Personalized and Invertible Face De-identification by Disentangled Identity Information Manipulation

Jingyi Cao[1], Bo Liu[2], Yunqian Wen[1], Rong Xie[1], and Li Song[1]

[1]The Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2]School of Computer Science, University of Technology Sydney
{cjycaojingyi,wenyunqian,xierong,song_li}@sjtu.edu.cn, bo.liu@uts.edu.au

Abstract

The popularization of intelligent devices including smartphones and surveillance cameras results in more serious privacy issues. De-identification is regarded as an effective tool for visual privacy protection with the process of concealing or replacing identity information. Most of the existing de-identification methods suffer from some limitations since they mainly focus on the protection process and are usually non-reversible. In this paper, we propose a personalized and invertible de-identification method based on the deep generative model, where the main idea is introducing a user-specific password and an adjustable parameter to control the direction and degree of identity variation. Extensive experiments demonstrate the effectiveness and generalization of our proposed framework for both face de-identification and recovery.

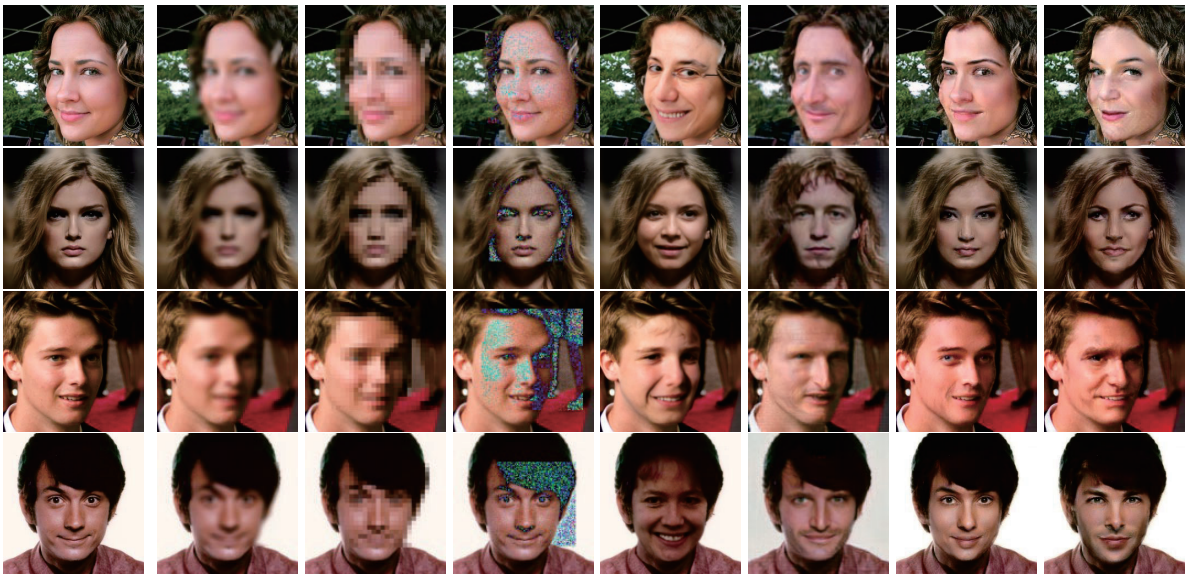
1. Introduction

The widespread use of handheld devices such as smartphones and digital cameras is conducive to image production, and the development of social media promotes wide dissemination and easy access to images along with the increasingly common applications of computer vision technology and deep learning. The above factors lead to serious threats to image privacy and security.

Most importantly, face images are generally considered to contain abundant private information. The earliest techniques obfuscated privacy-sensitive information by pixel-level processing which have been proved vulnerable and poor effects on utility [23]. Recent GAN-based methods like [10, 16] improve the quality and utility of de-identification results remarkably. What's more, the research on disentangled representations [5, 18] contributes to transforming the identity information without changing the other facial attributes, which makes it possible that the de-identified results keep visual similarity with the original.

Most de-identification methods only focus on the protection phase, which can help to protect identity in surveillance for normal situations or uploading images on social media. Considering that when looking for the identity in criminal investigations or sharing pictures with close friends, it is hoped to use the original image instead of the de-identified. Therefore, how to restore the original image is also a critical task. Moreover, notice that the tradeoff between privacy and utility poses a major challenge for all privacy-preserving methods, and different levels of privacy are required in different scenarios. We believe that an ideal comprehensive de-identification method should: a) avoid deteriorating non-sensitive information like facial expression, behavior and so on, b) control the degree of privacy protection according to application, c) be able to restore the original image under security conditions.

To achieve the above targets, this paper proposes a personalized and invertible face de-identification method. The main framework can be summarized in the following three stages: (1) extract disentangled identity and attributes and ensure the attributes unchanged during the de-identification process, (2) calculate the protected or restored identity with the identity modification module based on the password p and privacy level parameter d, (3) implement image reconstruction. As shown in Fig. 1, compared with existing de-identification methods, our approach can retain more similarities with the original. Different from the generative adversarial network conditioned on passwords proposed by Gu *et al*. [8], which needs to retrain the network for different passwords, our encryption process is relatively independent of the deep generative network, so that the password form can be defined more flexibly, the complexity will be reduced greatly and the scope of identity changes can be infinitely expanded. Different from k-Same family algorithms [6, 7, 17] which can provide privacy guarantees and control



(a) Original (b) Blur (c) Pixelation (d) Noise (e)DeepPrivacy (f)Gu *et al*. (g)Ours(d=0) (h)Ours(d=9)

Figure 1: De-identification results compared with existing methods, where (b),(c),(d) are traditional methods and (e),(f) are based on deep learning. From left to right: the original image, Gaussian Blur (s=8), pixelation (8×8), Gaussian noise (σ=15), DeepPrivacy [10], Gu *et al*. [8] and our de-identified results with the minimum and maximum privacy level d.

privacy protection levels for the entire datasets, our method can control the extent of identity variation for each image.

In summary, our main contributions are as follows:

- A general framework that can transform identity of the input while ensuring the other attributes keep similar.
- Personalized de-identification results can be generated with the user-specific password and the degree of identity variation can be controlled.
- The original image can be restored if and only if the corresponding encryption parameters are provided.
- Experimental results show that compared with existing methods, our approach can generate de-identified results with better performance of both privacy and utility, in addition to better-quality recovery results.

2. Related work

In this section, we discuss related work that constitutes the foundations and the motivations of our present work.

2.1. Face De-identification

Traditional face de-identification methods simply use blurring, masking, or pixelation. These methods mainly focus on obfuscating sensitive information directly, which may bring unpleasant artifacts and great harm to image utility. The k-Same family algorithms, based on the k-anonymity [21], can guarantee that each de-identified image is associated with k images to limit the probability of being recognized to $1/k$. There are some improvements based on k-Same [17], for example, k-Same-Select [6] aimed at preserving facial attributes and k-Same-M [7] tried to remove displeasing artifacts owing to misalignment by introducing the Active Appearance Model (AAM). Thanks to the advances in deep generative models, more novel GAN-based de-identification methods have been proposed to produce higher-quality images. DeepPrivacy [10] generates the whole face region to protect full of sensitive information. Recently, there have been some recoverable de-identification methods. Yamac *et al*. [25] introduces a reversible privacy-protection compression method combined with a multi-level encryption scheme for video surveillance applications. Gu *et al*. [8] described a generative adversarial network trained with both images and pre-defined passwords, which can reconstruct the original using the inverse password. A reversible de-identification method for low-resolution video was proposed in [19], which can generate a realistic de-identified stream that contains all the information required in reconstruction. Due to these recoverable methods are mainly based on conditional information, trained for specific passwords, or required additional information about the protection process, the application flexibility and protection security will be affected, while our method can lift such restrictions.

2.2. Disentangled Representation

Many different levels of supervised learning methods have been proposed to learn disentangled representations. Hu *et al*. [9] used two feature blocks mixing and unmix-

ing autoencoders to learn image representations without any data domain knowledge. The research on disentangling identity and attributes has also received extensive attention in the task of de-identification [1] and face-swapping [13] in the hope to focus on identity transform while keeping other attributes unaffected. Most of them follow a similar manner that using a pre-trained face recognition network to infer identity representations and guide the training of generator. Cho *et al*. [3] proposed a network that disentangles latent vectors to identity representations and facial features. Chen *et al*. [2] constructed a variational generative adversarial network VGAN-based disentanglement network for identity switching and expression consistency. Nitzan *et al*. [18] designed a method to disentangle identity-related embeddings and generate synthesis results based on both identity-reference image and attribute-reference image with StyleGAN. Gong *et al*. [5] established a twofold architecture called replacing and restoring variational autoencoders $\left(\mathrm{R}^2\mathrm{VAEs}\right)$ and based on the strategy of factor invariantion to ensure the identity-independent information can be completely disentangled.

3. Problem Formulation

Our identity conversion algorithm mainly possesses de-identification $\mathcal{F}$ and restoration $\mathcal{F}^{-1}$, which both require the input of source face image X, the user-specific password p, and a privacy level parameter d. The password can determine the direction of identity variation and d can control the variation degree. Inspired by Gu *et al*. [8], we mathematically formulate our problem in this section.

De-identification. In order to achieve the effectiveness of identity protection, we aim that the protected image will have different identity information from the original, which can be formulated as,

$$I(\mathcal{F}(X,p,d)) \neq I(X), \tag{1}$$

where $\mathcal{F}(X,p,d)$ indicates the de-identified X with parameters p and d, $I(X)$ represents the identity of image X. Considering the utility of de-identified results, we hope that $\mathcal{F}(X,p,d)$ looks similar to X as well as the face region and keypoints can still be detected by face detector.

Diversity. We can set different passwords p to generate diverse de-identification results, which can promote the security of identity-protection.

$$I(\mathcal{F}(X,p_1,d)) \neq I(\mathcal{F}(X,p_2,d)), p_1 \neq p_2. \tag{2}$$

Controllability. We can control the similarity between de-identified image and the original by the adjustable parameter d as,

$$D(\mathcal{F}(X,p,d_1),X) > D(\mathcal{F}(X,p,d_2),X), d_1 > d_2, \tag{3}$$

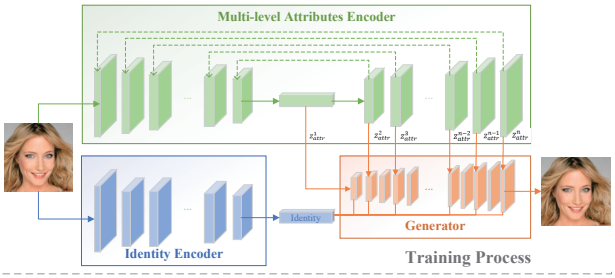


Figure 2: The framework of training process, which includes the identity encoder, the multi-level attributes encoder and the generator.

where $D(X,Y)$ means the identity distance between image X and Y, and the larger distance indicates lower similarity.

Recoverability. If the user takes the de-identified result $\mathcal{F}(X,p,d)$, corresponding password p and d as input, the origin image X can be restored successfully, which can be formulated as,

$$\mathcal{F}^{-1}(\mathcal{F}(X,p,d),p,d) = \hat{X}, I(X) = I(\hat{X}). \tag{4}$$

However, If the attacker tries to restore without the right identity encryption password, he can only get the image with another identity instead of the original one.

$$\mathcal{F}^{-1}(\mathcal{F}(X,p_1,d_1),p_2,d_2) = \hat{Y}, I(X) \neq I(\hat{Y}), \tag{5}$$

where $p_1 \neq p_2, d_1 \neq d_2$. In addition to the above, we also expect that both the de-identified and the restored results have high image quality and satisfactory visual perception.

4. Our Approach

The framework of training process is shown in Fig. 2 and that of protection process and recovery process is presented in Fig. 3, which mainly consists of two encoders E_{id} and E_{attr}, an identity modification module M and a generator G. In the first stage, we extract the image representations and disentangle them into identity $\boldsymbol{z}_{id}$ and attributes $\boldsymbol{z}_{attr}$. Secondly, we calculate the protected identity $\boldsymbol{z}_{new}$ or the restored $\hat{\boldsymbol{z}}_{id}$ by the identity modification module M. Finally, G generates de-identified results based on $\boldsymbol{z}_{new}$ and $\boldsymbol{z}_{attr}$ (or the restored based on $\hat{\boldsymbol{z}}_{id}$ and $\hat{\boldsymbol{z}}_{attr}$). Each part will be described in detail in this section.

4.1. Network Architecture

Identity Encoder. Similar to most researches on disentangled representations of identity and attributes, we use a pre-trained face recognition model as the identity encoder E_{id} and the identity representation $\boldsymbol{z}_{id} = E_{id}(X)$ for the given image X is taken from the last feature vector before the final Fully Connected layer.

Attribute Encoder. In order to retain better details of attributes like expression, pose, illumination and so on, we employ a U-Net-like structure and represent the attributes representations as multi-level feature maps which can be formulated as,

$$\boldsymbol{z}_{attr} = \left\{ \boldsymbol{z}^1_{attr}, \boldsymbol{z}^2_{attr}, \cdots \boldsymbol{z}^n_{attr} \right\}, \tag{6}$$

where $\boldsymbol{z}^k_{attr} (k = 1, 2, \cdots, n)$ means the k-th attributes embedding obtained from k layer of the U-Net decoder.

Identity Modification Module. The identity modification module mainly edits the identity embedding with latent space manipulation. As most of the state-of-the-art face recognition or verification models such as ArcFace [4], CosFace [24], and SphereFace [14] all convert identity features to the hyperspherical space, and use cosine similarity based on angles, which motivates us that rotating the identity vector is a more effective way to change identity information compared with other vector operations like translation. Considering the feasibility of restoration, we hope to introduce a definite process instead of introducing randomness like Gaussian noise. Therefore, we realize de-identification process $\boldsymbol{z}_{new} = M(\boldsymbol{z}_{id}, p, d)$ or restoration process $\hat{\boldsymbol{z}}_{id} = M^{-1}(\hat{\boldsymbol{z}}_{new}, p, d)$ by changing the phase of identity embedding. In more details, during the protection process, we first extract a reference vector $\boldsymbol{z}_r = f(p)$ from the pre-defined reference identity vector library, where f indicates the mapping relation between $\boldsymbol{z}_r$ and the password p. Each reference identity $\boldsymbol{z}_r$ is obtained by randomly selecting k different identities from the training set to combine, which aims to ensure that there is no real corresponding identity and avoid identity leakage. The new identity representation $\boldsymbol{z}_{new}$ after $\boldsymbol{z}_{id}$ rotation with the degree of θ on the hyperplane can be formulated as,

$$\boldsymbol{z}_{new} = \boldsymbol{z}_{id} \cos\theta + \boldsymbol{z}_{90} \sin\theta, \tag{7}$$

where $\boldsymbol{z}_{90}$ is the component vector decomposed from $\boldsymbol{z}_r$ and form a set of orthogonal bases with $\boldsymbol{z}_{id}$, which determines the direction of rotation and $\boldsymbol{z}_{new}$ may correspond to the identity of an unreal person. The function $\theta = g(d)$ is designed to control of the degree of identity variation with privacy level d. In recovery phase, we can calculate the original identity with the inverse operations and more detailed calculations will be introduced in Section 4.4.

Generator. The generator is required to implement image reconstruction based on $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{attr}$. Previous researches [1] have shown that simple embedding concatenation may result in relatively fuzzy results. To solve the problem, the novel *Adaptive Attentional Denormalization* (AAD) layers [13] have been proposed to improve feature integration in multiple levels. We employ cascaded n-AAD Residual Blocks in the generator to adjust attention regions of $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{attr}$ so that they can participate in synthesizing different parts.

4.2. Training Process

In training process, the identity encoder E_{id} is frozen while the others are trainable, where attributes encoder E_{attr} is trained to embed attributes representations disentangled from $\boldsymbol{z}_{id}$ and the generator G is trained to reconstruct the original image with $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{attr}$.

We use identity consistency loss $\mathcal{L}_{id}$ to make sure the identity of generated image $X' = G(\boldsymbol{z}_{id}, \boldsymbol{z}_{attr})$ still keeps the same.

$$\mathcal{L}_{id} = 1 - \frac{E_{id}(X') \cdot E_{id}(X)}{\|E_{id}(X')\|_2 \cdot \|E_{id}(X)\|_2} \tag{8}$$

We also define attributes consistency loss $\mathcal{L}_{attr}$ which can be formulated as,

$$\mathcal{L}_{attr} = \frac{1}{2} \sum_{k=1}^{n} \left\| \boldsymbol{z}^k_{attr}(X') - \boldsymbol{z}^k_{attr}(X) \right\|_2^2. \tag{9}$$

If the restored result X' is generated with the same $\boldsymbol{z}_{id}$ and $\boldsymbol{z}_{attr}$, it should be similar to the original image as possible. We set pixel-level $\mathcal{L}_2$ distance as reconstruction loss,

$$\mathcal{L}_{rec} = \frac{1}{2} \|X' - X\|_2^2. \tag{10}$$

We take advantage of adversarial learning to train the framework and introduce adversarial loss $\mathcal{L}_{adv}$ to constrain the generated results indistinguishable from real images. To promote the image quality, it is necessary to expand the perception range of the discriminator, so we adopt m multi-scale discriminators [22] with hinge loss functions for different resolution versions of the generated image.

$$\mathcal{L}_{adv}(X'_m, X_m) = \log\left(D(X_m)\right) + \log\left(1 - D(X'_m)\right), \tag{11}$$

where X_m indicates the low-resolution image after m-th downsampling.

The total loss function is the weighted sum of the above losses, which can be formulated as,

$$\mathcal{L}_{total} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_{attr}\mathcal{L}_{attr} + \lambda_{rec}\mathcal{L}_{rec}. \tag{12}$$

where λ_{adv}, λ_{id}, λ_{attr} and λ_{rec} are the tradeoff parameters.

4.3. Protection Process

In the protection phase, our approach takes the original image X, user-set password p and privacy level parameter d as input. The goal is to generate a specific de-identification image with p and d whose identity has been protected while other attributes remain the same.

For the original image X, we firstly get identity embedding $\boldsymbol{z}_{id} = E_{id}(X)$ and attributes embedding $\boldsymbol{z}_{attr} = E_{attr}(X)$. The de-identification identity representation $\boldsymbol{z}_{new} = M(\boldsymbol{z}_{id}, p, d)$ can be formulated as,

$$M(\boldsymbol{z}_{id}, p, d) = \bar{\boldsymbol{z}}_{id} \cdot \cos g(d) + \boldsymbol{z}_{90} \cdot \sin g(d), \tag{13}$$

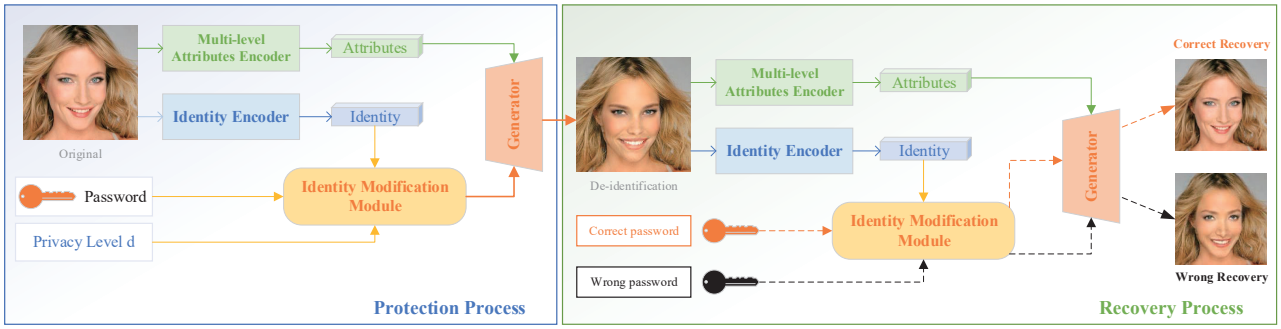


Figure 3: The framework of protection process and recovery process.

where

$$\boldsymbol{z}_{90} = f(p) - (\bar{\boldsymbol{z}}_{id} \cdot f(p)) \cdot \bar{\boldsymbol{z}}_{id}, \tag{14}$$

$\bar{\boldsymbol{z}}_{id}$ represents the normalized $\boldsymbol{z}_{id}$, and $f(p)$ is the reference identity corresponds to the password p.

Finally, we generate the de-identification result as,

$$\mathcal{F}(X, p, d) = G(\boldsymbol{z}_{new}, \boldsymbol{z}_{attr}). \tag{15}$$

4.4. Recovery Process

In the recovery phase, our approach can restore the de-identified image $\mathcal{F}(X, p, d)$ to the original image X only when the right password and privacy level are provided, which mainly differs from the protection process in the identity modification module M. For the de-identified image $\mathcal{F}(X, p, d)$, we extract $\hat{\boldsymbol{z}}_{new}$ and $\hat{\boldsymbol{z}}_{attr}$ with pre-trained encoders. The restored identity embedding $\hat{\boldsymbol{z}}_{id} = M^{-1}(\hat{\boldsymbol{z}}_{new}, p, d)$ can be calculated as,

$$M^{-1}(\hat{\boldsymbol{z}}_{new}, p, d) = \frac{\hat{\boldsymbol{z}}_{new} - f(p) \cdot \sin g(d)}{\cos g(d) - A \cdot \sin g(d)}, \tag{16}$$

where

$$A = \frac{\cos^2 g(d) - (\hat{\boldsymbol{z}}_{new} - f(p) \cdot \sin g(d)) \cdot \hat{\boldsymbol{z}}_{new}}{\sin g(d) \cdot \cos g(d)} \tag{17}$$

and $\hat{\boldsymbol{z}}_{new} = E_{id}(\mathcal{F}(X, p, d))$. The restored image $\hat{X}$ can be formulated as,

$$\hat{X} = G(\hat{\boldsymbol{z}}_{id}, \hat{\boldsymbol{z}}_{attr}). \tag{18}$$

5. Experiments

5.1. Implementation Details

Datasets. We train the network using CelebA-HQ [11] dataset, which is derived from CelebA [15] containing 30k upscale images of celebrity faces. Randomly choose 27k images for training while the others for test. Each image has been aligned and cropped to 256×256 covering the whole face region. In addition, we also test the generalization ability on FFHQ [12] and CASIA-WebFace [26].

Experimental Settings. We use the pre-trained ArcFace [4] as identity encoder E_{id}, and set the number of attributes representations $n = 8$ in Eq.(6). We train our network using Adam with $\beta_1 = 0$, $\beta_2 = 0.999$, and set the learning rate as 4×10^{-4}. The tradeoff parameters in Eq.(12) are set to $\lambda_{adv} = 0.1$, $\lambda_{id} = 5$ and $\lambda_{attr} = \lambda_{rec} = 10$. We define p as a six-digit password, each reference identity $\boldsymbol{z}_r$ is calculated by random $k = 10$ different identities and define $f(p)$ as one-to-one mapping. Based on testing on CelebA-HQ and considering both privacy protection effectiveness and image quality, and it cannot be restored when $\theta = 90°$, we define the relationship between θ and d as

$$g(d) = \begin{cases} 70 + d \times 5 & d \in [0, 4), \\ 70 + (d+1) \times 5 & d \in [4, 9]. \end{cases}$$

5.2. Evaluation Results

5.2.1 De-identification

Different Passwords. We evaluate the diversity of our approach by generating different de-identification results with different passwords. The qualitative results are shown in Fig. 4. It can be seen that our method can transform the identity into different identities in a large range which is determined with the password p.

Different Privacy Level. We evaluate the controllability by testing with different privacy levels d and present the qualitative results in Fig. 6. When d increases, the identity difference expands while the de-identified results can still share a similar appearance with the original in general, and most of them have successfully deceived the face verification model which we will provide the quantitative evaluation in the following part.

Quantitative Evaluation. We evaluate the performance of our approach from the perspectives of both privacy protection and image utility. Here we present the definition or explanation of the metrics we use.

(1) Privacy Protection: Almost all face verification models judge whether two images have the same identity by comparing identity embedding distance, so that we define

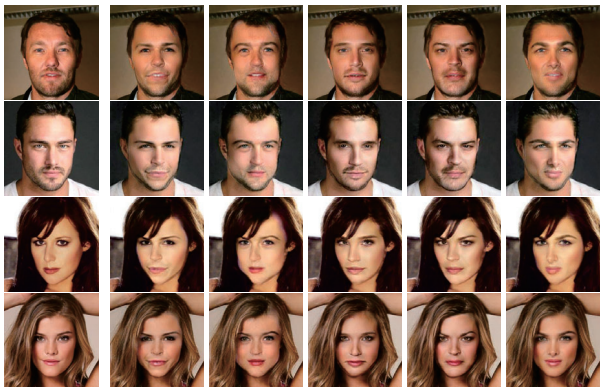


Figure 4: Various de-identification results. The leftmost column represents the original image and the last five columns present diverse de-identified results with different passwords. Particularly, the images of each column share the same password.

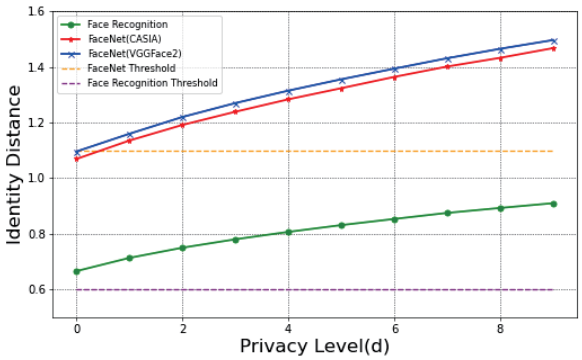


Figure 5: Identity Distance(*Id-dis*). Larger distance illustrates better de-identification effects. When identity distance exceeds the threshold, the face verification model believes the identity has been varied.

identity distance (*Id-dis*) and **successful protection rate (*SR*)** for protection effects evaluation. Id-dis indicates the distance between identity-vectors e_{id} extracted from the face recognition model, which can be formulated as

$$\textit{Id-dis} = D(e_{id}(X), e_{id}(\mathcal{F}(X, p, d)). \tag{19}$$

SR means the proportion of successful de-identification as

$$SR = 1 - \frac{1}{N}\sum_{i=1}^{N} f_{ver}(X, \mathcal{F}(X, p, d)), \tag{20}$$

when Id-dis $> \tau$, it considers two identities different as $f_{ver} = 0$ and otherwise $f_{ver} = 1$, N is the number of testing. We respectively use the Face Recognition library, FaceNet trained on CASIA and FaceNet trained on VGGFace2 for evaluation where the specific forms of D are all Euclidean distance.

	Face Recognition	FaceNet (CASIA)	FaceNet (VGGFace2)
DeepPrivacy [10]	0.74623 / 0.939	1.19684 / 0.734	1.22889 / 0.816
Gu *et al.* [8]	**0.82234** / 0.961	1.14419 / 0.704	1.16245 / 0.695
Ours	0.79195 / **0.975**	**1.24421** / **0.913**	**1.27270** / **0.928**

Table 1: Privacy evaluation of de-identification results, where the values in the table indicate identity distance and successful de-identified rate *Id-dis / SR*. We choose the threshold of Face Recognition Library as $\tau = 0.6$ and the threshold of FaceNet as $\tau = 1.1$ according to [20].

	DR↑	*Pixel-dis*↓				
		Face	Landmarks	Eyes	Nose	Mouth
DeepPrivacy [10]	**1.0**	5.005	2.506	1.502	1.799	3.288
Gu *et al.* [8]	0.8585	0.925	2.346	1.810	1.906	2.139
Ours	0.9973	**0.225**	**1.969**	**1.236**	**1.546**	**1.900**

Table 2: Utility evaluation of de-identification results. The face region is detected with *OpenCV* and landmarks are detected with *dlib*.

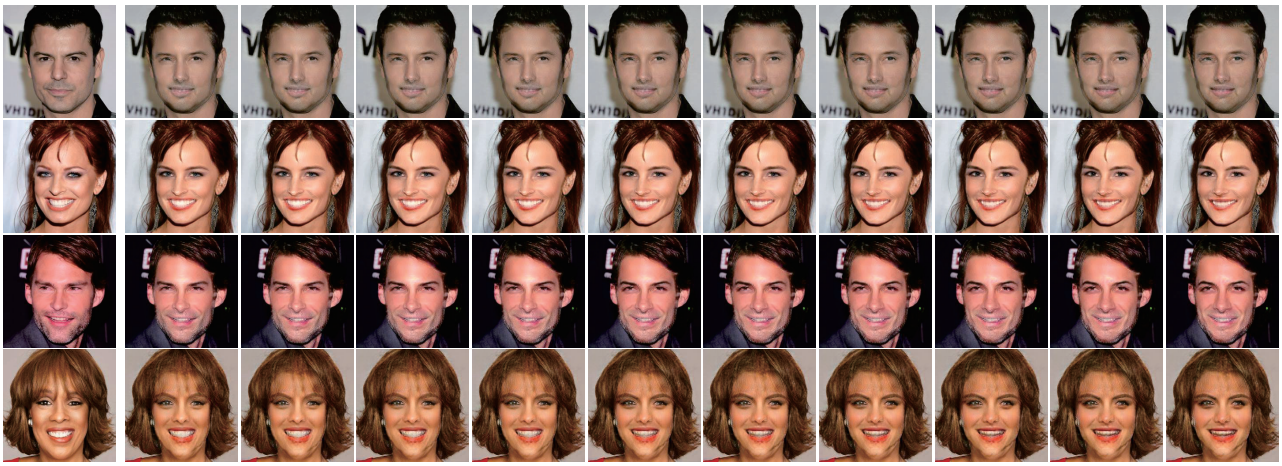
(2) Image Utility: We define the rate of the face in de-identified images can be detected as **face detectability (*DR*)** in Eq.(21) to measure the utility for computer vision tasks.

$$DR = \frac{1}{N}\sum_{i=1}^{N} f_{det}(\mathcal{F}(X, p, d)), \tag{21}$$

if the face can be detected, $f_{det} = 1$ and otherwise $f_{det} = 0$. We also detect face region and landmarks to calculate the **pixel-level distance (*pixel-dis*)** from the original image.

We randomly select several images from CelebA-HQ and de-identify them with random passwords p and privacy levels d. The privacy evaluation compared with DeepPrivacy [10] and Gu *et al.* [8] represents in Table 1, which can be concluded that our method is more effective for identity protection with both larger identity distance and higher successful rate. We also generate the de-identification results using random passwords with each privacy level, and the variation of identity distance with d is shown in Fig. 5.

In Table 2, we apply computer vision algorithms on the de-identified images and compare the difference of pixel-level in face region, landmarks, eyes, nose and mouth between the de-identification results and the original, as well as the detection rate of the de-identified. *Landmarks* indicates the mean distance of the total 68 keypoints while *Eyes/Nose/Mouth* represents that of keypoints corresponding to each facial area. The utility evaluation proves that our method can guarantee the consistency of the face region and landmarks better, and most de-identified faces can be detected, which proves that it guarantees better utility for identity-agnostic computer vision tasks. We also show the tradeoff between privacy and utility in Fig. 7. Increasing the level of privacy protection will increase the pixel difference, which means the utility of the image will be reduced.



X (a) $d = 0$ (b) $d = 1$ (c) $d = 2$ (d) $d = 3$ (e) $d = 4$ (f) $d = 5$ (g) $d = 6$ (h) $d = 7$ (i) $d = 8$ (j) $d = 9$

Figure 6: The leftmost column represents the original image and the rest indicate the de-identified results with different privacy level (From left to right, the privacy level parameter d increases form 0 to 9).

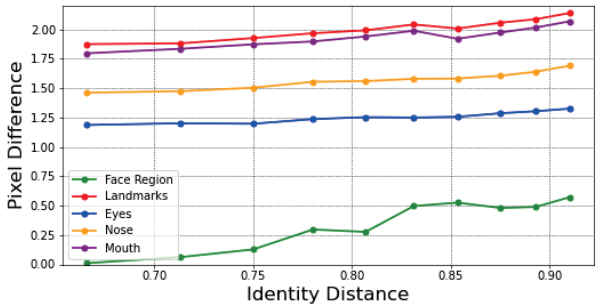


Figure 7: The tradeoff between privacy and utility of the de-identified results. The abscissa represents the identity distance measured by the Face Recognition library, and the ordinate is the pixel difference of face region and keypoints.

	Face Recognition	FaceNet (CASIA)	FaceNet (VGGFace2)
Incorrect Recovery	0.794 / 0.904	1.243 / 0.854	1.257 / 0.879
Correct Recovery	0.228 / 0.035	0.368 / 0.035	0.401 / 0.035

Table 3: Id-dis/SR evaluation for incorrect/correct recovery.

5.2.2 Recovery

The restored results with correct or wrong passwords are presented in Fig. 8. When the attacker tries to recover the de-identified image with wrong passwords, he can still get a good-quality face image but cannot obtain the original identity information, which may confuse him and achieve a more reliable protection.

While our framework is trained on CelebA-HQ, the generalization results tested on FFHQ and CASIA-WebFace are shown in Fig. 10, and it comes to the conclusion that our approach can apply to a wider images range. In order

(a)X (b)p_1 (c)p_2 (d)$p_2|p_1$ (e)$p_1|p_2$ (f)$\hat{X}$

Figure 8: Recovery Results. X: the original image, $p_{1,2}$: two different de-identified results, $p_m|p_n$: use p_n to restore the image de-identified with p_m, $\hat{X}$: the correct recovery.

	LPIPS↓	PSNR↑	SSIM↑	MAE↓
Blur	0.242	**28.396**	0.802	**0.026**
Pixelation	0.447	23.159	0.671	0.040
Noise	0.264	22.163	0.701	0.046
Gu *et al.*	0.186	27.602	0.827	0.029
Ours	**0.062**	27.501	**0.902**	0.031

Table 4: Comparison of the restored image quality.

to keep consistent with the model input, we first convert all test images to the size of 256×256 before feeding the model. The small artifacts are considered due to image distortion caused by interpolation or misalignment.

We compare de-identification results, wrong recovery and correct recovery with [8] on both CelebA-HQ and CA-

Figure 9: Compare results with Gu *et al*. [8]. The left are from CelebA-HQ while the right are from CASIA-WebFace. For the same input image, the upper row is our results, and the lower row is the results generated by [8].

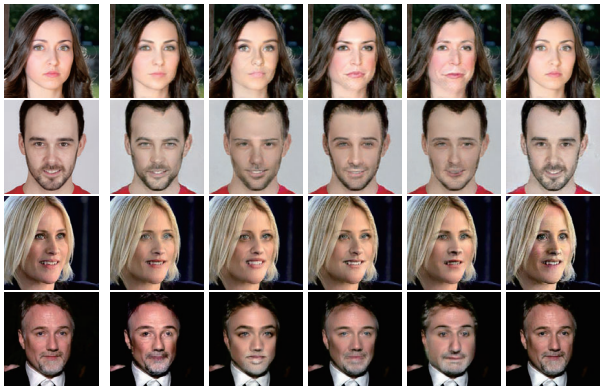


Figure 10: FFHQ and CASIA generalization results with the model trained on CelebA-HQ. The upper two lines are from FFHQ while the lower are CASIA. X: original image, $p_{1,2,3}$: the de-identified results with three different passwords, $\hat{Y}$: wrong recovery, $\hat{X}$: correct recovery.

SIA shown in Fig. 9, which shows our de-identified results can retain more similarity with the original. Identity evaluation of incorrect and correct recovery are shown in Table 3, where the recovery is effective when using correct password while wrong passwords will generate a different identity with a high probability. We evaluate the recovery quality in Table 4 using LPIPS (Learned perceptual image patch similarity) [27] distance to measure perceptual similarity, PSNR (Peak signal-to-noise ratio) and MAE (Mean absolute error) to measure distortion at the pixel level, and SSIM (Structural similarity) to measure the structure similarity. We compare to three traditional methods and Gu *et al*. [8] as baselines, where we deblur by Wiener filter, remove pixelation by bilinear interpolation, and de-noise by non-local averaging. Based on the comparison, the restored images obtained by our method are closest to the original with high image quality.

6. Conclusion

In this paper, we propose a personalized and invertible de-identification method for privacy preservation. Our method first disentangles the representations of identity and attributes, encrypts or restores identity with latent space manipulation based on the password and the privacy level parameter, and finally reconstructs the de-identified or recovery image. In the protection phase, our approach can generate personalized de-identification results with different passwords and control the identity distance from the original by the privacy level parameter. In the recovery phase, our approach can restore if and only if the corresponding password is given, while the image with another identity will be generated when the attacker tries the wrong passwords. Experiments demonstrate the satisfactory performance in privacy protection and image utility of the de-identified results, as well as the quality of the restored, compared with the traditional or state-of-the-art methods. Generalizing the proposed framework to handle face images of different resolutions and different poses is part of our future work. Besides, the de-identification in videos is also a problem worthy of research.

Acknowledgements This work was supported by MoE-China Mobile Research Fund Project (MCM20180702), National Key R&D Project of China (2019YFB1802701), the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018.

[2] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. Vgan-based image representation learning for privacy-preserving facial expression recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[3] Durkhyun Cho, Jin Han Lee, and Il Hong Suh. Cleanir: Controllable attribute-preserving natural identity remover. *Applied Sciences*, 10:1120, 2020.

[4] 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 (CVPR)*, 2019.

[5] Maoguo Gong, Jialu Liu, Hao Li, Yu Xie, and Zedong Tang. Disentangled representation learning for multiple attributes preserving face deidentification. *IEEE transactions on neural networks and learning systems*, 2020.

[6] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating utility into face de-identification. In *PET'05 Proceedings of the 5th international conference on Privacy Enhancing Technologies*, pages 227–242, 2005.

[7] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 161–161, 2006.

[8] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European Conference on Computer Vision*, 2020.

[9] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. Disentangling factors of variation by mixing them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[12] 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 (CVPR)*, 2019.

[13] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[14] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

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

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

[18] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG)*, 39:1 – 14, 2020.

[19] Hugo Proença. The uu-net: Reversible face de-identification for visual surveillance video footage. *ArXiv*, abs/2007.04316, 2020.

[20] 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 (CVPR)*, 2015.

[21] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.

[22] Wei Tang, Gui Li, Xinyuan Bao, and Teng Li. Mscgan: Multi-scale conditional generative adversarial networks for person image generation. *2020 Chinese Control And Decision Conference (CCDC)*, pages 1440–1445, 2020.

[23] Nishant Vishwamitra, Bart Knijnenburg, Hongxin Hu, Yifang P Kelly Caine, et al. Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 39–47, 2017.

[24] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[25] Mehmet Yamac, Mete Ahishali, Nikolaos Passalis, Jenni Raitoharju, Bulent Sankur, and Moncef Gabbouj. Reversible Privacy Preservation using Multi-level Encryption and Compressive Sensing. *arXiv e-prints*, page arXiv:1906.08713, June 2019.

[26] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *ArXiv*, abs/1411.7923, 2014.

[27] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.