.com/tttianyu/RMT-GAN

Introduction

Face recognition is one of the core identity authentication techniques with wide applications (Huang and Chen 2022; Huang et al. 2023). Meanwhile its privacy protection has raised increasing research attention since unauthorised face recognition (FR) systems pose a significant threat to recent FR applications (Zhong and Deng 2022; Shan et al. 2020). When a large number of facial images are collected from social media, the identity and privacy of individuals are in danger of getting compromised by unauthorised FR systems (Hill 2020; Shoshitaishvili, Kruegel, and Vigna 2015). Consequently, there is an urgent need to defend applications against such FR systems in order to protect user privacy.

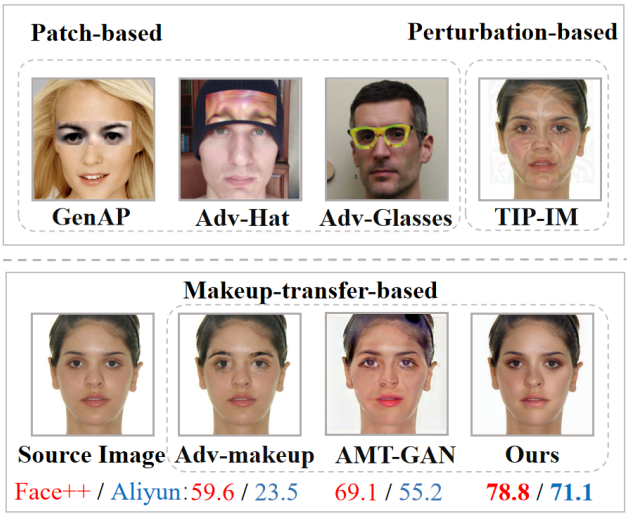

Figure 1: Comparison of our proposed RMT-GAN with existing adversarial attacks on FR systems with a black box attack setting. The numbers indicate the verification confidence of the target identity given by the commercial FR API Face++ and Aliyun. Our proposed method provides better visual quality and a higher attack success rate compared to current SoTA methods.

Recently, various methods have been proposed to enhance privacy protection. For example, data poisoning attacks on training data of malicious FR models are an effective way to alleviate the identity and privacy leakage problem (Shan et al. 2020; Cherepanova et al. 2021). In practice, however, it is difficult to inject such poisonous data in the training phase since attackers can hardly access the training datasets (Hu et al. 2022). Another strategy is to generate adversarial photos to avoid identity compromise (Cherepanova et al. 2021; Dong et al. 2019b). As illustrated in Figure 1, patch-based adversarial attacks (Sharif et al. 2019; Xiao et al. 2021) and perturbation-based attacks (Yang et al. 2021) are exploited to alleviate the privacy issue. They are capable of reducing the recognition confidence of true identities, thus yielding a

*Corresponding Authors.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

high attack success rate (ASR). However, their adversarial patterns are easily noticeable, limiting their application for facial identity protection on social media. To enhance imperceptibility, makeup transfer based models (MTMs) (Hu et al. 2022; Yin et al. 2021) are designed to embed adversarial patterns by generating more realistic photos. In principle, a reliable FR system is capable of recognising photos of the same identity with different makeup styles, while a MTM uses multi-task learning to provide flexibility that can embed pixel-level perturbations to enhance the imperception of adversarial patterns when transferring makeup styles.

Although recent makeup transfer-based protection methods achieve reasonable ASRs with good visual quality, some limitations hinder their application to real-world scenarios. First, to ensure control on identity protection, fixed target photos unrelated to both the protected and the makeup reference images are selected before attacking to inject adversarial signals to protect photos although this degrades both ASR and visual quality due to its strong constraint. Especially when a target face shares low geometric similarities to a protected face, the high non-linear mapping further increases the difficulty level of the transfer task. Second, once adversarial signals are embedded into their protected photos, none of these photos can be recognised by authorised FR systems (i.e., systems that have permission to perform face verification authorised by the users who have uploaded their identity images), thus failing the essential requirement of FR applications such as facial recognition payment, access control and facial check-in systems.

In this paper, we propose a Recoverable Makeup Transferred Generative Adversarial Network (RMT-GAN) as an effective solution to these problems. As illustrated in Figure 1, our method surpasses existing state-of-the-art (SoTA) makeup transfer protection techniques, and does so through two distinctive features. First, we design an adaptive target selection strategy to improve high-quality style transfer while maintaining control to avoid identity leakage by unauthorised FR systems. Second, we apply a cycle consistency loss in our model training to ensure the recovery of recognisable face photos. After enforcing this term, the reuse of adversarial photos is guaranteed with an additional recovery module.

Our novel contributions in this paper are as follows:

- we propose a new recoverable makeup transfer GAN for identity privacy protection. Our approach outperforms SoTA methods by a large margin on extensive experiments, including black-box attack evaluation in terms of ASR, and visual quality evaluation via a well-designed user study with various image examples;
- we present an adaptive target selection algorithm to improve ASR and visual quality of adversarial examples to confuse unauthorised FR systems;
- through an intuitive cycle consistency design, our model enables the recovery of face photos which are recognisable by authorised FR systems even if adversarial patterns are embedded. To our best knowledge, this is the first recoverable makeup method.

Related Work

Facial Privacy Protection

Development in FR techniques has seen rapid advancements (Schroff, Kalenichenko, and Philbin 2015; Deng et al. 2019) and deployment in various commercial systems such as Face++ and Aliyun. These approaches are further widely applied in various downstream tasks to provide enhanced security by verifying user identities. However, as a serious side-effect, automatic FR identity verification may (Hill 2020; Shoshitaishvili, Kruegel, and Vigna 2015). The wide use of social media makes it easy to obtain a large collection of face images with identity information, and this data can be exploited by unauthorised FR systems to track user behaviours or to abuse their identities in applications which require face verification.

To address the concern of privacy breach, obfuscation-based methods have been extensively studied. Conventional approaches here introduce image processing techniques, such as pixelation, darkening, blurring and occlusion (Wilber, Shmatikov, and Belongie 2016), to reduce recognition confidence of FR systems. Alternatively, generative adversarial network (GAN)-based algorithms generate obfuscation while maintaining high visual quality (Gafni, Wolf, and Taigman 2019; Sun et al. 2018). However, neither produce natural and realistic images. Although GAN-based methods produce more realistic face images, the synthesised facial appearance still looks unnatural due highlighting distinctive features.

Adversarial Attacks for Privacy Protection

Inspired by recent research on adversarial attacks (Madry et al. 2018; Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2014), several methods generate adversarial examples to defend face privacy (Hu et al. 2022; Yang et al. 2021; Yin et al. 2021). These approaches can be divided into white-box and black-box attacks. Earlier work focusses on white-box attacks, which embed adversarial signals when FR model parameters are accessible (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018). These are however not applicable against unauthorised third-party FR models since their parameters are unknown to privacy protection practitioners. In contrast, black-box attacks inject adversarial patterns in order to protect face identities when the unauthorised FR models are not transparent. Among these, query-based methods interactively retrieve responses from a black-box model which can be used to further refine adversarial examples (Dong et al. 2019b; Guo et al. 2021). However, in real-world scenarios, it is unlikely to be known which FR model an unauthorised party deploys and therefore transferable black-box methods, which seek adversarial examples to defend generalised FR models, are under extensive investigation (Xiao et al. 2021; Yang et al. 2021; Dong et al. 2018). In addition, recent work realizes the importance of untargeted attacks for the FR system in real-world scenarios and proposes a method performing both effective targeted and untargeted adversarial attacks (Zhou et al. 2024).

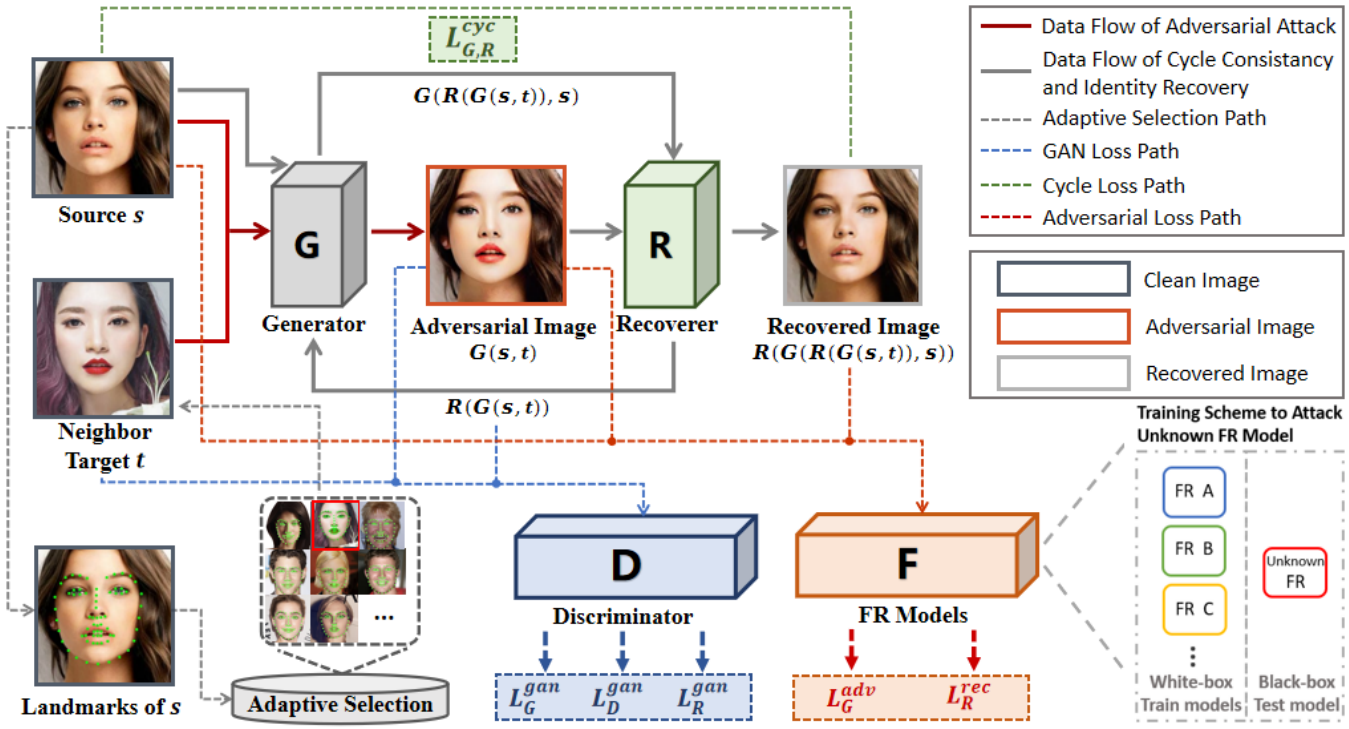

Figure 2: RMT-GAN training process: our framework consists of a generator G, a discriminator D, a recovery module R, and a set of white-box FR models F.

Makeup Transfer-based Adversarial Attacks

Makeup transfer, a style transfer task (Choi et al. 2018) that aims to synthesise realistic face images with different makeup styles on a source face, for adversarial attacks on face recognition systems was initially proposed in (Zhu, Lu, and Chiang 2019). Since the goal of adversarial attacks on FR systems is two-fold – achieving both a high ASR and photo-realistic images – it is intuitive to apply this multi-task training strategy on face privacy protection. Unlike the white-box setting in (Zhu, Lu, and Chiang 2019), Adv-Makeup (Yin et al. 2021) and AMT-GAN (Hu et al. 2022) employ a black-box attack setting which is more suitable for real-world scenarios. Adv-Makeup applies face makeup only to eye regions. As a patch-based method, it also generates artefacts around these regions. AMT-GAN proposes a regularisation module to reconcile the conflicts between adversarial noises and the cycle consistency loss in makeup transfer. Clip2Protect (Shamshad, Naseer, and Nandakumar 2023) uses CLIP model in the training process to ensure the similarity between real face images and makeup-transferred face, enhancing the makeup performances.

While the approach we propose in this paper shares similar ideas with Adv-Makeup and AMT-GAN, it is distinguished from and surpasses them due to the following features: (i) instead of setting fixed target faces, we present an adaptive target face selection strategy which can improve both the visual quality of adversarial examples and their ASRs; and (ii) our method is capable of recovering faces with original identities that can be recognised by authorised FR systems after attacks.

Method

Problem Formulation

To effectively protect user privacy, our RMT-GAN generates adversarial images via makeup transfer embedding. After the transferring stage (i.e., adversarial attacking), an unauthorised FR system will recognise an image as an incorrect target identity. For authorised FR systems, we introduce a makeup adversarial pattern removal module to recover images of recognisable identities.

For the transferring stage, we train a model to embed a style that minimises the distance between representations of a transferred face and a target face encoded by any unauthorised black-box FR systems, which can be expressed as

$$\min_{s_A} L_{adv} = D(F_k(s_A), F_k(t)), \tag{1}$$

where $D(\cdot)$ is a distance function, e.g., cross-entropy or cosine similarity, F_k represents a deep neural network (DNN) feature extractor of a face recognition model, s_A denotes an adversarial face image via makeup transfer, and t is the target image(in our method, the target image and the reference image are the same). Here, s_A appears to have the same identity with source image s and the same makeup style with t while being recognised as the reference t's identity by unauthorised FR.

To ensure correct recovery of the original identity after makeup adversarial attacks, we propose an identity recovery

Algorithm 1 Adaptive target image selection.

Input: source face image s
Parameter:68 landmarks of s l_s , landmarks images in target set T , L_2 distance between l_s and l_t $dis_{s,t}$, adaptive threshold τ , pre-defined ratio ra ,targets with distances less than τ $T_{<\tau}$
Output: adaptive target t

$l_s \leftarrow lm(s)$
$l_T \leftarrow lm(T)$
for t ∈ T **do**
  $dist_{s,t} \leftarrow \| l_s, l_t \|_2$
end for
$\tau \leftarrow Cal\_thre(ra)$
$T_{<\tau} \leftarrow t \in T$ **where** $dist_{s,t} < \tau$
$t \leftarrow random\_select(T_{<\tau})$

module to allow the identification from an authorised FR system. This module is co-trained with our makeup transfer network in order to guarantee its seamless connection with the previous transferring stage. This can be formulated as

$$\min_{s'} L_{rec} = D(F_k(s'), F_k(s)), \tag{2}$$

where s' represents the recovered images after removing adversarial makeup.

Adaptive Target Image Selection

A deep neural network normally performs well in learning a non-linear mapping function for the makeup transfer task even if the geometries of source and target faces are distinctive from each other. However, we can see from the result of AMT-GAN in Figure 1, artefacts are introduced that degrade the synthesis quality when simultaneously applying adversarial attacks and makeup transfer to transfer a source face to a pre-defined target face.

In our approach, we propose a simple yet effective strategy to select target faces, which is also capable of further improving the ASR as well as the visual quality of synthesised images. The workings of our adaptive target face selection module are given, in form of pseudo-code, in Algorithm 1. Noting that for each source image, we calculated the distance of each image in the reference image dataset from it and randomly selected one of the top 20% closest reference images as the adaptive target. Since target images are selected from faces sharing similar geometries to our source faces, we potentially enhance the synthesis quality when hiding adversarial signals. In addition, it also provides more flexibility to enforce additional constraints, e.g., the cycle-consistency recovery loss.

RMT-GAN

In Figure 2, we illustrate the training stage of our approach. Our proposed RMT-GAN consists of the following key components: a generator G to synthesise adversarial examples from source images; an identity recovery module R to recover images with original identities; a discriminator D to distinguish between real and makeup transferred face images; and an ensemble FR module F which contains a set of white-box FR models. We adopt the network architectures of the generator and discriminator in PSGAN (Jiang et al. 2020) for G and D. For R, since its main function is to remove adversarial signals and recover original identities, we adapt the residual-in-residual dense block (RRDB) network from (Hu et al. 2022).

In the following, we detail the various loss terms that we employ in our approach. As a two-player zero-sum game between generator and discriminator, the **GAN-based loss** term is expressed as

$$L_D^{gan} = -\log D(s) - \log(1 - D(G(s,t)), \tag{3}$$

and

$$L_G^{gan} = -\log(D(G(s,t))), \tag{4}$$

where L_D^{gan} and L_G^{gan} are the GAN losses of the makeup-transfer adversarial examples generator G and the discriminator D, respectively, s is the source image of the makeup transfer, and t is the target(makeup-reference) image.

The **FR loss** minimises the feature representation distance between a synthesised face and its corresponding target face. To enhance model generalisability in a black box configuration, we extract the feature representations from various FR systems in order to provide a solution capable of defending an arbitrary unauthorised FR system. The loss term is defined as

$$L_G^{adv} = \frac{1}{N}\sum_{k=1}^{N}(1 - \cos_s(F_k(t), F_k(T(G(s,t),p)))), \tag{5}$$

where $\cos_s(\cdot)$ represents the cosine similarity function to measure feature similarity, F_k is the feature extractor of the k-th of N FR models (which do not include the black-box model to be attacked), $T(\cdot)$ is a transformation function, and p is a predefined probability of whether to transform $G(s,t)$. As transformation functions, we choose image resizing and Gaussian noise addition.

The **identity recovery cycle consistency loss** is related to the recovery module R which is similar to mimicking a reversible process in order to recover a face of the original identity. The original cycle loss is obtained as

$$L^{cyc} = ||G(G(s,t),s) - s||_1, \tag{6}$$

where $||\cdot||_1$ represents the L_1 norm. To ensure seamless integration of our generator and recovery networks, we redesign the loss as

$$L_{G,R}^{cyc} = ||R(G(R(G(s,t)),s)) - s||_1. \tag{7}$$

The **identity recovery network FR loss** allows enhancing the recovered identity representations which can be recognised by any authorised FR system. Similar to the FR loss to confuse unauthorised systems, we minimise the feature distance between the recovered face and the face of original identity as

$$L_R^{rec} = \frac{\sum_{k=1}^{N}(1 - \cos_s(F_k(s), F_k(R(G(R(G(s,t)),s)))))}{N} \tag{8}$$

where $R(G(R(G(s,r)),s))$ has removed makeup and adversarial attacks.

	CelebA-HQ				
	MobileFace	FaceNet	IR152	IRSE50	Average
source	12.7	1.1	3.8	7.3	6.2
FGSM (Goodfellow, Shlens, and Szegedy 2015)	16.4 (+3.7)	1.2 (+0.1)	17.1 (+13.3)	42.2 (+34.9)	19.2 (+13.0)
PGD (Madry et al. 2018)	49.7 (+37.0)	2.3 (+1.2)	19.6 (+15.8)	45.8 (+38.5)	29.4 (+23.2)
MI-FGSM (Dong et al. 2018)	45.9 (+33.2)	2.6 (+1.5)	25.0 (+21.2)	45.8 (+38.5)	29.8 (+23.6)
TI-DIM (Dong et al. 2019a)	57.1 (+44.4)	15.3 (+14.2)	36.2 (+32.4)	63.6 (+56.3)	43.1 (+36.9)
TIP-IM (Yang et al. 2021)	79.0 (+66.3)	11.5 (+10.4)	50.0 (+46.2)	72.1 (+64.8)	53.2 (+47.0)
Adv-Makeup (Yin et al. 2021)	22.0 (+9.3)	1.4 (+0.3)	9.5 (+5.7)	17.2 (+9.9)	12.5 (+6.3)
AMT-GAN (Hu et al. 2022)	50.7 (+38.0)	16.6 (+15.5)	35.1 (+31.3)	77.0 (+69.7)	44.9 (+43.7)
CLIP2Protect (Shamshad, Naseer, and Nandakumar 2023)	75.2 (+62.5)	41.7 (+40.6)	48.4 (+44.6)	81.1 (+73.8)	61.6 (+55.4)
source*	19.4	1.6	4.0	9.0	8.5
RMT-GAN	93.4 (+74.0)	31.0 (+29.4)	51.8 (+47.8)	91.4 (+82.4)	66.9 (+58.4)
	LADN dataset				
	MobileFace	FaceNet	IR152	IRSE50	Average
source	5.1	0.6	3.6	2.7	3.0
FGSM (Goodfellow, Shlens, and Szegedy 2015)	10.2 (+5.1)	1.2 (+0.6)	3.9 (+0.3)	11.4 (+8.7)	6.7 (+3.7)
PGD (Madry et al. 2018)	11.1 (+ 6.0)	2.1 (+1.5)	4.5 (+0.9)	13.2 (+10.5)	7.7 (+4.7)
MI-FGSM (Dong et al. 2018)	45.0 (+39.9)	6.3 (+5.7)	25.6 (+22.0)	48.9 (+46.2)	31.5 (+28.5)
TI-DIM (Dong et al. 2019a)	48.3 (+43.2)	22.1 (+21.5)	34.2 (+30.6)	56.4 (+53.7)	40.3 (+37.3)
TIP-IM (Yang et al. 2021)	55.3 (+50.2)	15.1 (+14.5)	50.5 (+46.9)	45.3 (+42.6)	41.6 (+38.6)
Adv-Makeup (Yin et al. 2021)	22.4 (+17.3)	1.0 (+0.4)	10.0 (+6.4)	29.6 (+26.9)	15.8 (+12.8)
AMT-GAN (Hu et al. 2022)	72.4 (+67.3)	32.1 (+31.5)	49.1 (+45.5)	89.6 (+86.9)	60.8 (+57.8)
CLIP2Protect (Shamshad, Naseer, and Nandakumar 2023)	79.9 (+74.8)	47.9 (+47.3)	53.3 (+49.7)	91.6 (+88.9)	68.2 (+65.2)
source*	24.6	6.6	5.7	10.8	11.9
RMT-GAN	99.4 (+74.8)	76.9 (+70.3)	80.8 (+75.1)	98.5 (+87.7)	88.9 (+77.0)

Table 1: 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). Here, * means that the ASR of the source image of our method is different from other methods because of the adaptive target image selection, and the numbers in brackets indicate the improvements compared to the rates before makeup transfer.

To ensure realistic recovery, we also enforce the **discriminator loss**

$$L_R^{gan} = -\log(D(R(G(s,t)))) \tag{9}$$

as an extra loss term to train the recovery network.

We include two more constraints as **auxiliary losses**. Histogram matching is typically used to match the colour distribution of reference while maintaining the content information from source image (Li et al. 2018). To ensure its holistic image consistency, we exploit this term, expressed as

$$L_G^{his} = ||G(s,t) - HM(s,t)||_2, \tag{10}$$

where $HM(s,t)$ represents histogram matching and $||\cdot||_2$ the L_2 norm. We further add a self-reconstruction loss to ensure G and R maintain the original face structure of the source image as it is important for the generator to avoid distortion of face attributes. The self-reconstruction loss is defined as

$$L_{G,R}^{sr} = ||R(G(s,s) - s)||_1 + LPIPS(R(G(s,s)), s), \tag{11}$$

where $LPIPS(\cdot)$ is the learned perceptual image patch similarity (Zhang et al. 2018), which we employ to measure the visual similarity between two images.

The **total losses** for D, G, and R are then obtained as

$$L_D = L_D^{gan}, \tag{12}$$

$$L_G = L_G^{gan} + \lambda_{adv} L_G^{adv} + \lambda_{cyc} L_{G,R}^{cyc} + \lambda_{his} L_G^{his} + \lambda_{sr} L_{G,R}^{sr}, \tag{13}$$

and

$$L_R = L_R^{gan} + \lambda_{cyc} L_{G,R}^{cyc} + \lambda_{rec} L_R^{rec} + \lambda_{sr} L_{G,R}^{sr}, \tag{14}$$

respectively, where the λs represent loss weight hyper-parameters.

More details of network architecture and the training procedure are provided in Supplementary Materials due to the page limitation.

Experiments

Experimental Settings

Datasets Following (Hu et al. 2022; Li et al. 2018; Chen et al. 2019), we use the Makeup Transfer (MT) dataset (Li et al. 2018) as our training dataset, which contains 1,115 non-makeup and 2,719 makeup images. Our test data we draw from two public datasets, 500 non-makeup/500 makeup images from CelebA-HQ (Karras et al. 2018) and 333 non-makeup/302 makeup images from the LADN dataset (Gu et al. 2019) as source/reference images.

SoTA Benchmarks We select FGSM (Goodfellow, Shlens, and Szegedy 2015), PGD (Madry et al. 2018), MI-FGSM (Dong et al. 2018), TI-DIM (Dong et al. 2019a), TIP-IM (Yang et al. 2021), Adv-Makeup (Yin et al. 2021), AMT-GAN (Hu et al. 2022) and Clip2Prtoect (Shamshad, Naseer, and Nandakumar 2023) as our benchmark models for comparison. FGSM and PGD are classic gradient-based attacks, while Adv-Makeup, AMT-GAN and Clip2Prtoect are makeup-transfer based methods for attacking FR systems (and thus closer related to our method).

Unauthorised FR Models Following (Yin et al. 2021) and (Hu et al. 2022), we conduct experiments to attack four popular FR models, IR152 (He et al. 2016), IRSE50 (Hu, Shen, and Sun 2018), FaceNet (Schroff, Kalenichenko, and

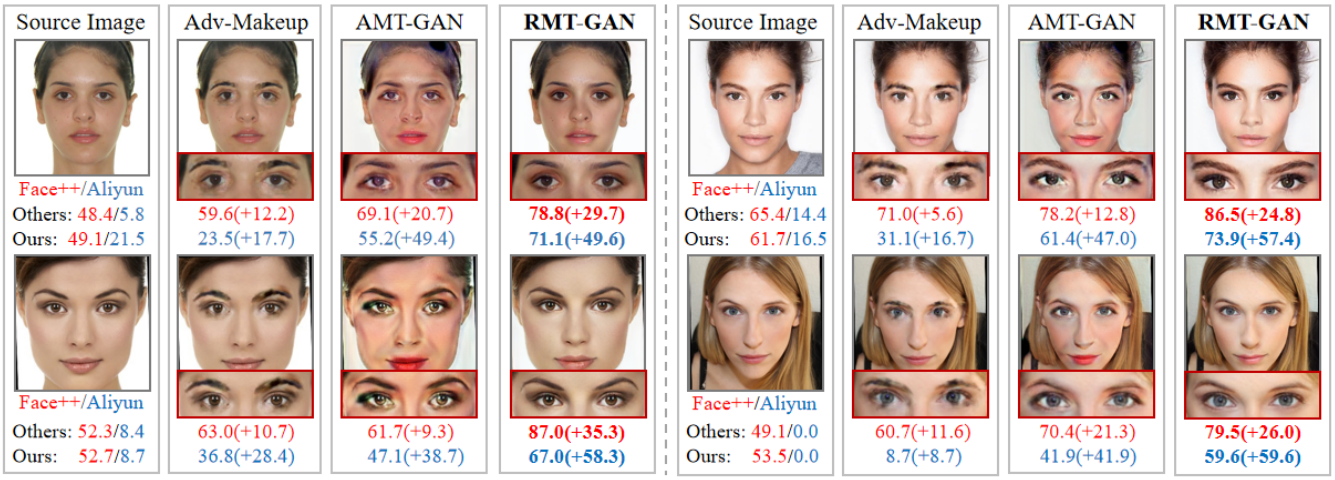

Figure 3: Examples of adversarial images generated by Adv-Makeup, AMT-GAN and our RMT-GAN. Given are also the verification confidences of the target identity from Face++ (red) and Aliyun (blue).

Philbin 2015), and MobileFace (Deng et al. 2019), as well as two commercial FR APIs, Face++ and Aliyun, as unauthorised black-box models for testing.

Evaluation Metrics To evaluate the performance to confuse unauthorised FR models, we calculate the attack success rate (ASR) as used in (Hu et al. 2022; Xiao et al. 2021; Yin et al. 2021), and report the ASR at FAR@0.01 for black-box testing. For commercial FR APIs, we directly obtain the confidence scores returned by the FR servers, with a high confidence score indicating that the server believes the two input images to be of the same identity. To assess visual quality, we use FID (Heusel et al. 2017), PSNR and SSIM (Wang et al. 2004). FID measures the distance between two data distributions, which is often used to evaluate whether a generated dataset looks as realistic as the source dataset. In contrast, PSNR and SSIM are widely-used metrics to measure the visual similarity between two images. For identity recovery evaluation, we employ two metrics, the ASR of the recovered image by removing adversarial signals, and the visual similarity between the recovered image and its corresponding source image.

Results and Comparison with SoTA

Black-box Attack Evaluation We assess the attack success rate under a black-box attacking configuration using a leave-one-out approach where we use three FR models as training models and the left-out model as a black-box unauthorised FR system for testing. Since our adaptive selected target images are different from the fixed target image used in benchmark methods, we further compare the increments of ASR after attack for fair comparisons. The results on both CelebA-HQ and LADN-datasets are reported in Table 1.

From Table 1, it is evident that our proposed RMT-GAN outperforms the other benchmark methods by achieving the best average ASR to confuse the black-box unauthorised FR models. The reason for this is the use of our well-designed adaptive target selection strategy, which dentifies the most suitable target image for makeup transfer.

	FID(↓)	PSNR(↑)	SSIM(↑)
CelebA-HQ			
Adv-Makeup (Yin et al. 2021)	3.630	36.164	0.996
AMT-GAN (Hu et al. 2022)	23.227	20.204	0.796
CLIP2Protect (2023)	37.117	19.354	0.603
RMT-GAN	21.254	21.171	0.811
LADN dataset			
Adv-Makeup (Yin et al. 2021)	10.689	31.199	0.991
AMT-GAN (Hu et al. 2022)	34.440	19.504	0.787
CLIP2Protect (2023)	51.923	19.643	0.595
RMT-GAN	29.777	21.656	0.800

Table 2: Image quality results of adversarial examples.

Visual Quality Evaluation We compare the visual quality of the images of our method to other makeup transfer-based methods, i.e., AMT-GAN, Adv-Makeup and Clip2Protect. The obtained FID, PSNR and SSIM results of the generated adversarial images are given in Table 2 for both datasets. RMT-GAN outperforms AMT-GAN and Clip2Protect because our adaptive selection of target identity leads to significantly improved image quality. Since Adv-Makeup only transfers the eye regions which take up only a small part of the entire face, it gives the best FID/PSNR/SSIM results. However, as can be seen in Figure 3, it introduces unwanted artefacts. Further, we have demonstrated that our method has achieved much higher ASR when compared to Adv-makeup method.

We also conduct two evaluation experiments by applying the methods to two commercial FR APIs, Face++ and Aliyun, as unauthorised black-box FR systems. Figure 3 shows some adversarial images generated by the three makeup transfer-based methods and their returned confidence values from the two FR systems. More adversarial examples are provided in Supplementary Materials.

As can be seen, the adversarial examples generated by our method are more natural compared to those of the other two methods. In particular, the adversarial examples generated

	Adv-Makeup	AMT-GAN	RMT-GAN
mean score	55.083±15.031	62.762±15.891	83.283±5.612
best rate(%)	4.0	8.0	88.0

Table 3: Results of conducted user study.

	source	adversarial	recovered
CelebHQ-A			
MobileFace	19.4	93.4	17.2
FaceNet	1.6	31.0	1.6
IR152	4.0	51.8	3.2
IRSE50	9.0	91.4	7.6
LADN dataset			
MobileFace	24.6	99.4	21.0
Facenet	6.6	76.9	8.4
IR152	5.7	80.8	5.1
IRSE50	10.8	98.5	12.0

Table 4: Identity recovery results in terms of ASR. ASR results for all models are trained on three FR models and tested on the hold-out model.

by AMT-GAN exhibit pseudo-shadows while Adv-Makeup generates obvious perturbations in the eye regions. Clearly, our method achieves the best confidence values to confuse the unauthorised FR system.

We conducted a user study with 25 participants with normal vision (14 male, 11 female; aged 21 to 52, average age 27.8 ± 6.7) to further compare the synthesised image quality of the three makeup transfer-based methods. Participants were informed about the purpose of the study but not about any of our hypotheses. Participants were given the dataset of source images and three result sets generated by the three methods, and asked to rate the images according to their naturalness, visual quality, and makeup-transfer effectiveness while setting the source images as a benchmark on a score of 100. The results are given in Table 3.

As is apparent from there, our RMT-GAN clearly outperforms the other two methods, yielding a mean score of 83.283, which is much higher than those of Adv-Makeup (55.083) and AMT-GAN (62.762), while 88.0% of participants found our generated images to be superior to those of the other approaches in terms of image naturalness, visual quality, and makeup transfer effectiveness.

Identity Recovery Evaluation We evaluate identity recovery of adversarial images by calculating the ASR of original images being recognised as target images, the ASR of adversarial images being recognised as target images, and the ASR of images after recovery being recognised as target images, and list the obtained results in Table 4. As we can see from there, the ASRs of source images and the ASRs of the images after recovery are similar, while significantly lower compared to ASRs of adversarial images that attack different black-box FR models, confirming that our identity recovery module works well. While there are some deviations of the outputs from the recovery network, resulting in some slightly lower/higher ASRs compared to the original images, these differences are not significant. In addition, visual examples of recovery images are provided in our Supplementary Materials.

	CelebA-HQ			
	MobileFace	FaceNet	IR152	IRSE50
source (fixed)	12.7	1.1	3.8	7.3
RMT-GAN (fixed)	54.8 (+42.1)	19.3 (+18.2)	33.7 (+29.9)	58.4 (+51.1)
source (adaptive)	19.4	1.6	4.0	9.0
RMT-GAN (adaptive)	93.4 (+74.0)	31.0 (+29.4)	51.8 (+47.8)	91.4 (+82.4)
	LADN dataset			
	MobileFace	FaceNet	IR152	IRSE50
source (fixed)	5.1	0.6	3.6	2.7
RMT-GAN (fixed)	63.2 (+58.1)	44.2 (+37.6)	47.7 (+42.0)	69.3 (+66.6)
source (adaptive)	24.6	6.6	5.7	10.8
RMT-GAN (adaptive)	99.4 (+74.8)	76.9 (+70.2)	80.8 (+75.1)	98.5 (+87.7)

Table 5: Target selection ablation results in terms of ASR. ASR results for all models are trained on three FR models and tested on the hold-out model. Here, the fixed/adaptive in the brackets means the target images are fixed or selected adaptively.

	FID(↓)	PSNR(↑)	SSIM(↑)
CelebA-HQ			
fixed	24.626	20.573	0.810
adaptive	21.254	21.171	0.811
LADN dataset			
fixed	33.508	20.879	0.797
adaptive	29.777	21.654	0.800

Table 6: Target selection ablation results.

Ablation Study

In this section, we conduct an ablation study to evaluate the effectiveness of our designed adaptive target selection algorithm by evaluating the black-box ASR and image quality. The impacts of the auxiliary losses are provided in our Supplementary Materials. As is apparent from Table 5, using adaptively selected targets for makeup transfer leads to significantly higher ASR results compared to fixed targets. In addition, as confirmed in Table 6, the image quality of the generated adversarial images is also higher when employing adaptive target selection. Our adaptive selection strategy avoids a domain conflict between the target image and the makeup transfer image, thus yielding improved ASRs and more natural adversarial images.

Conclusions

In this paper, we have proposed RMT-GAN, a novel recoverable adversarial image generation method based on makeup transfer for protecting face privacy. We design an adaptive target face selection scheme, which sets the makeup reference image and the identity target image as the same to enhance both attack success rate and visual quality. In addition, our model introduces an identify recovery module to ensure the reuse of adversarial images for authorised systems, such as access control and facial check-in systems. Extensive experiments demonstrate the effectiveness of the each designed component and the superiority of RMT-GAN compared to other state-of-the-art methods in terms of attack and makeup performances. In future work, we aim to design an effective area restrict strategy for makeup methods to allow more flexibility on local makeup transfer, and improve the effectiveness of our method on various unauthorised FR systems.

Acknowledgments

This research is supported by the Natural Science Foundation of Hunan Province, China (2022GK5002, 2024JK2015, 2024JJ5440), the National Natural Science Foundation of China (62472446), and the Special Foundation for Distinguished Young Scientists of Changsha (kq2209003), the foundation of State Key Laboratory of High Performance Computing, National University of Defense Technology (202401-13), and the High Performance Computing Center of Central South University.

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.

Cherepanova, V.; Goldblum, M.; Foley, H.; Duan, S.; Dickerson, J. P.; Taylor, G.; and Goldstein, T. 2021. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019a. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4312–4321.

Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; and Zhu, J. 2019b. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7714–7722.

Gafni, O.; Wolf, L.; and Taigman, Y. 2019. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9378–9387.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

Gu, Q.; Wang, G.; Chiu, M. T.; Tai, Y.-W.; and Tang, C.-K. 2019. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10481–10490.

Guo, Y.; Wei, X.; Wang, G.; and Zhang, B. 2021. Meaningful adversarial stickers for face recognition in physical world. *CoRR*, arXiv–2104.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hill, K. 2020. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*, 170–177. Auerbach Publications.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Hu, S.; Liu, X.; Zhang, Y.; Li, M.; Zhang, L. Y.; Jin, H.; and Wu, L. 2022. 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*, 15014–15023.

Huang, B.; Wang, Z.; Wang, G.; Jiang, K.; Han, Z.; Lu, T.; and Liang, C. 2023. PLFace: progressive learning for face recognition with mask bias. *Pattern Recognition*, 135: 109142.

Huang, Y.-H.; and Chen, H. H. 2022. Deep face recognition for dim images. *Pattern Recognition*, 126: 108580.

Jiang, W.; Liu, S.; Gao, C.; Cao, J.; He, R.; Feng, J.; and Yan, S. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5194–5202.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

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.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–823.

Shamshad, F.; Naseer, M.; and Nandakumar, K. 2023. CLIP2Protect: Protecting Facial Privacy Using Text-Guided Makeup via Adversarial Latent Search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20595–20605.

Shan, S.; Wenger, E.; Zhang, J.; Li, H.; Zheng, H.; and Zhao, B. Y. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Security Symposium*.

Sharif, M.; Bhagavatula, S.; Bauer, L.; and Reiter, M. K. 2019. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3): 1–30.

Shoshitaishvili, Y.; Kruegel, C.; and Vigna, G. 2015. Portrait of a privacy invasion. *Proc. Priv. Enhancing Technol.*, (1): 41–60.

Sun, Q.; Tewari, A.; Xu, W.; Fritz, M.; Theobalt, C.; and Schiele, B. 2018. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 553–569.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wilber, M. J.; Shmatikov, V.; and Belongie, S. 2016. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–9. IEEE.

Xiao, Z.; Gao, X.; Fu, C.; Dong, Y.; Gao, W.; Zhang, X.; Zhou, J.; and Zhu, J. 2021. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11845–11854.

Yang, X.; Dong, Y.; Pang, T.; Su, H.; Zhu, J.; Chen, Y.; and Xue, H. 2021. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3897–3907.

Yin, B.; Wang, W.; Yao, T.; Guo, J.; Kong, Z.; Ding, S.; Li, J.; and Liu, C. 2021. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Interafional Joint Conference on Arifcial intelligence*, 1252–1258.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhong, Y.; and Deng, W. 2022. OPOM: Customized Invisible Cloak towards Face Privacy Protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, F.; Zhou, Q.; Yin, B.; Zheng, H.; Lu, X.; Ma, L.; and Ling, H. 2024. Rethinking Impersonation and Dodging Attacks on Face Recognition Systems. In *ACM Multimedia 2024*.

Zhu, Z.-A.; Lu, Y.-Z.; and Chiang, C.-K. 2019. Generating adversarial examples by makeup attacks on face recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, 2516–2520. IEEE.