arXiv:2305.13548v1 [cs.CV] 22 May 2023

# Attribute-Guided Encryption with Facial Texture Masking

Chun Pong Lau
clau13@jhu.edu
Jiang Liu
jiangliu@jhu.edu
Rama Chellappa
rchella4@jhu.edu

Johns Hopkins University
Baltimore, USA

**Abstract**

The increasingly pervasive facial recognition (FR) systems raise serious concerns about personal privacy, especially for billions of users who have publicly shared their photos on social media. Several attempts have been made to protect individuals from unauthorized FR systems utilizing adversarial attacks to generate encrypted face images to protect users from being identified by FR systems. However, existing methods suffer from poor visual quality or low attack success rates, which limit their usability in practice. In this paper, we propose Attribute Guided Encryption with Facial Texture Masking (AGE-FTM) that performs a *dual manifold* adversarial attack on FR systems to achieve both good visual quality and high black-box attack success rates. In particular, AGE-FTM utilizes a high fidelity generative adversarial network (GAN) to generate natural *on-manifold* adversarial samples by modifying facial attributes, and performs the facial texture masking attack to generate imperceptible *off-manifold* adversarial samples. Extensive experiments on the CelebA-HQ dataset demonstrate that our proposed method produces more natural-looking encrypted images than state-of-the-art methods while achieving competitive attack performance. We further evaluate the effectiveness of AGE-FTM in the real world using a commercial FR API and validate its usefulness in practice through an user study.

## 1 Introduction

The rise of deep neural networks has enabled the tremendous success of facial recognition (FR) systems [3, 6, 24]. However, the widely deployed FR systems also pose significant threat to personal privacy as billions of users have publicly shared their photos on social media. By large-scale social media photo analysis, FR systems can be used for detecting user relationships [28], stalking victims [29], stealing identities [18], and performing massive government surveillance [8, 20, 22]. It is critical to develop facial privacy protection techniques to protect individuals from unauthorized FR systems.

Recently, several works [2, 10] used adversarial attacks to generate encrypted face images to protect users from being identified by FR systems. However, existing attacks on FR systems [2, 10, 14, 26, 35, 36, 38] suffer from poor visual quality or low attack success rates. Noise-based methods such as Lowkey [2] and TIP-IM [35] create unpleasant noise

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.




(a) Original (b) TIP-IM [35] (c) AMT-GAN [10] (d) AGE-FTM (ours)

Figure 1: Comparison with other black-box adversarial attacks on FR systems. The number on the left bottom corner is the face verification confidence score with a target image of other person from a commercial FR API Face++.

patterns on face images(see Fig. 1(b)). Patch attack-based methods such as Adv-Hat [14] and Adv-Glasses [26] add unexplaniable and conspicuous changes to the source images. Makeup-based methods [10, 36] drastically change the makeup styles of source images and might be biased towards female users [10] (see Fig. 1(c)). From a user-centered perspective, the aforementioned approaches are not desirable in practice, as everyone wants to post their best-looking photos on social media – not unattractive photos, even though they might be encrypted for protecting personal privacy.

An ideal facial identity encryption algorithm should only create *natural* or *imperceptible* changes to the source images, while being able to fool the FR systems. To achieve this goal, we propose Attribute Guided Encryption with Facial Texture Masking (AGE-FTM) that performs dual manifold adversarial attacks on FR systems. We utilize a high fidelity generative adversarial network (GAN) to learn the natural face manifold, and propose the Attributed Guided Encryption (AGE) strategy that uses the attribute vectors to guide the *on-manifold* attacks, which ensures the encrypted images are *natural-looking*. Unlike makeup-based methods that induce severe artifacts on male faces image, changing attribute vectors such as "age" and "smile" does not have significant gender bias and hence both the protected male and female images look naturally. In addition, AGE-FTM performs Facial Texture Masking (FTM) Attack that generates *off-manifold* adversarial samples by only perturbing the texture regions of the images. Human perception systems are less sensitive to changes in high-frequency components compared to low-frequency components [32], which ensures the perturbations generated by FTM to be *imperceptible*. The overview of AGE-FTM is shown in Fig. 2.

In summary, the contributions of this paper are as follows:

- We propose AGE-FTM, which performs dual manifold adversarial attacks on FR systems. The off-manifold attack FTM only perturbs pixels in the hair texture region. On the other hand, the on-manifold attack AGE applies attribute vectors to guide the latent space attack. These two components ensure AGE-FTM to create natural and imperceptible changes for facial identity protection.
- Our extensive experiments on the CelebA-HQ dataset demonstrate that AGE-FTM produces more natural looking encrypted images than state-of-the-art methods while achieving competitive attack performance.
- We further evaluate the effectiveness of AGE-FTM in the real world using a commercial FR API Face++[1] and validate its usefulness in practice through a user study.

[1] https://www.faceplusplus.com/

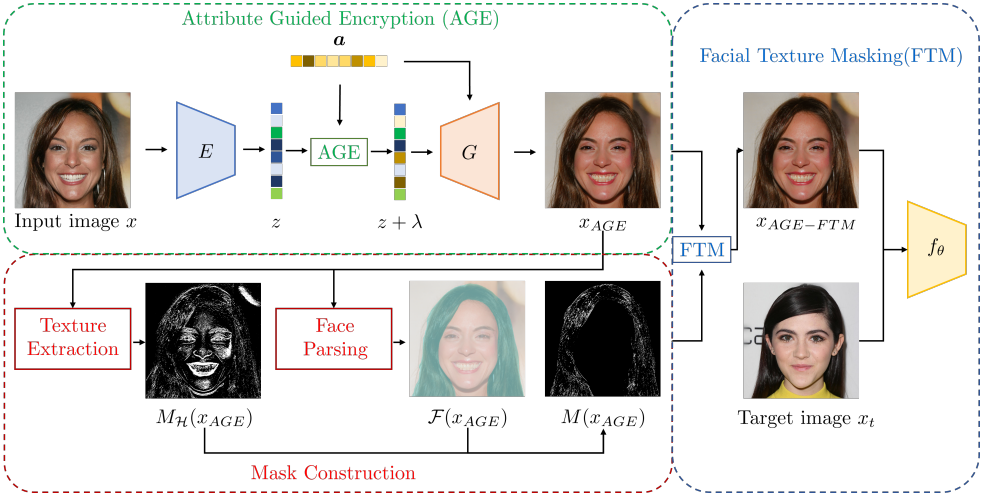



Figure 2: Overview of AGE-FTM. The face image $x$ is passed through the AGE, which is an on-manifold attack. The corresponding latent code $z$, which is encoded by $E$, is perturbed within the latent space and guided by the attribute vector. Then the intermediate image $x_{AGE}$, is passed through the texture extraction module and the face parsing module to obtain the texture mask and face parsing map respectively. These two masks will be combined as a hair texture mask $M$. Then $x_{AGE}$ will be passed through FTM, which is a off-manifold attack restricted on the mask $M$, to obtain the resultant protected image $x_{AGE-FTM}$

# 2 Related Work

## 2.1 Adversarial Attacks on Face Recognition

Many studies have been proposed to attack FR systems, including both poisoning [25] and evasion [2, 10, 14, 26, 35, 36] attacks. Poisoning attacks require injecting poisoned face images into the training sets of FR systems, which is unlikely to achieve for individual users. Evasion attacks, especially transferable black-box attacks, are more practical for protecting facial image privacy, as they only require perturbing the source images to fool the FR systems. Existing attacks on FR systems [2, 10, 14, 26, 35, 36] suffer from poor visual quality or low attack success rates, which limit their usability in the real world. Noise-based methods such as Lowkey [2] and TIP-IM [35] create unexplainable noise patterns on face images. Patch attack-based methods such as Adv-Hat [14] and Adv-Glasses [26] add unnatural and conspicuous changes to the source images. Makeup based-methods [10, 36] drastically change the makeup styles of source images and might be biased towards female users [10]. In contrast, the proposed AGE-FTM creates natural and imperceptible changes to the source images and achieves competitive attack performance.

## 2.2 Off- and On-Manifold Attacks

Recently, several works aim to use a low-dimensional underlying data manifold to attack neural networks by creating on-manifold adversarial samples [11, 17, 30, 31]. On-manifold adversarial samples are adversarial samples constrained to lie on data manifolds and are obtained by perturbing inputs in the latent space learned by generative models. Adversarial samples computed in the image space are considered as off-manifold [31]. On-manifold

adversarial samples have been used to break models trained by adversarial training [30] as well as deep generative model-based defense methods [1] such as DefenseGAN [21], Analysis by Synthetics [23] and MoG-VAE [4]. Compared to off-manifold attacks, on-manifold attacks generate more natural-looking and less conspicuous adversarial samples as they come from a learned natural image manifold. [15] proposed Joint Space Threat Model, which is a dual-manifold attack that combines both on-manifold and off-manifold perturbations.

# 3 Attribute Guided Encryption with Facial Texture Masking (AGE-FTM)

## 3.1 Problem Formulation

In this section, we formulate the problem of adversarial attacks on FR systems. Suppose the face images $x \in \mathcal{X} := \mathbb{R}^{H\times W\times C}$ are drawn from an underlying distribution $\mathbb{P}_X$, where $H$, $W$ and $C$ are the height, width and the number of channels of the image respectively. Let $f_\theta$ be a parameterized model which maps any image in $\mathcal{X}$ to a feature vector $y$ in $f_\theta(\mathcal{X})$. An *accurate* FR system can map two images $x_1$ and $x_2$ with the same identity to features that are very close in the feature space, i.e. $\mathcal{D}(f_\theta(x_1), f_\theta(x_2)) \leq \tau$, where $\mathcal{D}$ is a distance function and $\tau$ is a threshold. A successful attack fools the FR system to map an adversarial image $\hat{x}$ with the same identity as $x$ to a feature that is far away from the feature of $x$, i.e. $\mathcal{D}(f_\theta(\hat{x}), f_\theta(x)) \geq \tau$. Another type of successful attack is the targeted attack, which is also called the impersonation attack in FR. It aims to perturb the source image such that it is identified as the target identity, i.e. $\mathcal{D}(f_\theta(\hat{x}), f_\theta(x_t)) \leq \tau$ where $x_t$ is the target image. From [17], *exact manifold assumption (EMA)* is that there exists a generative model $G$ such that there exists a latent representation $z$ in the latent space $\mathcal{Z} := \mathbb{R}^d$ for every image $x$.

**Off- and On-Manifold Robustness** To encrypt the facial information against malicious FR models in a practical scenario, we consider the impersonation attack in a black-box setting. We use multiple FR models to craft the attack and denote the set of FR systems as $\mathcal{A}_f$. Formally, we aim to solve the following problem:

$$\min_{\hat{x}} \mathcal{D}(f_\theta(\hat{x}), f_\theta(x_t)) = \sum_{f_\theta \in \mathcal{A}_f} 1 - \cos\left(f_\theta(\hat{x}), f_\theta(x_t)\right), \tag{1}$$

where $x_t$ is the target image and $\mathcal{D}$ is the cosine similarity loss. We consider the adversarial samples perturbed within the image space and the latent space as off-manifold and on-manifold adversarial samples respectively. Mathematically,

$$\min_{\delta\in\Delta} \mathcal{D}(f_\theta(x+\delta), f_\theta(x_t)) \text{ and } \min_{\lambda\in\Lambda} \mathcal{D}(f_\theta(G(z+\lambda)), f_\theta(x_t)), \tag{2}$$

where $x+\delta$ and $G(z+\lambda)$ are the off-manifold and on-manifold adversarial samples respectively, $\Delta = \{\delta : \|\delta\|_p < \varepsilon\}$ and $\Lambda = \{\lambda : \|\lambda\|_p < \eta\}$.

## 3.2 Facial Texture Masking Attack

Off-manifold attacks in the image space such as PGD [19] are usually strong and imperceptible. However, when it comes to FR, in particular for black-box attacks, it becomes less effective. Moreover, it is easier to perceive the existence of the attack when it comes to face images than in other images. According to [32], human perception systems are less sensitive to changes in high-frequency components compared to low-frequency components. In other words, edges or textures of the face image are less disturbing to the human perception systems. On the

other hand, faces of people tend to be smooth piecewise regions. It is more noticeable when there is a perturbation in those regions. This motivates us to propose Facial Texture Masking (FTM) Attack, which consists of two components.

The first one is that we do not attack the whole image but only the textured regions in the image. This makes the attack hard to notice from the perspective of the human visual system. There are many methods to extract the textured regions in the images, such as edge detection algorithms or truncation in the frequency domain using Fourier transform or Wavelet transform. To reduce computational cost, we are intentionally not using a deep learning approach to extract textures. Instead, we apply *unsharp masking* technique to extract the textures. Suppose $\mathcal{B}$ is a Gaussian blur operator. We can obtain the binary mask $M_{\mathcal{H}}$ by

$$M_{\mathcal{H}} = \mathbb{1}\{\mathcal{H}(x) > \gamma\},\ \mathcal{H}(x) = |x - \mathcal{B}(x)| \tag{3}$$

where $\mathbb{1}$ is the characteristic function, $\mathcal{H}$ is the high frequency component that generally consists of edges, and $\gamma$ is some predefined threshold. Then we can restrict the off-manifold perturbation within the high frequency components. Taking the PGD attack as an example, we only perturb the pixels that are in the high frequency regions, i.e. $\mathcal{A}_{\mathcal{H}} = \{x_{ij} : M_{\mathcal{H}}(x_{ij}) = 1\}$. The iteration steps become:

$$\delta_{k+1} = \varepsilon_{iter} \cdot sign\left(\nabla_{\delta_k|\mathcal{A}_{\mathcal{H}}} \mathcal{D}(f_\theta(x+\delta_k), f_\theta(x_t))\right),\ \nabla_{\delta_k|\mathcal{A}_{\mathcal{H}}}(x) = \nabla_{\delta_k}(x) \odot M_{\mathcal{H}}, \tag{4}$$

where $\varepsilon_{iter}$ is the attack step size at each iteration and $\odot$ is the Hadamard product. We denote this attack as Texture Masking Attack (TMA).

Although the TMA significantly reduces the perceptibility of the attack, we can still notice the perturbations at the face portion if we zoom in and look carefully. To further improve the imperceptibility of the attack, we observe that perturbations within the hair region are barely noticable. Therefore, we leverage a face parsing algorithm to locate the hair region. Suppose $\mathcal{F}$ be the face parsing algorithm, which is a function $\mathcal{F} : \mathcal{X} \to \mathcal{C}$ that maps each pixel $x_{ij}$ in a face image $x$ in $\mathcal{X}$ to a discrete component label $c$ in $\mathcal{C} := \{1, \cdots, |\mathcal{C}|\}$. Then the new binary mask $M$ is

$$M = \mathbb{1}\left\{\{\mathcal{H}(x) > \gamma\} \cap \{x_{ij} : \mathcal{F}(x_{ij}) = c_{\text{hair}}\}\right\}, \tag{5}$$

where $\{x_{ij} : \mathcal{F}(x_{ij}) = c_{\text{hair}}\}$ is the set containing the pixels in the hair portion. Denote $\mathcal{A} = \{x_{ij} : M(x_{ij}) = 1\}$. The formulation of the FTM attack becomes

$$\min_{\delta \in \Delta} \mathcal{D}(f_\theta(x + \delta \odot M), f_\theta(x_t)), \tag{6}$$

and the iteration steps in the FTM attack will be

$$\delta_{k+1} = \varepsilon_{iter} \cdot sign\left(\nabla_{\delta_k|\mathcal{A}} \mathcal{D}(f_\theta(x+\delta_k), f_\theta(x_t))\right). \tag{7}$$

## 3.3 Attributed Guided Encryption

Makeup attacks such as [10, 36] are shown to be effective for FR systems. These attacks can synthesize makeup in some parts of the face, such as the eye region in [36] or the whole face in [10]. However, there are some drawbacks. Makeup attacks usually require a makeup dataset for training. This leads to better visual performance for female because of the imbalance of gender in the makeup dataset. Moreover, in order to increase the attack strength, makeup attacks usually produce heavy makeup on the original images which may be undesirable. Therefore, we propose the Attributed Guided Encryption (AGE), which uses the attribute vectors in the generator to guide the on-manifold attack. We apply HFGI [33], which is a state-of-the-art method that uses a high fidelity GAN and can invert and edit high resolution

face images effectively. HFGI consists of an encoder $E$ to encode the images and uses a generator $G$ to edit the high resolution face images. Then the AGE formulation is

$$\min_{\lambda \in \Lambda} \mathcal{D}(f_\theta(G(E(x)+\lambda;\boldsymbol{a})), f_\theta(x_t)), \tag{8}$$

To obtain the perturbation, we optimize it by

$$\lambda_{k+1} = \eta_{iter} \cdot sign\left(\nabla_{\lambda_k} \mathcal{D}(f_\theta(G(E(x)+\lambda_k;\boldsymbol{a}_k), f_\theta(x_t)))\right), \tag{9}$$

where $\boldsymbol{a}_k = \frac{k\boldsymbol{a}}{N}$, $\eta_{iter}$ is the attack step size at each iteration, and $N$ is the total iteration number. For each step, intuitively the latent vector $z$, which represents the general latent representation, is perturbed by $\lambda_k$ and guided by $\boldsymbol{a}_k$, the specific semantic latent direction, such as attributes "age" and "smile". Therefore, combining FTM and AGE, we propose the AGE-FTM, which is a dual manifold attack consisting of both off- and on-manifold attacks,

**AGE-FTM optimization:**

$$\min_{\lambda \in \Lambda,\ \delta \in \Delta} \mathcal{D}(f_\theta(G(E(x)+\lambda;\boldsymbol{a}) + \delta \odot M, f_\theta(x_t)). \tag{10}$$

# 4 Experiments

## 4.1 Settings

**Datasets** To generate high quality face images , the FFHQ dataset [13] is used for training following [33]. To train the face parsing algorithm, the CelebAMask-HQ dataset [16] is used. CelebAMask-HQ is a large-scale face image dataset that has 30000 high-resolution face images, whose masks are manually-annotated including all facial components and accessories such as skin, nose and eyes. For testing, we choose CelebA-HQ dataset [12] and FFHQ dataset [13], which are popular face image datasets with high resolution. For CelebA-HQ, we use a subset of 1000 face images with different identities following [10]. For FFHQ, we randomly select 1000 images.

**Evaluation details** We compare the performance of the proposed method with different adversarial attacks specifically for FR: 1) Perturbation-based attack TIP-IM [35]; 2) Makeup-based attack Adv-Makeup [36]; and 3) Style transfer attack AMT-GAN [10]. To evaluate the attack ability of different impersonation attacks, we follow [34] to use attack success rate (ASR). We show the ASR at FAR@0.01 in a black-box setting. To evaluate the image quality of adversarial images, we use FID [7], which measures the distance of the data distribution.

**Implementation details** For texture extraction, the blurring operator $\mathcal{B}$ is a Gaussian smoothing operator with kernal size $19 \times 19$ and the standard deviation is 5. The threshold for texture extraction $\gamma$ is 0.003. For face parsing, we used a pretrained BiSeNet [32] trained on the CelebAMask-HQ dataset [16]. For the generator $G$, we used a pre-trained HFGI [33] trained on the FFHQ dataset [13]. To obtain the attribute vectors, we apply InterfaceGAN [27] to obtain the attribute vectors to edit the generated images.

We set $\varepsilon = 16/255$, $\varepsilon_{iter} = 2/255$ with fifty iteration steps for the off-manifold attack FTM. We set $\eta = 0.1$, $\eta_{iter} = 0.02$ with $N = 10$ iteration steps for the on-manifold attack AGE. The pretrained FR models that we use for black box attacks includes IR152 [5], IRSE50 [9], Facenet [24], and Mobileface [3]. All the experiments are conducted on one NVIDIA RTX 2080Ti GPU.

| | Black-box Attack Ability | | | | Image Quality | Black-box Attack Ability | | | | Image Quality |
|---|---|---|---|---|---|---|---|---|---|---|
| | IRSE50 | IR152 | Facenet | Mobileface | FID (↓) | IRSE50 | IR152 | Facenet | Mobileface | FID (↓) |
| Clean | 7.29 | 3.80 | 1.08 | 12.68 | - | 4.4 | 2.5 | 1.7 | 5.2 | - |
| TIP-IM [35] | 54.4 | 37.2 | **40.7** | 48.7 | 38.74 | **52.3** | 27.5 | **12.50** | **55.1** | 39.86 |
| AMT-GAN [40] | **77.0** | 35.1 | 16.6 | 50.7 | 34.44 | 19.9 | 16.1 | 3.7 | 30.4 | 35.23 |
| AGE-TMA (ours) | 69.1 | **51.0** | 13.2 | **78.0** | **27.61** | 47.7 | **36.6** | 11.2 | 53.9 | 32.04 |
| AGE-FTM (ours) | 51.0 | 34.7 | 13.1 | 63.6 | 27.63 | 28.0 | 27.4 | 11.0 | 34.0 | **31.51** |

Table 1: Quantitative evaluation of the proposed method and other compared methods. Left: CelebA-HQ. Right: FFHQ.




(a) Original (b) TIP-IM [35] (c) AMT-GAN [40] (d) AGE-FTM (ours)

Figure 3: Visualization of the proposed method AGE-FTM attack with other compared methods. Top: CelebA. Bottom: FFHQ.

## 4.2 Main Results

**Black-box attack ability** The black-box attack results on four different pre-trained FR models are shown in Table 1. For each target model, the other three FR models are used to craft the attacks. From Table 1, we observe both AGE-TMA and AGE-FTM have a strong black-box attack ability. AGE-TMA has the strongest attack ability among all the methods in the CelebAMask-HQ dataset and is comparable TIP-IM in the FFHQ dataset. In both datasets, the proposed method is significantly better than AMT-GAN.

**Image quality** The image quality results evaluated with FID are shown in Table 1. To measure the naturalness of the attacked images, we also show the FID results. Note that the proposed methods have the best FID results. This demonstrates that the adversarial images generated by AGE-FTM have a more natural appearance, which is the goal of this paper. The visualization is in Fig. 3. We can observe similar results from FTM in Table 2. Since FTM only attacks the textures on the hair region, it is a local attack. All the pixels which are not on the hair region are not perturbed. Therefore, it achieves good image quality results.

| Method | Clean | TMA | FTM | AGE | AGE-TMA | AGE-FTM |
|---|---|---|---|---|---|---|
| FID (↓) | - | 1.33 | 0.19 | 27.64 | 27.61 | 27.63 |
| ASR (↑) | 12.68 | 38.9 | 18.1 | 53.0 | 78.0 | 63.6 |

Table 2: Ablation studies of AGE-FTM. PSNR, SSIM, FID and ASR are evaluated with CelebA-HQ datase. ASR at FAR@0.01 is evaluated in a black-box setting with the targeted model Mobileface.



(a) Original (b) TMA (c) TMA Mask (d) FTM (e) FTM mask

Figure 4: Visualization of the proposed FTM attack. For TMA, it is harder to notice as we only perturb the texture of the image. To further improve the imperceptibility, FTM only perturbs the texture of the hair portion of the face image. This makes it even harder to distinguish whether it is attacked or not.

## 4.3 Ablation Studies

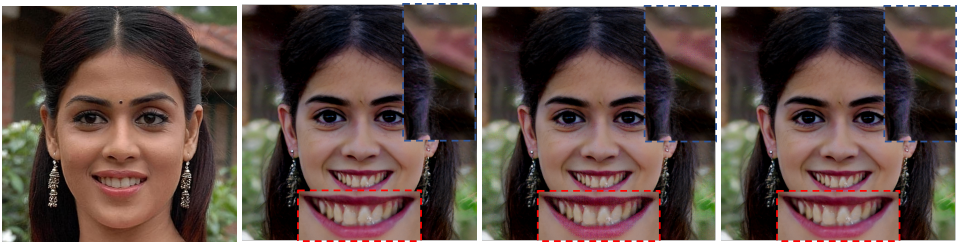
We show the ablation studies evaluating the image quality and the attack ability of the proposed method in Table 2. The ASR is evaluated at FAR@0.01 with Mobileface as the targeted model. We trained five models for evaluating the performance of the proposed methods AGE-FTM.

**FTM** To compare the facial masking component, we have 1) the TMA that only attacks the textures of the image; and 2) the FTM attack that only attacks the textures of the hair portion. The visualization is shown in Fig. 4. From Table 2, we can observe FTM achieves the best image quality. On the other hand, we can see there is a trade-off between the image quality and the attack ability. This is reasonable as the total number of perturbed pixels is reduced significantly for the case of TMA and FTM (See Fig. 4(d) and (f)). In particular for FTM, since the attack is restricted to the hair portion, the attack is weakened.

**AGE** To compare the AGE component, we have 1) AGE which uses the attribute vector to guide the on-manifold attack; 2) AGE-TMA that is a dual manifold attack containing AGE and TMA; and 3) AGE-FTM that is our ultimate dual manifold attack containing AGE and FTM. The visualization is shown in Fig. 5. From Table 2, we can see that AGE itself achieves descent ASR. Similar to makeup attacks, AGE also changes the style/attribute of the images. However, the proposed AGE is different from other makeup attacks as the attribute vectors, which codes the semantic latent vector directions, are used to regularize the on-manifold perturbation. The noise added in TMA and FTM is hardly noticeable (See Fig. 5(c) and (d)).

## 4.4 Attack Ability on Commercial FR API

We evaluate the attack performance by collecting and averaging the confidence score from the commercial FR API Face++ with the CelebA-HQ dataset. The average confidence score for the clean images, TIP-IM [65], AMT-GAN [10], and the proposed method AGE-FTM are



(a) Original (b) AGE (c) AGE-TMA (d) AGE-FTM

Figure 5: Visualization of the proposed FTM attack. For PGD attack, all the pixels are perturbed and it is easy to notice the perturbation. For TMA, it is harder to notice as we only perturb the texture of the image. To further improve the imperceptibility, FTM only perturbs the texture of the hair portion of the face image. This makes it even harder to distinguish whether it is attacked or not.

33.67%, 57.06%, 54.17%, and 54.96% respectively. This shows that the proposed method has a satisfactory attack ability even for the commercial API while achieving good image quality.

### 4.5 User Study

We conducted a user study to evaluate the subjective image quality of our proposed method and baselines. The participants include forty five social media users. We showed the participants encrypted images of ten subjects from the CelebA-HQ dataset generated by TIP-IM [35], AMT-GAN [10], and AGE-FTM (ours), and the participants were asked to choose up to three images for each person that they will like to use for sharing on social media if they were the person. On average, 68.89% of the participants chose the image generated by AGE-FTM, 33.56% chose AMT-GAN, and 6.22% chose TIP-IM. The results demonstrate that our proposed method has higher image quality than the baselines and is more likely to be used in practice.

### 4.6 Limitations and Future direction

AGE-FTM involves two components, AGE and FTM. AGE perturbs the latent space guided with attribute vectors to perform an on-manifold attack. However, it requires a generator that has a semantic latent space and generates a high-quality face image. A good face alignment is necessary for these face image GAN models. In other words, AGE does not work without face alignment. This would be a future direction that uses a different generator, such as flow-based or diffusion models that do not require face alignment.

## 5 Conclusion

In this paper, we propose AGE-FTM that performs dual manifold adversarial attack on FR systems for facial privacy protection. AGE-FTM utilizes the image manifold information learned by GAN models to create *natural* changes to the source images guided by facial attributes, and adds *imperceptible* perturbations through facial texture masking. Our extensive experiments on the CelebA-HQ dataset demonstrate that AGE-FTM achieves state-of-the-art black-box attack performance and good visual quality, which can be useful for facial privacy protection in real-world applications.

# References

[1] Yanzhi Chen, Renjie Xie, and Zhanxing Zhu. On breaking deep generative model-based defenses and beyond. In *International Conference on Machine Learning*, 2020.

[2] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[4] Partha Ghosh, Arpan Losalka, and Michael J Black. Resisting adversarial attacks using Gaussian mixture variational autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 541–548, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Yonghao He, Dezhong Xu, Lifang Wu, Meng Jian, Shiming Xiang, and Chunhong Pan. LFFD: A light and fast face detector for edge devices. *arXiv preprint arXiv:1904.10633*, 2019.

[7] 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. *Advances in neural information processing systems*, 30, 2017.

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

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[10] 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. *arXiv preprint arXiv:2203.03121*, 2022.

[11] Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

[12] 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.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[14] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021.

[15] Chun Pong Lau, Jiang Liu, Hossein Souri, Wei-An Lin, Soheil Feizi, and Rama Chellappa. Interpolated joint space adversarial training for robust and generalizable defenses. *arXiv preprint*, 2021.

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks. In *Advances in Neural Information Processing Systems*, 2020.

[18] Taylor Kay Lively. Facial recognition in the US: Privacy concerns and legal developments. https://www.asisonline.org/security-management-magazine/monthly-issues/security-technology/archive/2021/december/facial-recognition-in-the-us-privacy-concerns-and-legal-deve Accessed: 5/12/2021.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[20] Paul Mozur and Aaron Krolik. A surveillance net blankets china's cities, giving police vast powers. *The New York Times*, 17, 2019.

[21] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[22] ADAM SATARIANO. Police use of facial recognition is accepted by british court. *The New York Times*, 4, 2019.

[23] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. In *International Conference on Learning Representations*, 2018.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[25] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1589–1604, 2020.

[26] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019.

[27] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

[28] Yan Shoshitaishvili, Christopher Kruegel, and Giovanni Vigna. Portrait of a Privacy Invasion. *Proceedings on Privacy Enhancing Technologies*, 2015(1):41–60, 2015. doi: 10.1515/popets-2015-0004.

[29] Maya Shwayder. Clearview ai's facial-recognition app is a nightmare for stalking victims. *Digital Trends*, 2020.

[30] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[31] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] Ee-Leng Tan and Woon-Seng Gan. *Perceptual Image Coding with Discrete Cosine Transform*. Springer Publishing Company, Incorporated, 2015. ISBN 9789812875426.

[33] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[34] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11845–11854, 2021.

[35] 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 IEEE/CVF International Conference on Computer Vision*, pages 3897–3907, 2021.

[36] 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.

[37] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[38] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16: 1452–1466, 2020.