Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

7-2022

A3GAN: Attribute-aware anonymization networks for face de-identification

Liming ZHAI

Qing GUO

Xiaofei XIE
Singapore Management University, xfxie@smu.edu.sg

Lei MA

Yi Estelle WANG

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Artificial Intelligence and Robotics Commons, and the OS and Networks Commons

Citation

ZHAI, Liming; GUO, Qing; XIE, Xiaofei; MA, Lei; WANG, Yi Estelle; and LIU, Yang. A3GAN: Attribute-aware anonymization networks for face de-identification. (2022). *Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 2022 October 10-14*. 5303-5313.
Available at: https://ink.library.smu.edu.sg/sis_research/7495

This Conference Proceeding Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Liming ZHAI, Qing GUO, Xiaofei XIE, Lei MA, Yi Estelle WANG, and Yang LIU

This conference proceeding article is available at Institutional Knowledge at Singapore Management University: https://ink.library.smu.edu.sg/sis_research/7495

A^{3}GAN: Attribute-Aware Anonymization Networks for Face De-identification

Liming Zhai
Nanyang Technological University
Singapore
limingzhai@whu.edu.cn

Qing Guo*
Nanyang Technological University
Singapore
tsingqguo@ieee.org

Xiaofei Xie
Singapore Management University
Singapore
xfxie@smu.edu.sg

Lei Ma
University of Alberta & Kyushu University
Canada
ma.lei@acm.org

Yi Estelle Wang
Continental Automotive Singapore Pte. Ltd.
Singapore
Estelle.wang@continental.com

Yang Liu*
Zhejiang Sci-Tech University & Nanyang Technological University
China
yangliu@ntu.edu.sg

ABSTRACT

Face de-identification (De-ID) removes face identity information in face images to avoid personal privacy leakage. Existing face De-ID breaks the raw identity by cutting out the face regions and recovering the corrupted regions via deep generators, which inevitably affect the generation quality and cannot control generation results according to subsequent intelligent tasks (*e.g.*, facial expression recognition). In this work, for the first attempt, we think the face De-ID from the perspective of attribute editing and propose an *attribute-aware anonymization network (A^{3}GAN)* by formulating face De-ID as a joint task of *semantic suppression and controllable attribute injection*. Intuitively, the semantic suppression removes the identity-sensitive information in embeddings while the controllable attribute injection automatically edits the raw face along the attributes that benefit De-ID. To this end, we first design a multi-scale semantic suppression network with a novel *suppressive convolution unit (SCU)*, which can remove the face identity along multi-level deep features progressively. Then, we propose an attribute-aware injective network (AINet) that can generate De-ID-sensitive attributes in a controllable way (*i.e.*, specifying which attributes can be changed and which cannot) and inject them into the latent code of the raw face. Moreover, to enable effective training, we design a new anonymization loss to let the injected attributes shift far away from the original ones. We perform comprehensive experiments on four datasets covering four different intelligent tasks including face verification, face detection, facial expression recognition, and fatigue detection, all of which demonstrate the superiority of our face De-ID over state-of-the-art methods.

*Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '22, October 10–14, 2022, Lisboa, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547757

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Face de-identification; Facial attribute; Controllability

ACM Reference Format:
Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. 2022. A^{3}GAN: Attribute-Aware Anonymization Networks for Face De-identification. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3503161.3547757

1 INTRODUCTION

Privacy leakage has become a major concern with the explosive growth of social media usage in today's digital era [29]. Many notable privacy protection laws such as General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Personal Data Protection Act (PDPA) and Personal Information Protection Act (PIPA), have been enacted by governments in the last few years [35]. The face information is of paramount importance to personal privacy [5], and is typically protected by face de-identification (De-ID) [23], which aims to remove identifying characteristics from original face images, evading both human and machine recognizers.

Although advances have been achieved, the existing face De-IDs are still far from satisfactory. Early face De-ID methods only focus on the *anonymity* (the anonymized face is unidentifable), and are usually accomplished by some naïve image processing operations, such as pixelation [27], blurring [28] and black-out [32], which are visually unpleasant and also hinder downstream vision tasks (*e.g.* face detection and facial attribute analysis).

Recent face De-ID methods first cut out the face regions and then replace the original faces with new generated faces [7, 12, 22, 37, 38], focusing more on the *realism* (the anonymized face should be photo-realistic). However, generating new and natural faces to seamlessly fit the original image scenes is challenging, and visual artifacts or distortions are often introduced to the anonymized faces (see the result of CIAGAN [22] in Fig. 1). Besides, these face De-ID methods also neglect the fact that the anonymized faces may be used to the downstream applications and their inherent flaw (*i.e.*, they cannot flexibly control the semantic information of anonymized faces) makes the generated faces impossible to support such applications.

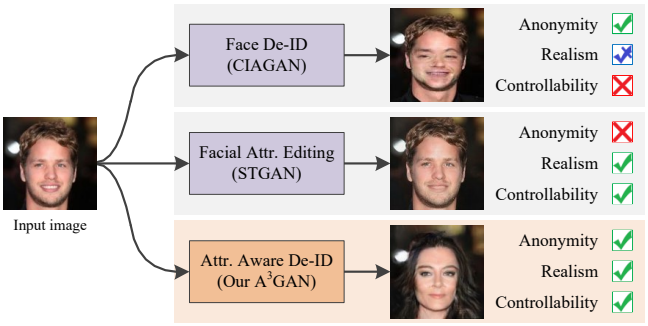

Figure 1: Comparison of conventional face De-ID (CIAGAN [22]), facial attribute editing (STGAN [20]) tasks and our face De-ID A^{3}GAN.

For example, these face De-ID methods likely fail to anonymize the attendees while keeping their facial expressions unchanged in an online meeting. Or they unsatisfactorily remove the face identity while preserving the gaze direction and the mouth state, which are essential for fatigue detection in a driving scenario. These real-world applications require stricter control over the anonymized faces, indicating the importance of *controllability* for face De-ID. Although some works [7, 22] introduce the control information to the face De-ID process, their purpose is just to improve the diversity of anonymized faces, which still cannot satisfy the requirements of the above downstream tasks. As shown in Fig. 1, the facial expression and facial attributes in the anonymized face of the state-of-the-art CIAGAN [22] is changed unpredictably.

To address the above problems, we propose to use facial attribute manipulation to anonymize the faces. The motivation is that the face identity is a high-level semantic representation determined by the combination of specific facial attributes, and thus the change of face identity can be fulfilled through manipulating the facial attributes, which is conducive to maintaining the image quality and also allows more granular control for anonymization. However, existing facial attribute editing methods [47] aim to modify the facial attributes while preserving the original face identity, which is utterly different from the objective of face De-ID (see the result of STGAN [20] in Fig. 1). Besides, one might seemingly rightfully argue that the face De-ID is easy to reach by simply accumulating multiple changed facial attributes. Nevertheless, as evidenced by our experiment (see Fig. 2 in Sec. 3.2), the changes of multiple facial attributes may be a conflict with each other and easily cause unreasonable face semantics or degraded image quality. As a result, there is an urgent need for new techniques to leverage the facial attributes for face De-ID.

To this end, we devise an *attribute-aware anonymization generative adversarial network (A^{3}GAN)* for face De-ID. We divide the face De-ID into two tasks: original identity suppression and controllable attribute injection, and design a novel encoder-decoder network with two new modules, *i.e.*, *suppressive convolutional unit (SCU)* and *attribute-aware injective network (AINet)*, to handle the two tasks. In particular, the SCU is able to reduce the identity information within the deep embeddings of the encoder and the AINet is to inject alternative controllable face attributes into the latent code at bottleneck. The encoder embeddings and updated latent code are fed into the decoder to produce realistic and high-quality faces. Moreover, to allow effective training, we design a new anonymization loss using an attribute mask vector to let the injected attributes shift far away from the original ones selectively. As a result, the proposed method simultaneously satisfies the three face De-ID requirements, i.e., anonymity, realism, and controllability (see Fig. 1). We conduct a large-scale evaluation of A^{3}GAN. Apart from the face verification and face detection that emphasizes anonymity and realism, we also use facial expression recognition and fatigue detection to heavily test the controllability of our proposed method. All experiments demonstrate its advantages over existing face De-ID methods.

2 RELATED WORK

Face de-identification. In recent years, the generative adversarial network (GAN) has dominated the study of face De-ID. Many GAN-based methods regard the face De-ID as an image inpainting problem [6], in which the face region is firstly blacked out, and then the face-missing image is seamlessly filled with a new face generated by GAN with the help of face landmarks. The blacking out face is to reduce the original identity, and the face landmarks are used for pose preservation. Under this paradigm, Sun *et al.* [37, 38] propose the two-stage face De-ID frameworks, which first generate a new face identity based on landmarks or a parametric face model, and then blend the rendered head into the background image using a GAN. Hukkelas *et al.* [12] use a conditional GAN for face De-ID, and considers the pose information to ensure high-quality faces. Maximov *et al.* [22] also adopt a conditional GAN, where a different identity label is used as a condition to control the face identity, and additionally explore an identity guidance discriminator to further increase the anonymity. These methods cannot control the semantic details in anonymized faces, thus limiting their adaptation to downstream tasks. In this paper, we go beyond this paradigm, and anonymize the faces more granularly by manipulating facial attributes instead of total face generation.

There are also some face De-ID methods related to facial attributes. The methods in [14, 19] anonymize the faces while retaining all facial attributes, but they lack controllability and flexibility. In contrast, we perform the face anonymization optionally either by preserving all facial attributes or changing desired facial attributes.

Facial attribute editing. Facial attribute editing aims at altering the semantic attributes in a face image according to target attributes. Most facial attribute editing methods focus on the attribute independence problem, in which the modified facial attributes should not affect non-target attributes or attribute-irrelevant information. Typical solutions include attribute-independent latent representation [31], spatial attention mechanism [17, 45], auxiliary classifier for facial attributes [9], selective transfer unit and difference attribute vector [20], style skip connections [3], disentanglement of latent semantics [36], and latent space factorization model [42], *etc.* All these methods preserve the face identity when modifying the attributes, while our purpose is to change the face identity via facial attribute manipulation.

Another type of facial attribute manipulation is used for privacy protection [2, 24], which adds imperceptible adversarial noises to attribute regions to fool facial attribute classifiers. However, these methods can only mislead machines and do not change the visually visible face identity which is still recognizable to human eyes.

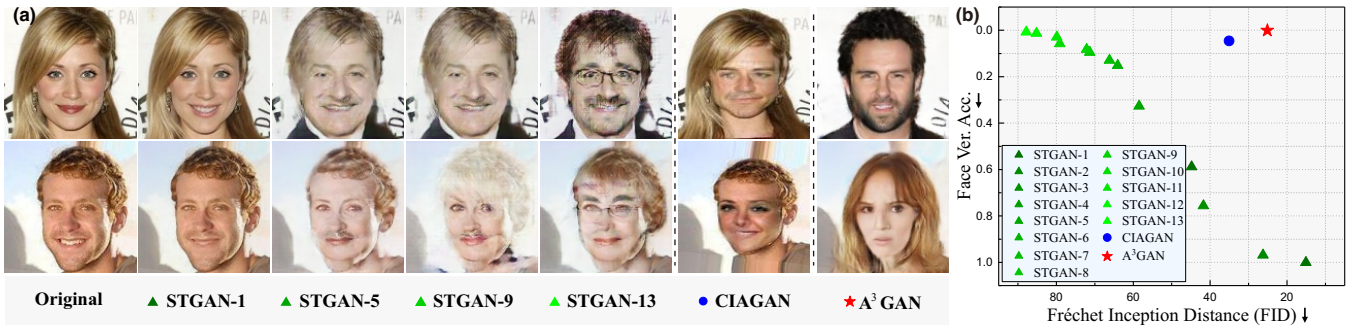

Figure 2: Using the SOTA attribute editing method (*i.e.* STGAN [20]), the SOTA face De-ID method (*i.e.*, CIAGAN [22]), and our method A³GAN to perform face anonymization on the CelebA dataset [21]. STGAN-M is to use STGAN to change M attributes of the input faces and M ranges from 1 to 13. We visualize two examples in (a), and also present the image quality metric (*i.e.*, FID [10]) and the De-ID metric (*i.e.*, face verification accuracy [11]) on the CelebA dataset in (b).

3 MOTIVATION

3.1 Face De-ID as Attribute Editing

Facial attribute editing. Following existing face attribute editing works [3, 20], we can describe an original (source) human face via a series of attributes (*e.g.*, hair color, gender, age) and indicate whether a face contains these attributes through a vector, *e.g.*, $\mathbf{a}_\mathrm{s} \in \{0,1\}^N$ where the i-th element is denoted as $\mathbf{a}_\mathrm{s}^i$ and N is the number of attributes. If we have $\mathbf{a}_\mathrm{s}^i = 1$ for a face, it means that the face contains the i-th attribute and vice versa. Given a target attribute vector $\mathbf{a}_\mathrm{t} \in \{0,1\}^N$, face attribute editing aims to transfer an input face image to a new but realistic face that follows the attributes indicated by the vector $\mathbf{a}_\mathrm{t}$. Existing methods achieve this goal via GANs with specific generators and loss functions. In general, we can formulate the face attribute editing via an encoder-decoder network (*e.g.*, U-Net [33]). Given an input face image $\mathbf{I}$, we first conduct the encoder via

$$\{\mathbf{F}_\mathrm{en}^l\}_{l=1}^L = \phi(\mathbf{I}) = \phi_L(\cdots\phi_2(\phi_1(\mathbf{I}))), \tag{1}$$

where $\phi(\cdot)$ is the encoder with L layers $\{\phi_l(\cdot)\}_{l=1}^L$ and $\mathbf{F}_\mathrm{en}^l$ is the feature extracted from the l-th layer, *i.e.*, $\mathbf{F}_\mathrm{en}^l = \phi_l(\mathbf{F}_\mathrm{en}^{l-1})$. The $\mathbf{F}_\mathrm{en}^L$ is also known as the latent code, which is the output of the last layer of $\phi(\cdot)$ (*i.e.*, $\phi_L(\cdot)$). The decoder takes the target attribute vector $\mathbf{a}_\mathrm{t}$ and $\{\mathbf{F}_\mathrm{en}^l\}_{l=1}^L$ as inputs and generate a new face $\hat{\mathbf{I}}$ via

$$\hat{\mathbf{I}} = \phi^{-1}(\{\mathbf{F}_\mathrm{en}^l\}_{l=1}^L, \mathbf{a}_\mathrm{t}) = \phi_L^{-1}([\cdots\phi_2^{-1}([\phi_1^{-1}([\mathbf{F}_\mathrm{en}^L, \mathbf{a}_\mathrm{t}]), \mathbf{F}_\mathrm{en}^{L-1}])]), \tag{2}$$

where $\phi^{-1}(\cdot)$ denotes the decoder with L layers (*i.e.*, $\{\phi_l^{-1}(\cdot)\}_{l=1}^L$), and '$[\cdot,\cdot]$' denotes a concatenation operation. To generate natural faces meeting the target attributes, some works add new modules in the encoder or loss functions to train a powerful generator [17, 20].

Face De-ID via attribute editing. Intuitively, we can implement face De-ID via the above attribute editing straightforwardly. We take the state-of-the-art (SOTA) attribute editing method (*i.e.*, STGAN [20]) as an example: ❶ Given a face image we want to anonymize, we obtain its source attribute vector $\mathbf{a}_\mathrm{s}$ by using a pre-trained attribute classifier [21]. ❷ We specify a target attribute vector $\mathbf{a}_\mathrm{t}$ by inverting some elements of $\mathbf{a}_\mathrm{s}$. Specifically, we set an index set $\mathcal{T} = \{i | i \in [1, N]\}^M$ containing M indexes and representing M attributes should be inverted, that is, $\mathbf{a}_\mathrm{t}^j = 1 - \mathbf{a}_\mathrm{s}^j, \forall\, j \in \mathcal{T}$. Other elements of $\mathbf{a}_\mathrm{t}$ are the same as those of $\mathbf{a}_\mathrm{s}$. ❸ We use the STGAN and the specified $\mathbf{a}_t$ to map the input face to a new one that is regarded as the anonymized face. With the above process, each set $\mathcal{T}$ corresponds to a face De-ID method. When we randomly select M indexes from $[1, \ldots, N]$ for $\mathcal{T}$, the STGAN-based De-ID method is to change M attributes of the input face and we denote it as STGAN-M. Then, we can test the attribute editing-based De-ID methods with different M as well as the SOTA face De-ID method [22] on the public dataset and discuss their advantages and limitations, as presented in the following subsection.

3.2 Observations and Challenges

We use the above STGAN-based methods with different M (*i.e.*, $\{\text{STGAN-}M | M = [1, \ldots, 13]\}$) and the SOTA De-ID method (*i.e.*, CIACAN [22]) to conduct the face De-ID experiment on the CelebA dataset [21]. Here, we consider total $N = 13$ attributes for $\mathbf{a}_\mathrm{s}$ and $\mathbf{a}_\mathrm{t}$, as did in [20]. We use face verification accuracy [11] and Fréchet inception distance (FID) [10] to evaluate the effectiveness of ID removal and image quality, respectively. We show the results in Fig. 2 and observe that: ❶ Editing a few attributes (*e.g.*, STGAN-$\{1, 2, 3, 4\}$) can generate high-quality faces (*i.e.*, low FID values) but fails to remove the identity information effectively (*i.e.*, high face verification accuracy). ❷ Editing more attributes (*e.g.*, STGAN-$\{5, \ldots, 13\}$) can lead to effective identity changing (*i.e.*, low face verification accuracy) while decreasing the face quality significantly (*i.e.*, high FID values). ❸ Although the De-ID method CIAGAN [22] achieves effective face DeID results, it generates much lower quality faces than STGAN-$\{1, 2\}$.

According to the above observations, we find that existing facial attribute editing methods cannot be used for face De-ID trivially. Only the editing method with elaborated target attribute vectors can generate natural and realistic faces but fails to remove the identity information effectively. In addition, the SOTA face De-ID method can anonymize the face identity effectively but usually introduces undesired artifacts and distortions. These findings and observations motivate us to design a new face De-ID framework based on attribute editing to take advantage of its natural and realistic generation capability. To this end, we need to address two key challenges: how to design the editing generator be able to remove face identity effectively? and how to generate elaborated target attribute information that helps to generate natural and realistic faces automatically?

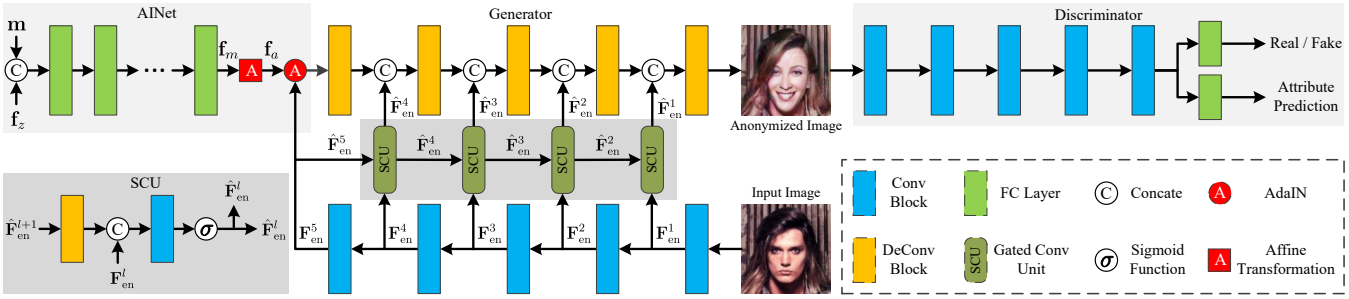

Figure 3: Architecture of the proposed A^{3}GAN. Our main contributions are suppressive convolution units (SCUs) for filtering out original identity information and the attribute-aware injective network (AINet) for injecting new identity information. The architecture details are also shown in Appendix A.

4 METHODOLOGY

4.1 Overview

We propose a *suppresive convolutional unit (SCU)* and an *attribute-aware injective network (AINet)* to handle the two challenges, respectively, which lead to our method A^{3}GAN containing three steps: multi-level feature extraction and original identity suppression, automatic target attribute generation and injection, and suppressed & attribute-injected feature decoding. Specifically, given an input face we want to anonymize, the A^{3}GAN first extracts multi-layer deep features via Eq. (1) and get $\{\mathbf{F}^l_{\text{en}}\}^L_{l=1}$. Instead of feeding these features to the decoder directly, we propose to suppress the identity information in $\{\mathbf{F}^l_{\text{en}}\}^L_{l=1}$ explicitly via a novel module (*i.e.*, SCU). For the l-th feature $\mathbf{F}^l_{\text{en}}$, we can transfer it through the SCU

$$\hat{\mathbf{F}}^l_{\text{en}} = \text{SCU}(\mathbf{F}^l_{\text{en}}, \hat{\mathbf{F}}^{l+1}_{\text{en}}), \tag{3}$$

where SCU takes the l-th feature and the $(l+1)$-th transferred feature as inputs and maps $\mathbf{F}^l_{\text{en}}$ to a new counterpart (*i.e.*, $\hat{\mathbf{F}}^l_{\text{en}}$) in which the identity information is suppressed. Note that, the deeper feature from $(l+1)$-th layer (*i.e.*, $\hat{\mathbf{F}}^{l+1}_{\text{en}}$) captures more semantic information than $\hat{\mathbf{F}}^l_{\text{en}}$ and can guide the identity suppression at the l-th layer. For the last layer, we have $\hat{\mathbf{F}}^L_{\text{en}} = \mathbf{F}^L_{\text{en}}$. We present the connection between SCUs in Fig. 3 and it is a coarse-fine suppression process. We will detail the principle and structure of SCU in Sec. 4.2.

With the SCU, we can obtain identity-suppressed features (*i.e.*, $\{\hat{\mathbf{F}}^l_{\text{en}}\}^L_{l=1}$). Then, following the pipeline of attribute editing, we should feed the features $\{\hat{\mathbf{F}}^l_{\text{en}}\}^L_{l=1}$ and a manually specified attribute vector (*e.g.*, $\mathbf{a}_\text{t}$ in Sec. 3.2) to the decoder and generate a new face. However, the manually specified discrete vector cannot balance the generation quality and face De-ID as discussed in Sec. 3.2. To address this issue, we propose the attribute-aware injective network (AINet) to automatically generate the target attribute information in the latent space instead of using a discrete vector. Specifically, the AINet takes an attribute latent feature $\mathbf{f}_z$ and an attribute mask vector $\mathbf{m}$ as inputs and output another attribute latent feature $\mathbf{f}_\text{a}$. The latent $\mathbf{f}_z$ is randomly sampled from the latent space of attributes. We can represent the above process as

$$\mathbf{f}_\text{a} = \text{AINet}([\mathbf{f}_z, \mathbf{m}]), \tag{4}$$

where $\mathbf{m} = \{0,1\}^{N\times 1}$ determines which attributes could be injected into the latent feature of the input face, that is, the mask vector provides a way to control which attributes could be used for De-ID. For example, if $\mathbf{m}^i = 1$, the decoder can change the face along the i-th attribute. *Note that, the changing is optional and our AINet allows to change or preserve the specified attributes. In contrast, if* $\mathbf{m}^i = 0$, *the modifications on the i-th attribute are not allowed.* As a result, $\mathbf{f}_\text{a}$ contains the accessible attribute information. Intuitively, the vector $\mathbf{m}$ specifies which attributes could be changed to the original face. Its functionality is totally different from the target attribute vector $\mathbf{a}_\text{t}$ indicating which attributes should be included in the generated face. We detail the AINet in Sec. 4.3.

After obtaining $\mathbf{f}_\text{a}$, we fuse it with the latent feature $\hat{\mathbf{F}}^L_{\text{en}}$ and then feed them to the decoder to generate the anonymized face

$$\hat{\mathbf{I}} = \phi^{-1}(\{\hat{\mathbf{F}}^l_{\text{en}}\}^L_{l=1}, \mathbf{f}_\text{a}) = \phi^{-1}_L([\cdots \phi^{-1}_2([\phi^{-1}_1([\hat{\mathbf{F}}^L_{\text{en}}, \mathbf{f}_\text{a}]), \hat{\mathbf{F}}^{L-1}_{\text{en}}])]). \tag{5}$$

With the SCU and AINet, the A^{3}GAN can generate anonymized and realistic faces following the mask vector $\mathbf{m}$. The whole pipeline is displayed in Fig. 3. During the testing process, given an input face and $\mathbf{m}$, the anonymized face is automatically generated by changing the modifiable attributes specified by $\mathbf{m}$. Intuitively, the whole architecture is similar with the U-Net shape where the skip connections are replaced with SCUs for identity removal along different feature levels and the latent code is injected with the required attribute information.

4.2 Suppressive Convolutional Unit (SCU)

As the designing in Sec. 4.1, the SCUs are desired to remove the original identity information progressively while keeping the identity-irrelevant facial information along the encoder to ensure both anonymity and realism.

Motivated by the selective transfer unit (STU) in [20], we design the suppressive convolutional unit (SCU) to filter the extracted features that contain identity information. The STU adopts a GRU-based structure to select attribute-irrelevant encoder features and transfer them to match the requirements of attribute editing. However, the STU is not suitable for face De-ID: *First*, the STU incorporates a target attribute vector to ensure the target attributes occur in new faces. This hard constraint strengthens the editing capability but easily degrades the image quality for multiple attribute changes (see the results of STGAN-5/9/13 in Fig. 2). *Second*, the GRU is originally used to solve the long dependency issues in sequential data, and the state transition and gate mechanisms in GRU implicitly conduct the information selection, complicating the process and also increasing the computational cost.

Beyond STU, we propose the SCU that comprises a transposed convolution and a gated convolution to transfer each encoder feature (see Fig. 3). Specifically, for the l-th layer, a SCU receives two

inputs: one is the l-th encoder feature $\mathbf{F}_{\text{en}}^{l}$ and another is the suppressed feature $\hat{\mathbf{F}}_{\text{en}}^{l+1}$ from a higher layer. For the initial suppressed feature, $\hat{\mathbf{F}}_{\text{en}}^{L} = \mathbf{F}_{\text{en}}^{L}$. The SCU is formulated as

$$\hat{\mathbf{F}}_{\text{en}_\text{T}}^{l+1} = \mathbf{W}_{\text{T}}^{l} \otimes_{\text{T}} \hat{\mathbf{F}}_{\text{en}}^{l+1}, \tag{6}$$

$$\hat{\mathbf{F}}_{\text{en}}^{l} = \sigma\left(\mathbf{W}_{\text{f}}^{l} \otimes \left[\mathbf{F}_{\text{en}}^{l}, \hat{\mathbf{F}}_{\text{en}_\text{T}}^{l+1}\right]\right) \odot \mathbf{F}_{\text{en}}^{l}, \tag{7}$$

where $\mathbf{W}_{\text{T}}^{l}\otimes_{\text{T}}$ denotes the transposed convolution with weights $\mathbf{W}_{\text{T}}^{l}$, and $\mathbf{W}_{\text{f}}^{l}\otimes$ denotes the gated convolution with weights $\mathbf{W}_{\text{f}}^{l}$. The function $\sigma(\cdot)$ is the Sigmoid function, and $\odot$ represents the entry-wise product operation.

The gated convolution in Eq. (7) can be regarded as learning a point-wise intensity map that indicates the relevance score of the face identity in encoder features, thus suppressing the original identity information and passing the identity-irrelevant information to the decoder and the SCU at the shallow layer.

Compared with STU, the SCU has the following two advantages: *First*, the SCU is to suppress the identity information in the original image instead of incorporating specified attributes like STU. Without hard constraints on attribute synthesizing, the SCU is conducive to generating more realistic faces (see the examples of A^{3}GAN in Fig. 2). *Second*, we use gated convolutions [44] to remove original identity information in an explicit manner, which provides a learnable dynamic selection mechanism for identity information removal and is more computationally efficient than GRU adopted by the STU. The parameters of SCU are 57.8% of those of STU for the same network architecture settings. The importance of SCU on anonymity and realism is validated in the ablation study section.

4.3 Attribute-aware Injective Network (AINet)

Instead of manually setting target attribute vectors [9, 20], we propose an attribute-aware injective network (AINet) to automatically generate the targeted attribute information in the latent space. Specifically, the AINet consists of three components, a fully-connected neural network (FCNet) with six FC layers and ReLU activations, a learnable affine transform (AffNet), and an adaptive instance normalization (AdaIN). With $\mathbf{f}_z$ and $\mathbf{m}$, the FCNet predicts the target attribute feature denoted as $\mathbf{f}_{\text{m}}$ in the attribute latent space, which is suitable for face De-ID and meets the requirement of $\mathbf{m}$. The AffNet is to align the predicted feature $\mathbf{f}_{\text{m}}$ to match the channels of the feature $\mathbf{F}_{\text{en}}^{L}$ and predict the transformed latent feature $\mathbf{f}_{\text{a}}$. Finally, the AdaIN fuses the $\mathbf{f}_{\text{a}}$ and the latent feature $\mathbf{F}_{\text{en}}^{L}$ of the input face for subsequent decoding, and the implementation of AdaIN is the same as that in [15]:

$$\text{AdaIN}(\mathbf{f}_{\text{a}}, \mathbf{F}_{\text{en}}^{L}) = \mathbf{f}_{\text{a,s}} \frac{\mathbf{F}_{\text{en}}^{L} - \mu(\mathbf{F}_{\text{en}}^{L})}{\sigma(\mathbf{F}_{\text{en}}^{L})} + \mathbf{f}_{\text{a,b}} \tag{8}$$

where $\mathbf{f}_{\text{a,s}}$ and $\mathbf{f}_{\text{a,b}}$ are the scale and bias of $\mathbf{f}_{\text{a}}$ obtained by the AffNet, and $\mu(\mathbf{F}_{\text{en}}^{L})$ and $\sigma(\mathbf{F}_{\text{en}}^{L})$ are the channel-wise mean and variance of $\mathbf{F}_{\text{en}}^{L}$ for normalization.

We now explain the above components of AINet. The FCNet aims to map the initial latent feature $\mathbf{f}_z$ to the attribute-oriented feature $\mathbf{f}_{\text{m}}$ that contains new identity information for face De-ID. Compared to the manually determined target attribute vector $\mathbf{a}_{\text{t}}$ used in facial attribute editing, the learned $\mathbf{f}_{\text{m}}$ can be treated as a calibrated version, which provides more reasonable target attributes and avoids the image quality degradation due to multiple attribute changes. The facial attribute mask vector $\mathbf{m}$ is to guide the generation of $\mathbf{f}_{\text{m}}$. If the $\mathbf{m}$ is not provided, the FCNet with only $\mathbf{f}_z$ as input will result in unpredictable facial attributes. Our ablation study shows that adding the mask vector can provide a more accurate control over the facial attributes and identity.

For the AdaIN, it realizes the attribute injection from a style transfer view. The prior works [9, 20] simply concatenate the target attribute vectors to the encoder features, which cannot guarantee the anonymity of generated faces. The AdaIN directly transfers the faces with the learned attribute information, helping to increase the anonymity of faces. The effectiveness of AdaIN on face De-ID is validated in our ablation study.

4.4 Loss Functions

To train our De-ID generator effectively, we construct a discriminator to cover two tasks (see Fig. 3): to determine whether the input face image is real or fake and to estimate whether the attributes are included in the input face image or not. The first task is a binary classification task and represented as $\text{D}_{\text{cls}}(\cdot)$, and the second one is formulated as multiple binary classification tasks and named as $\text{D}_{\text{att}}(\cdot)$. We consider five loss functions to make the generated face image $\hat{\mathbf{I}}$ anonymous, realistic and controllable, meet the attributes specified by $\mathbf{m}$, and conduct De-ID effectively.

Attribute loss function for the original image (*i.e.*, $\mathcal{L}_{\text{att1}}$). We set an attribute loss function for the original image to empower the discriminator to predict the attributes of the input image correctly. Specifically, given an input image $\mathbf{I}$, we combine N binary cross-entropy loss functions

$$\mathcal{L}_{\text{att1}}(\mathbf{I}, \mathbf{a}_{\text{s}}) = -\sum_{i=1}^{N} [\mathbf{a}_{\text{s}}^{i} \log \text{D}_{\text{att}}^{i}(\mathbf{I}) + (1 - \mathbf{a}_{\text{s}}^{i}) \log(1 - \text{D}_{\text{att}}^{i}(\mathbf{I}))] \tag{9}$$

where $\mathbf{a}_{\text{s}}$ is a vector indicating the original source attributes $\mathbf{a}_{\text{s}}^{i}$ contained in $\mathbf{I}$, and $\text{D}_{\text{att}}^{i}(\mathbf{I})$ is the prediction of the i-th attribute.

Attribute-aware anonymization loss function for the generated image (*i.e.*, $\mathcal{L}_{\text{att2}}$). Straightforwardly, when the input is a generated image, we can also force its attributes to be the same with the target attributes like Eq. (9). However, in this work, the target attribute information is not specified and unknown. We only provide a constraint on the target attributes via AINet, *i.e.*, the attribute mask vector $\mathbf{m}$ in Eq. (4), which indicates whether an attribute is changeable or not. To make the generated image follow the $\mathbf{m}$ constraint and achieve effective face anonymization, we first drive a regularization attribute vector $\mathbf{a}_{\text{m}}$ according to $\mathbf{m}$ and the original attribute vector $\mathbf{a}_{\text{s}}$

$$\mathbf{a}_{\text{m}} = \text{xnor}(\mathbf{a}_{\text{s}}, \mathbf{m}), \tag{10}$$

where 'xnor$(\cdot)$' represents the element-wise exclusive-NOR operation. Intuitively, the $\mathbf{a}_{\text{m}}$ is the inverse of $\mathbf{a}_{\text{s}}$ at the attributes $\{i|\mathbf{m}^{i} = 0\}$ but keeps the same with $\mathbf{a}_{\text{s}}$ at $\{i|\mathbf{m}^{i} = 1\}$. As a result, $\mathbf{a}_{\text{m}}$ indicates which attributes the generated image cannot contain and which attributes should be preserved. For example, if $\mathbf{a}_{\text{m}}^{i} = 1$, it means the generated image $\hat{\mathbf{I}}$ should not contain the i-th attribute; if $\mathbf{a}_{\text{m}}^{i} = 0$, we desire the i-th attribute to occur in the generated image $\hat{\mathbf{I}}$. The above design encourages the generated attributes to

be greatly different from the original ones and leads to effective face DeID.

To this end, we propose to minimize the inverse of the cross-entropy loss functions based on $\mathbf{a}_\mathrm{m}$, *i.e.*,

$$\mathcal{L}_{\mathrm{att2}}(\hat{\mathbf{I}}, \mathbf{a}_\mathrm{m}) = (-\sum_{i=1}^{N}[\mathbf{a}_\mathrm{m}^i \log \mathrm{D}_{\mathrm{att}}^i(\hat{\mathbf{I}}) + (1-\mathbf{a}_\mathrm{m}^i)\log(1-\mathrm{D}_{\mathrm{att}}^i(\hat{\mathbf{I}}))] + \epsilon)^{-1}, \quad (11)$$

where ϵ is a minimal constant to avoid dividing by zero.

Anonymization loss function for face De-ID (*i.e.*, $\mathcal{L}_{\mathbf{deid}}$). We set an anonymization loss to make sure the face De-ID objective achieved. Specifically, we adopt the ArcFace loss [4] and formulate the de-identification loss as

$$\mathcal{L}_{\mathrm{deid}}(\mathbf{I}, \hat{\mathbf{I}}) = \mathrm{Sim}(\varphi(\mathbf{I}), \varphi(\hat{\mathbf{I}})) \quad (12)$$

where $\varphi(\cdot)$ is a pretrained ArcFace model to extract face features, and $\mathrm{Sim}(\cdot,\cdot)$ denotes the Cosine similarity loss.

Binary cross-entropy loss function for the discriminator (*i.e.*, $\mathcal{L}_{\mathbf{cls}}$). We also add a binary-entropy loss function for the discriminator, *i.e.*, $\mathcal{L}_{\mathrm{cls}}(\mathbf{I}, \hat{\mathbf{I}})$, which enables the discriminator to determine whether the input is the original image or the generated one and encourage the generator to produce realistic face images.

Adversarial loss function for the generator (*i.e.*, $\mathcal{L}_{\mathbf{adv}}$). We further adopt the adversarial loss function used in Wasserstein GAN [8] denoted as $\mathcal{L}_{\mathrm{adv}}(\hat{\mathbf{I}})$ to enhance the capability of generating realistic face images (see the definition of $\mathcal{L}_{\mathrm{adv}}$ in Appendix B).

Overall, our final loss function is formulated as

$$\mathcal{L} = \lambda_1\mathcal{L}_{att1} + \lambda_2\mathcal{L}_{att2} + \lambda_3\mathcal{L}_{deid} + \lambda_4\mathcal{L}_{adv} + \mathcal{L}_{cls} \quad (13)$$

where λ_1, λ_2, λ_3, and λ_4 are hyper parameters for training.

5 EXPERIMENTS

5.1 Experimental Setup

Tasks and Datasets. We use four tasks on four datasets to thoroughly evaluate our A^{3}GAN. For the anonymity, the anonymized faces should not be identified by human and machine recognizers, so we perform face verification on CelebA [21] to assess the De-ID performance. For the realism, the anonymized faces should look natural, so we perform face detection on WIDER FACE [43] to verify whether the anonymized faces can be easily detected like real faces. For the controllability, we conduct facial expression recognition and fatigue detection on ExpW [46] and NTHU-DDD [41], respectively, to examine whether the facial details can be controlled flexibly. Our A^{3}GAN is trained on CelebA and tested on all the above datasets to verify its adaptability to different datasets.

Baselines. We adopt three types of baselines for comparisons. Since our A^{3}GAN is based on facial attribute manipulation, we use two typical facial attribute editing methods, STGAN [20] and L2M-GAN [42] (each has three variants), as the first type of baseline. The second type is early face De-ID methods, *i.e.*, 8×8 pixelation [27], 9×9 Gaussian blur [28] and black-out [32]. The third type is the state-of-the-art deep generative face De-ID methods, including DeepPrivacy [12] and CIAGAN [22].

Implementation Details. The A^{3}GAN is implemented using PyTorch [30] and trained on a single NVIDIA RTX 3090 GPU. During the training, the batch size is set to 32, and the Adam [16] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ is used as the optimizer. The learning rate starts from 0.0002 and is divided by 10 after 80 epochs. The model training is finished at 120 epochs. For the hyper-parameters, the constant in Eq. (11) is set as $\epsilon = 1$, and the balance parameters in Eq. (13) are set as $\lambda_1 = 1$, $\lambda_2 = 30$, $\lambda_3 = 10$, and $\lambda_4 = 1$.

5.2 Comparison with State-of-the-art methods

Face Verification vs. Image Quality. We use FaceNet [34] and CurricularFace [11] to evaluate the anonymity of face De-ID methods with the evaluation metric of verification accuracy, and a lower verification accuracy represents a higher anonymity. We adopt BRISQUE [25] and FID [10] to evaluate the realism of anonymized faces, and lower values of BRISQUE and FID denote higher image quality. The performance comparison of SOTA facial attribute editing methods, face De-ID methods and our A^{3}GAN in terms of face verification and image quality is shown in Fig. 4, and the corresponding numerical results are also reported in Appendix E.

In Fig. 4, the STGAN-1/3/5 and L2M-GAN-1/3/5 denote the STGAN and L2M-GAN with one/three/five changed facial attributes. The detail of changed facial attributes is provided in Appendix D. The verification accuracy values of STGAN-1 and L2M-GAN-1 are very high (nearly 1), but decrease with increasing the number of facial attributes, since changing more facial attributes will deteriorate the image quality (see also Fig. 2). This again verifies the contradiction of anonymity and realism for traditional facial attribute editing methods, indicating their unsuitability for face De-ID.

The naïve face De-ID methods especially the pixelation and black-out have a high anonymity, but at a cost of a significant drop in image quality, since the face regions are all seriously distorted or removed (see the examples in Fig. 7). DeepPrivacy and CIAGAN both achieve considerable anonymization performance with all verification accuracy values less than 5%, and their image quality is also greatly improved compared to naïve face De-ID methods.

Similar to the variants of STGAN and L2M-GAN, A^{3}GAN-1/3/5/10 represents changing one, three, five and all ten attributes, but A^{3}GAN-0 denotes no changing of attributes. All the five versions of A^{3}GAN achieve remarkable anonymization performance, prevailing over all the competitors by a sizeable margin. The verification accuracy values of A^{3}GAN-5/10 are even zeros, which mean a perfect face De-ID when more attributes are changed. We can also see that the A^{3}GAN with little attribute changes has better image quality than previous face De-ID methods. Although A^{3}GAN-10 has lower image quality scores, this is just a stress testing, and in practice, we do not need to modify all facial attributes for anonymization.

Face Verification vs. Face Detection. We employ SSH [26] and DSFD [18] to evaluate the effect of face De-ID on face detection with average precision (AP), for which larger values mean better face detection and thus more natural anonymized faces. The performance comparison of SOTA facial attribute editing methods, face De-ID methods and our A^{3}GAN in terms of face verification and face detection is shown in Fig. 5.

We observe that the STGAN-1 and L2M-GAN-1 have high AP values, and gradually decrease with more changed facial attributes, which is consistent with that in Fig. 4. For the three naïve face De-ID methods, the coarse anonymized faces can still be detected to a certain extent. This is because they only obfuscate the face

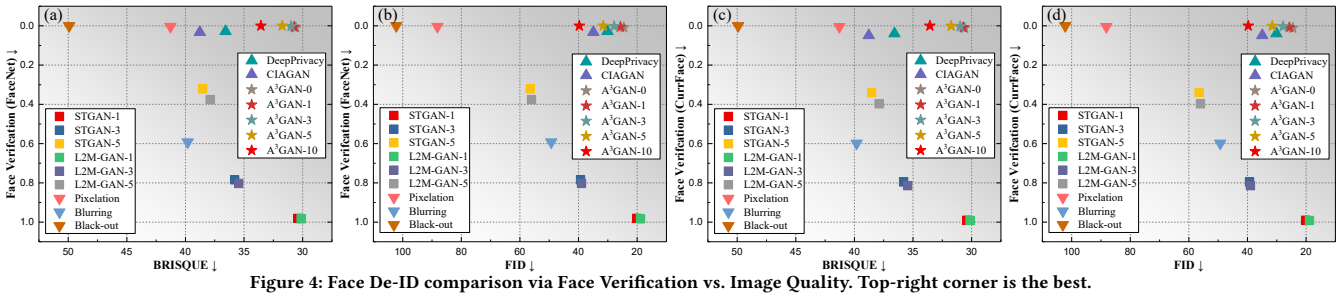

Figure 4: Face De-ID comparison via Face Verification vs. Image Quality. Top-right corner is the best.

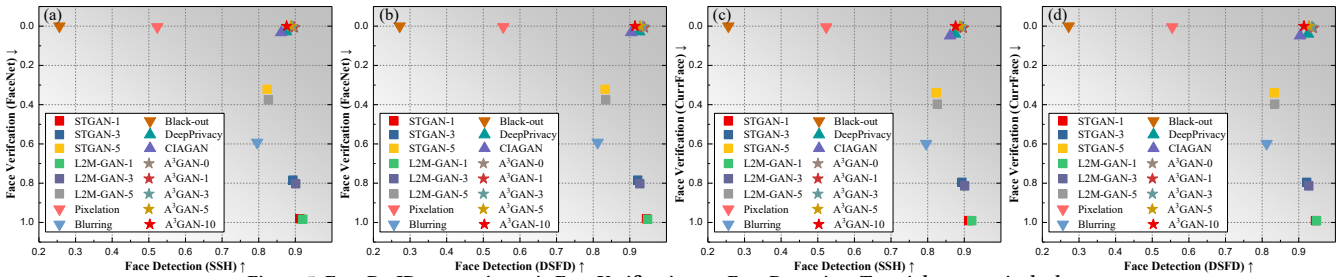

Figure 5: Face De-ID comparison via Face Verification vs. Face Detection. Top-right corner is the best.

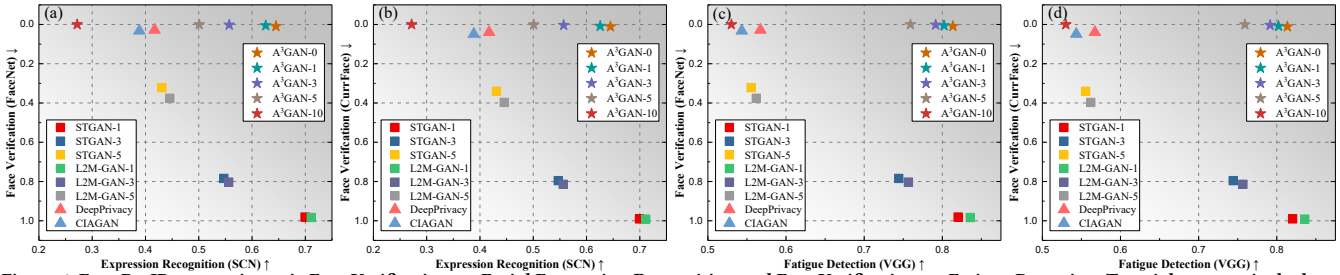

Figure 6: Face De-ID comparison via Face Verification vs. Facial Expression Recognition and Face Verification vs. Fatigue Detection. Top-right corner is the best.

region, but the head region is remaining detectable to face detectors. For the DeepPrivacy and CIAGAN, their AP values are all increased compared to the naïve face De-ID methods, but still inferior to A^{3}GAN due to the visual distortions caused by total face generation.

Our A^{3}GAN obtains higher AP than all face De-ID methods. Note that, the AP of A^{3}GAN does not decrease quickly with the increasing of changed attribute numbers (the A^{3}GAN-10 even modifies all the facial attributes), and this can be attributed to the special design of anonymization loss (see the explanation of Eq. (11)).

Face Verification vs. Facial Expression Recognition and Face Verification vs. Fatigue Detection. To evaluate the controllability of face De-ID methods, we investigate whether or not the face De-ID can precisely control the facial expression, and we utilize an expression recognition network SCN [40] considering seven basic facial expressions for this experiment, in which a larger recognition accuracy denotes a better preservation of facial expressions. The performance comparison of SOTA facial attribute editing methods, face De-ID methods and our A^{3}GAN in terms of face verification and facial expression recognition is shown in Fig. 6 (a) and (b).

Fatigue detection is another task to evaluate the controllability of face De-ID, which should not affect some constraint facial regions, such as eyes and mouth. We build a fatigue detection model based on a VGG network in [1], and adopt the detection accuracy as a metric. The performance comparison in terms of face verification and fatigue detection is shown in Fig. 6 (c) and (d). Since pixelation, blurring and black-out heavily disturb the facial details and also should be correctly detected by face detection first, we do not report their results for these two tasks.

The STGAN and L2M-GAN with little changed attributes have higher accuracy for facial expression recognition and fatigue detection than face De-ID methods, since we avoid changing eyes/mouth-related attributes. (see Appendix D for changed attributes).

DeepPrivacy and CIAGAN do not consider the facial details during the anonymization process, so they obtain low accuracy values, implying that they cannot be applied to facial expression recognition and fatigue detection tasks.

Our A^{3}GAN achieves better trade-off between anonymity and controllability, since it can control the facial attributes flexibly and circumvent the problem of DeepPrivacy and CIAGAN. The recognition accuracy and detection accuracy of A^{3}GAN-0 are comparable to those of STGAN and L2M-GAN, but drop apparently for A^{3}GAN-10, because all facial attributes including eyes and mouth-related attributes are modified.

5.3 Qualitative Evaluation

We display various types of anonymized faces in Fig. 7 for visualization comparison. We can observe that our A^{3}GAN provides a higher visual quality for face De-ID than state-of-the-art methods. In contrast, there are noticeable artifacts on the anonymized faces of DeepPrivacy and CIAGAN (see the yellow arrows in Fig. 7). For

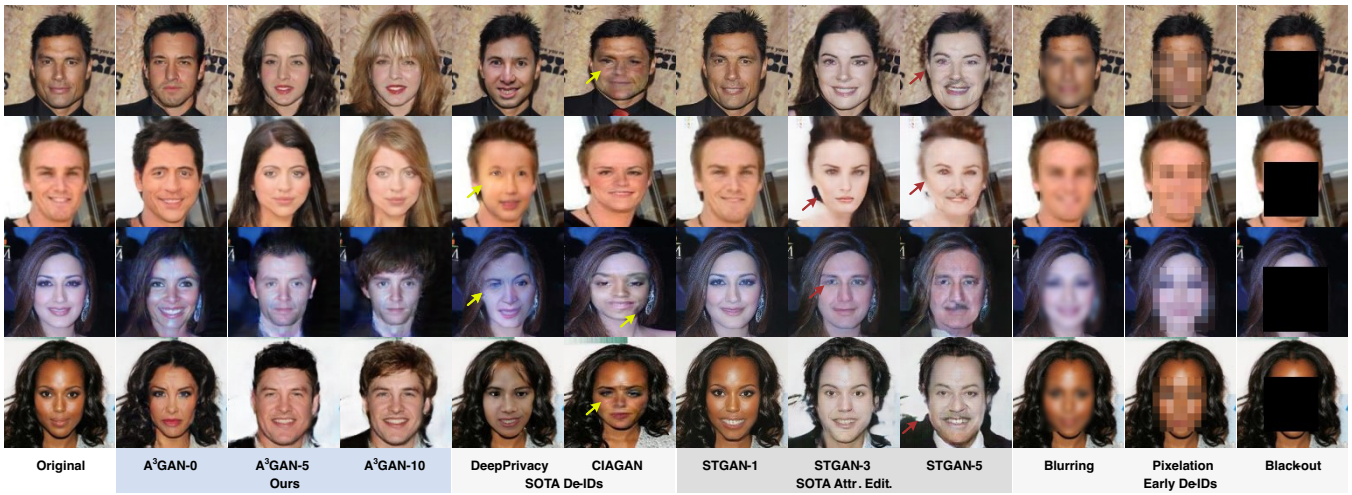

Figure 7: Visual comparisons of A^{3}GAN with different changing attributes, SOTA face De-IDs (*i.e.*, DeepPrivacy [12], CIAGAN [22]), SOTA attribute editting method (*i.e.*, STGAN [20]), Blurring, Pixelation and Black-out. Yellow and red arrows highlight the visual artifacts in anonymized faces.

the faces generated by STGAN, the anonymity and realism still cannot be satisfied simultaneously (see the red arrows in Fig. 7), as discussed in Sec. 3.2. Besides, our A^{3}GAN also has a powerful generative capability. For example, it can transfer a blurry face into a clear face (see row 2 in Fig. 7). When transferring a female to a male, or conversely, the long hair will be replaced by generated background or the image background will be rendered with long hair. Furthermore, our A^{3}GAN also exhibits high diversity of anonymized faces by tuning facial attribute mask vectors (see the gender or face style changes).

5.4 Ablation Study

Our A^{3}GAN has two important modules, namely a SCU for suppressing original identity information and an AINet for controlling injected identity information. The AINet also has some sub-modules, such as facial attribute mask vector and AdaIN. To verify their importance to face De-ID, we perform an ablation study, in which we compare the performance of A^{3}GAN with and without these modules on face verification, fatigue detection, and image quality. The configuration here is the same as that of A^{3}GAN-5. The ablation results are reported in Table 1, in which A^{3}GAN w/ STU denotes replacing the SCU with STU, A^{3}GAN w/o SCU/mask denote removing the corresponding modules in original A^{3}GAN, and A^{3}GAN w/o AdaIN denotes replacing the AdaIN with concatenation operation.

Comparing the results in the first row and those in the last row, We observe that the SCU indeed contributes more to the anonymity and realism than the STU. We also observe from the second row that removing the SCU increases the verification accuracy of A^{3}GAN from 0.05 to 2.92, demonstrating the effectiveness of SCU in suppressing original identity information. The A^{3}GAN without SCU also increases the accuracy of the fatigue detection task, since the preservation of original identity contains more facial expression information that is needed for fatigue detection.

The A^{3}GAN without mask vector (A^{3}GAN w/o mask) has little effect on the face verification, but it loses the ability to control the facial attributes, so its detection accuracy on fatigue detection task decreases drastically.

Table 1: Ablation study of A^{3}GAN on different vision tasks. The top three results are highlighted in red, purple, and green, respectively.

Different A^{3}GAN Models	Face Ver.	Fatigue Det.	Image Quality
	CurrFace [11] ↓	VGG [1] ↑	FID [10] ↓
A^{3}GAN w/ STU	**2.39**	73.32	32.01
A^{3}GAN w/o SCU	2.92	**76.10**	**31.24**
A^{3}GAN w/o mask	**1.06**	53.06	32.29
A^{3}GAN w/o AdaIN	3.95	**74.39**	**31.90**
A^{3}GAN	**0.05**	**75.92**	**31.55**

The results in the fourth row indicate that the learned attribute information should be fused with the encoder features by using AdaIN other than concatenation operation.

6 CONCLUSION

In this paper, we studied the problems of face De-ID from a micro perspective. We proposed an attribute-aware anonymization network (A^{3}GAN) by using facial attribute manipulations instead of total face generation. In particular, we designed a novel generative network consisting of a suppressive convolutional unit (SCU) and an attribute-aware injective network (AINet) to filter out original identity information and introduce new identity information, respectively. This moderate De-ID fashion increases the anonymity while maintaining the realism of anonymized faces. Furthermore, we design a new loss function with an attribute mask to control the changes or preservation of facial attributes, enabling more granularly anonymization and more diverse anonymized faces. The comprehensive experiments on four types of vision tasks have demonstrated the state-of-the-art performance of our A^{3}GAN.

7 ACKNOWLEDGMENTS

This work was supported by the National Research Foundation, Singapore under its AI Singapore Programme (Award No: AISG2-RP-2020-019), Singapore National Cybersecurity R&D Program No. NRF2018NCR-NCR005-0001, National Satellite of Excellence in Trustworthy Software System No. NRF2018NCR-NSOE003-0001, and NRF Investigatorship No. NRF-NRFI06-2020-0001. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC).

REFERENCES

[1] Bhakti Baheti, Suhas Gajre, and Sanjay Talbar. 2018. Detection of distracted driver using convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1032–1038.

[2] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. 2018. Anonymizing k-facial attributes via adversarial perturbations. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 656–662.

[3] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. SSCGAN: Facial attribute editing via style skip connections. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. 414–429.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 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.

[5] Julia Dietlmeier, Joseph Antony, Kevin McGuinness, and Noel E O'Connor. 2021. How important are faces for person re-identification?. In *International Conference on Pattern Recognition (ICPR)*. IEEE, 6912–6919.

[6] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. 2020. Image inpainting: A review. *Neural Processing Letters* 51, 2 (2020), 2007–2028.

[7] Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9378–9387.

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of Wasserstein GANs. arXiv:arXiv preprint arXiv:1704.00028

[9] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. AttGAN: Facial attribute editing by only changing what you want. *IEEE transactions on image processing* 28, 11 (2019), 5464–5478.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[11] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. CurricularFace: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5901–5910.

[12] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. DeepPrivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*. 565–578.

[13] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. 448–456.

[14] Amin Jourabloo, Xi Yin, and Xiaoming Liu. 2015. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*. 278–285.

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[16] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[17] Jeong-gi Kwak, David K Han, and Hanseok Ko. 2020. CAFE-GAN: Arbitrary Face Attribute Editing with Complementary Attention Feature. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 524–540.

[18] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. 2019. DSFD: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5060–5069.

[19] Yongxiang Li, Qianwen Lu, Qingchuan Tao, Xingbo Zhao, and Yanmei Yu. 2021. SF-GAN: Face De-identification Method without Losing Facial Attribute Information. *IEEE Signal Processing Letters* (2021).

[20] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. 2019. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3673–3682.

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

[22] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. CIAGAN: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5447–5456.

[23] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. 2021. Privacy–Enhancing Face Biometrics: A Comprehensive Survey. *IEEE Transactions on Information Forensics and Security* (2021).

[24] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2020. PrivacyNet: semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing* 29 (2020), 9400–9412.

[25] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[26] Mahyar Najibi, Pouya Samangouei, Rama Chellappa, and Larry S Davis. 2017. SSH: Single stage headless face detector. In *Proceedings of the IEEE international conference on computer vision*. 4875–4884.

[27] Carman Neustaedter, Saul Greenberg, and Michael Boyle. 2006. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 1 (2006), 1–36.

[28] Elaine M Newton, Latanya Sweeney, and Bradley Malin. 2005. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering* 17, 2 (2005), 232–243.

[29] José Ramón Padilla-López, Alexandros Andre Chaaraoui, and Francisco Flórez-Revuelta. 2015. Visual privacy protection methods: A survey. *Expert Systems with Applications* 42, 9 (2015), 4177–4195.

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[31] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. 2016. Invertible conditional GANs for image editing. arXiv:arXiv preprint arXiv:1611.06355

[32] Slobodan Ribaric, Aladdin Ariyaeeinia, and Nikola Pavesic. 2016. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication* 47 (2016), 131–151.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. 234–241.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[35] SecureTrust. 2021. Data Privacy. https://www.securetrust.com/data-privacy/. Accessed: 2021-08-31.

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

[37] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5050–5059.

[38] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. 2018. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European Conference on Computer Vision*. 553–569.

[39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. arXiv:arXiv preprint arXiv:1607.08022

[40] Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. 2020. Suppressing uncertainties for large-scale facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6897–6906.

[41] Ching-Hua Weng, Ying-Hsiu Lai, and Shang-Hong Lai. 2016. Driver drowsiness detection via a hierarchical temporal deep belief network. In *Asian Conference on Computer Vision*. 117–133.

[42] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. 2021. L2M-GAN: Learning To Manipulate Latent Space Semantics for Facial Attribute Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2951–2960.

[43] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5525–5533.

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4471–4480.

[45] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. 2018. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European conference on computer vision (ECCV)*. 417–432.

[46] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision* 126, 5 (2018), 550–569.

[47] Xin Zheng, Yanqing Guo, Huaibo Huang, Yi Li, and Ran He. 2020. A survey of deep facial attribute analysis. *International Journal of Computer Vision* 128, 8 (2020), 2002–2034.

A NETWORK ARCHITECTURE

Our A^{3}GAN has three sub-networks, *i.e.*, a generator, a discriminator and an attribute-aware injective network (AINet). Their network architectures are shown in Table 2-4. In the following tables, the Conv(in_dim, out_dim, k, s) and DeConv(in_dim, out_dim, k, s) denote the convolutional layer and transposed convolutional layer with input channels in_dim, output channels out_dim, kernel size k, and stride s. FC(in_dim, out_dim) represents a fully-connected layer with input channels in_dim, output channels out_dim. BN and IN denote batch normalization [13] and instance normalization [39], respectively. N is the number of facial attributes.

B DEFINITION OF ADVERSARIAL LOSS

The loss functions of Wasserstein GAN with gradient penalty (WGAN-GP) [8] is widely used in literature. To save space of the paper, we do not provide the detailed definitions of $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{adv}}$ in Sec. 4.4. We give their definitions as follows

$$\mathcal{L}_{\text{cls}}(\mathbf{I}, \hat{\mathbf{I}}) = -\mathbb{E}_{\mathbf{I}}[\mathrm{D}_{\text{cls}}(\mathbf{I})] + \mathbb{E}_{\tilde{\mathbf{I}}}[\mathrm{D}_{\text{cls}}(\hat{\mathbf{I}})] + \lambda \mathbb{E}_{\tilde{\mathbf{I}}}[\left(\left\|\nabla_{\tilde{\mathbf{I}}} \mathrm{D}_{\text{cls}}(\tilde{\mathbf{I}})\right\|_2 - 1\right)^2] \tag{14}$$

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{\tilde{\mathbf{I}}}[\mathrm{D}_{\text{cls}}(\hat{\mathbf{I}})] \tag{15}$$

where $\mathrm{D}_{\text{cls}}(\cdot)$ denotes the binary classifier branch of the discriminator for outputting real or fake predictions, $\mathbf{I}$ and $\hat{\mathbf{I}}$ are original real face image and anonymized face image, $\tilde{\mathbf{I}}$ is any point sampled between $\mathbf{I}$ and $\hat{\mathbf{I}}$, and λ is a gradient penalty coefficient with default value 10.

C ATTRIBUTE-AWARE ANONYMIZATION LOSS FUNCTION

In Sec. 4.4, we explain why the anonymization loss function for the generated image (*i.e.*, $\mathcal{L}_{\text{att2}}$) is given in a reciprocal form, namely the inverse of the cross-entropy loss function. Here, we provide a further explanation and also use experiment to validate our design.

The facial attribute editing methods [9, 20] encourage the discriminator to correctly recognize the target facial attributes, so they design the loss function for the generated image by minimizing the cross entropy with target attribute labels:

$$\mathcal{L}_{\text{att2}}(\hat{\mathbf{I}}, \mathbf{a}_{\text{t}}) = -\sum_{i=1}^{N} [\mathbf{a}_{\text{t}}^i \log \mathrm{D}_{\text{att}}^i(\hat{\mathbf{I}}) + (1 - \mathbf{a}_{\text{t}}^i) \log(1 - \mathrm{D}_{\text{att}}^i(\hat{\mathbf{I}}))] \tag{16}$$

where $\hat{\mathbf{I}}$ is the generated image, $\mathbf{a}_{\text{t}}$ is the target attribute with N attributes, and D_{att} is the multiple binary classifier branch of the discriminator.

However, this design of loss function has two drawbacks. First, it will cause unreasonable face semantics. For example, if we choose a "moustache" attribute in the target attribute label vector, there will be moustaches on the generated faces regardless of the gender. Second, it will force the target attributes with multiple attribute changes occur in the generated images, which easily lead to deteriorated image quality, as discussed in Sec. 3.2.

Table 2: Generator architecture.

Layer	Generator
E1	Conv(3,64,4,2), BN, Leaky ReLU
E2	Conv(64,128,4,2), BN, Leaky ReLU
E3	Conv(128,256,4,2), BN, Leaky ReLU
E4	Conv(256,512,4,2), BN, Leaky ReLU
E5	Conv(512,1024,4,2), BN, Leaky ReLU
D1	DeConv(1024,1024,4,2), BN, ReLU
D2	DeConv(1536,512,4,2), BN, ReLU
D3	DeConv(768,256,4,2), BN, ReLU
D4	DeConv(384,128,4,2), BN, ReLU
D5	DeConv(192,3,4,2), BN, ReLU

Table 3: Discriminator architecture.

Layer	Discriminator	
	Adv. Classifier	Attr. Classifier
1	Conv(3,64,4,2), IN, Leaky ReLU	
2	Conv(64,128,4,2), IN, Leaky ReLU	
3	Conv(128,256,4,2), IN, Leaky ReLU	
4	Conv(256,512,4,2), IN, Leaky ReLU	
5	Conv(512,1024,4,2), IN, Leaky ReLU	
6	FC(16384,1024), Leaky ReLU	FC(16384,1024), Leaky ReLU
7	FC(1024,1)	FC(1024,N), Sigmoid

Table 4: AINet architecture.

Layer	AINet
1	FC(512+N,512), Leaky ReLU
2	FC(512,512), Leaky ReLU
3	FC(512,512), Leaky ReLU
4	FC(512,1024), Leaky ReLU
5	FC(1024,1024), Leaky ReLU
6	FC(1024,1024), Leaky ReLU

Aiming at the above problems, we encourage the generator to synthesis new facial attributes that deviate from the source facial attributes with the condition of a mask vector. We achieve this by minimizing the inverse of the cross-entropy loss functions with the regularization attribute vector $\mathbf{a}_{\text{m}}$, in which the preserved attributes are flipped in advance (see Eq. (10)). The final anonymization loss function for the generated image is defined as Eq. (11)

Fig. 8 compares our loss design (Eq. (11)) and that in STGAN [20] (Eq. (16)). We observe that the female faces generated by STGAN have a moustache or a beard, but our A^{3}GAN can avoid these unreasonable attribute changes. Besides, our loss design can also maintain a higher image quality than than the STGAN when changing multiple facial attributes, and their visualization comparison is shown in Fig. 7.

D FACIAL ATTRIBUTES USED IN EXPERIMENTS

Our A^{3}GAN is based on facial attribute manipulations for face anonymization, and the controllability of A^{3}GAN can be realized

Table 5: Quantitative results of different face De-ID methods in terms of face verification, face detection, facial expression recognition, fatigue detection and image quality assessment. The top three results are highlighted in red, purple, and green, respectively.

De-identification Methods	Face Ver.		Face Det.		Exp. Recog.	Fatigue Det.	Image Quality	
	FaceNet [34] ↓	CurrFace [11] ↓	SSH [26] ↑	DSFD [18] ↑	SCN [40] ↑	VGG [1] ↑	BRISQUE [25] ↓	FID [10] ↓
Original	98.54	99.21	91.98	95.33	73.23	87.62	30.05	–
STGAN-1 [20]	98.12	99.06	**91.26**	**94.57**	**69.95**	**82.02**	**30.41**	**20.14**
STGAN-3 [20]	78.44	79.56	89.35	92.16	54.70	74.43	35.80	39.29
STGAN-5 [20]	32.18	34.02	82.30	83.25	43.10	55.54	38.52	56.51
L2M-GAN-1 [42]	98.36	99.15	**92.05**	**94.86**	**71.12**	**83.58**	**30.13**	**18.83**
L2M-GAN-3 [42]	80.33	81.42	90.11	92.74	55.63	75.67	35.46	38.98
L2M-GAN-5 [42]	37.59	39.71	82.65	83.43	44.56	56.20	37.89	56.05
Pixelation	0.36	0.44	52.32	55.46	–	–	41.28	88.16
Blurring	59.23	59.82	79.59	81.23	–	–	39.81	49.31
Black-out	**0.00**	**0.00**	25.59	27.18	–	–	49.92	102.18
DeepPrivacy [12]	2.81	3.95	87.48	92.55	41.72	56.72	36.58	30.12
CIAGAN [22]	3.26	4.87	86.15	90.37	38.85	54.34	38.77	34.95
A³GAN-1	0.52	0.81	89.66	93.83	62.57	80.22	30.72	25.69
A³GAN-3	0.20	0.33	89.37	93.62	55.71	79.15	30.98	27.83
A³GAN-5	**0.00**	**0.05**	88.82	92.79	55.09	75.92	31.75	31.55
A³GAN-0	0.86	1.07	**89.73**	**94.08**	**64.49**	**81.36**	**30.59**	**24.53**
A³GAN-10	**0.00**	**0.00**	87.63	91.42	27.18	53.01	33.56	39.74

Figure 8: Face de-identification using our A³GAN and SOTA facial attribute editing method STGAN [20]. For each column, we only change one facial attribute, *i.e.*, "Mustache" or "Beard".

by the facial attribute mask vector. We use different facial attribute mask vectors for different model configurations. A total of ten facial attributes are used in our experiments, including "Bangs", "Blond Hair", "Brown Hair", "Bushy Eyebrow", "Male", "Mouth Slightly Open", "Mustache", "Narrow Eyes", "Pale Skin", "Young". The mask vector is a ten-length binary vector, in which 0 and 1 denote preserving and changing the corresponding facial attribute, respectively. For example, if we set all the bits in the mask vector as zeros, the A³GAN will try to preserve all the facial attributes while changing the face identity. If we set all the bits in the mask vector as ones, the A³GAN will alter all the facial attributes during the anonymization process. These two examples correspond to the A³GAN-0 and A³GAN-10 in Fig. 4 - 7. For the A³GAN-1/3/5 in Fig. 4 - 6, the A³GAN-1 denotes changing the "Male" attribute, the A³GAN-3 denotes changing the "Male", "Mustache" and "Pale Skin" attributes, and the A³GAN-5 denotes changing the "Blond Hair", "Male", "Mustache", "Pale Skin" and "Young" attributes. The eyes and mouth related attributes are preserved for fatigue detection. For the A³GAN-5 in Fig. 7, we modify five conspicuous facial attributes for visualization, including "Male", "Mouth", "Mustache", "Narrow Eyes" and "Pale Skin".

For the baseline STGAN [20] in Fig. 4 - 6, it has three variants based on the number of changed facial attributes. The STGAN-1 denotes changing the "Male" attribute, the STGAN-3 denotes changing the "Male", "Mustache" and "Pale Skin" attributes, and the STGAN-5 denotes changing the "Blond Hair", "Male", "Mustache", "Pale Skin" and "Young" attributes. The STGAN-1/3/5 do not change the eyes and mouth related attributes. The changed attributes for L2M-GAN [42] are also the same. For the STGAN in Fig. 7, the STGAN-1 denotes changing the "Mouth" attribute, the STGAN-3 denotes changing "Mouth", "Male" and "Pale Skin" attributes, and the STGAN-5 denotes changing the "Male", "Mouth", "Mustache", "Pale Skin" and "Young" attributes.

E QUANTITATIVE EVALUATION RESULTS

In Sec. 5.2, we compare our A³GAN with previous face De-ID methods and facial attribute editing methods only using figures. We also report the numerical results in Table 5 for Fig. 4 - 6.