FaceRSA: RSA-Aware Facial Identity Cryptography Framework

Zhongyi Zhang, Tianyi Wei*, Wenbo Zhou*,
Hanqing Zhao, Weiming Zhang, Nenghai Yu
University of Science and Technology of China
{ericzhang@mail., bestwty@mail., welbeckz@, zhq2015@mail., zhangwm@, ynh@}ustc.edu.cn

Abstract

With the flourishing of the Internet, sharing one's photos or automated processing of faces using computer vision technology has become an everyday occurrence. While enjoying the convenience, the concern for identity privacy is also emerging. Therefore, some efforts introduced the concept of "password" from traditional cryptography such as RSA into the face anonymization and deanonymization task to protect the facial identity without compromising the usability of the face image. However, these methods either suffer from the poor visual quality of the synthesis results or do not possess the full cryptographic properties, resulting in compromised security. In this paper, we present the first facial identity cryptography framework with full properties analogous to RSA. Our framework leverages the powerful generative capabilities of StyleGAN to achieve megapixel-level facial identity anonymization and deanonymization. Thanks to the great semantic decoupling of StyleGAN's latent space, the identity encryption and decryption process are performed in latent space by a well-designed password mapper in the manner of editing latent code. Meanwhile, the password-related information is imperceptibly hidden in the edited latent code owing to the redundant nature of the latent space. To make our cryptographic framework possesses all the properties analogous to RSA, we propose three types of loss functions: single anonymization loss, sequential anonymization loss, and associated anonymization loss. Extensive experiments and ablation analyses demonstrate the superiority of our method in terms of the quality of synthesis results, identity-irrelevant attributes preservation, deanonymization accuracy, and completeness of properties analogous to RSA.

Introduction

In today's world, privacy has become a crucial concern, especially with regard to facial identity information. However, many computer vision tasks require uploading photos or videos, which may compromise user privacy. For instance, family cameras are used to monitor the behavior of infants and young children, but an attacker shouldn't gain access to facial identity information. At the same time, trusted users, such as family members, may require access to the original images. This presents a challenge since we need to design a special cryptographic system without affecting the use of facial images for other computer vision tasks.

Some traditional anonymization methods simply use pixel-level operations, such as downsampling, blurring, and masking. These methods will greatly impair the quality and usability of the image or could be easily reverted with advances in deep learning techniques. Recently, some approaches (Maximov, Elezi, and Leal-Taixé 2020; Hukkelås, Mester, and Lindseth 2019) have utilized generative networks to produce photo-realistic anonymized images, but these methods pay no attention to recovering the original images. Inspired by RSA (Rivest, Shamir, and Adleman 1978), a popular public-private key encryption method that offers provable security and computational privacy protection through mathematical complexity, Gu et al. (Gu et al. 2020) introduced the concept of password and reverse password into face anonymization and deanonymization. However, their approach has several limitations such as low-quality synthesis results and incomplete RSA properties. RiDDLE (Li et al. 2023) also proposes a face anonymization framework. However, their framework relies on the same password for both anonymization and deanonymization thus posing a risk of password leakage. Additionally, their method struggles to preserve identity-irrelevant information in the original image.

To address the aforementioned limitations, we introduce a novel method, FaceRSA, which is the first facial identity cryptography framework with full properties analogous to RSA. It is important to note that *our framework only provides RSA-aware properties in the domain of human faces and does not offer any security guarantees based on mathematical complexity like traditional cryptographic systems.* The security of our system is based on ambiguity, i.e., an attacker could not tell whether an image has been anonymized, rather than the security based on computational complexity like traditional cryptography. We also need to clarify that the passwords we use for face anonymization and deanonymization are not RSA keys.

Different from directly using conditional GAN (Chen et al. 2016) for anonymization at image level, we use the powerful representation ability of the pre-trained StyleGAN (Karras, Laine, and Aila 2019) by inverting the real image into the StyleGAN latent space through GAN inversion methods to realize face anonymization and de-

*Corresponding authors: Wenbo Zhou, Tianyi Wei anonymization. This design is based on the powerful representational ability of StyleGAN and the redundant nature of its latent space. We further refine the scheme of using a password to control the anonymization and deanonymization process to meet the properties of RSA, specifically: 1) *Locating the Identity-relevant Layers.* The different latent space layers of StyleGAN correspond to different levels of semantics from coarse to fine. We locate the latent layers that are most relevant to identity by examining the corresponding semantics of different layers, thus ensuring preserving identity-irrelevant attributes through anonymization and deanonymization. 2) *Password Converter.* To better control the anonymization process, we use a password converter to convert the discrete password into a 512-dimensional password vector to align the dimensions of the StyleGAN latent space. 3) *Modulation Model.* We use a modulation model so that the converted password vector can be used to explicitly control the change of the latent code, which enhances the controllability of our framework.

In order to ensure that our framework can possess all the properties analogous to RSA, we designed three types of loss functions: 1) Single anonymization loss is used to control the anonymization process under a single pair of encryption and decryption passwords. 2) Sequential anonymization loss is used to implement some extensive anonymization and deanonymization requirements when multiple pairs of encryption and decryption passwords are utilized. 3) Associated anonymization loss is used to ensure the existence of the equivalent password for both encryption and decryption processes.

Our framework has been evaluated through qualitative and quantitative experiments, demonstrating its superiority in terms of the quality of the synthesis images, preservation of identity-irrelevant information, deanonymization accuracy and properties analogous to RSA. We also conduct some extensive ablation experiments in the supplementary material to verify the effectiveness of our framework and loss function design.

Overall, the contributions of our method consist of the following:

- Our proposed framework is the first facial identity cryptography framework with full properties analogous to RSA, which supports megapixel-level facial identity anonymization and deanonymization.
- We choose to build such a facial identity cryptography system with the help of StyleGAN's latent space by proposing a mechanism to locate identity-related layers, designing the password mapper, and customizing three types of training losses.
- Extensive experiments and ablation studies are conducted to show the superiority of our method and the necessity of each new design.

Related Works

Face Anonymization

Simple human face anonymization methods using blurring, noise, masking, etc. on the face region may greatly destroy the usability of the image, so some work is devoted to the anonymization of a face image without compromising image quality. DeepPrivacy (Hukkelås, Mester, and Lindseth 2019) proposed an inpainting-based method to realize face de-identification, Li et al. (Li et al. 2021) found identity-aware face regions to remove original identity while keeping other attributes, IdentityDP (Wen et al. 2022) introduced the concept of differential privacy into de-identification to achieve measurable anonymization. However, these studies only focused on removing the identity of the original image and did not take into account for the possibility of recovering the original image.

Recent work has taken this deficiency into account and proposed methods that can recover the original image. For instance, FIT (Gu et al. 2020) introduced a discrete password-based method that controls the generated anonymized image and ensured that an attacker using a wrong password can only obtain a wrong but photo-realistic image. Cao et al. (Cao et al. 2021) separated identity and attribute information, and realize anonymization and deanonymization through controllable rotation of identity vector. Concurrent with our work, RiDDLE (Li et al. 2023) used a transformer structure to anonymize the image by a randomly sampled latent code.

Although these works consider recovering the original image, they may fail in complex anonymization scenarios e.g. deanonymizing the image that has been anonymized multiple times, which could impact their practical application as well as make the anonymized image distinguishable thus compromising its security. In contrast, our proposed FaceRSA builds a cryptosystem with full properties analogous to RSA, thereby overcoming these limitations. For a comparison of our work and some prior work, see Table 1.

Image Manipulation Based on StyleGAN

StyleGAN (Karras, Laine, and Aila 2019; Karras et al. 2020) is a powerful generative network that can generate high-resolution images on various data domains. Surprisingly, its latent space exhibits promising disentanglement properties (Collins et al. 2020; Jahanian, Chai, and Isola 2019; Abdal et al. 2021). As a result, many works (Patashnik et al. 2021; Wei et al. 2022a; Jiang et al. 2021; Sun et al. 2022; Wei et al. 2023) have utilized StyleGAN for various image manipulation tasks. StyleCLIP (Patashnik et al. 2021) realized text-guided image manipulation with the help of CLIP's powerful image-text representation capability (Radford et al. 2021). HairCLIP (Wei et al. 2022a) introduced a modulation module to achieve direct control of the hair condition input over the latent code. In this paper, we implement password-based identity manipulation with the powerful generative ability of StyleGAN and the great semantic decoupling of its latent space, which shares the same philosophy as the previously mentioned methods.

Preliminaries

RSA is a widely used encryption and decryption algorithm and has many excellent properties. As our approach is an RSA-aware system, the following properties should be sat-

Method	DeepPrivacy	IdentityDP	Li et al.	FIT	Cao et al.	RiDDLE	Ours
Face Anonymization	✓	✓	✓	✓	✓	✓	✓
Face Deanonymization	✗	✗	✗	✓	✓	✓	✓
Sequential Anonymization and Deanonymization	✗	✗	✗	✗	✗	✗	✓
Password Interchangeability	✗	✗	✗	✗	✗	✗	✓
Password Associativity	✗	✗	✗	✗	✗	✗	✓

Table 1: Comparisons between our approach and mainstream face anonymization methods in terms of functionality. Only our method supports all kind of scenarios.

isfied. Here I_* denotes using cryptography algorithm on image I, when multiple passwords exist in $*$, it means use cryptography algorithm with them in order, i.e $I_{e,d}$ means first use anonymization algorithm with password e on I and then use de-anonymization algorithm with password d on I_e. (e_i, d_i) represents the i-th pair of encryption and decryption passwords.

(a)Photo-realism. We hope that the cryptography algorithm will still generate a real-looking face, so an attacker cannot distinguish whether the face is anonymized or not, and it will not affect the usage of this image for computer vision tasks. Let Φ be the manifold of human faces, this property could be formed as:

$$\forall I \in \Phi, \forall e, \ s.t. \ I_e \in \Phi$$

(b)Anonymized with password. Let f represents the function that maps the facial image to the identity, the anonymization progress via encryption password e could be formalized as:

$$f(I_e) \neq f(I)$$

(c)Deanonymized with correct password. The original identity of I could be recovered with the correct decryption password d.

$$f(I_{e,d}) = f(I)$$

(d)Wrong deanonymized with wrong password. When a wrong password d' is given to deanonymize the image, the system will generate a new identity that is different from both the original image and the anonymized image.

$$f(I_{e,d'}) \neq f(I_e)$$
$$f(I_{e,d'}) \neq f(I) \quad where \quad d' \neq d \; and \; I_{e,d'} \in \Phi$$

(e)Diversity. Different identities should be generated when using different encryption passwords on a single image.

$$f(I_{e_1}) \neq f(I_{e_2}) \quad where \quad e_1 \neq e_2$$

(f)Cycle anonymized and deanonymized with paired passwords. When the anonymization operation is applied to the same image multiple times, the identity of the original image will be obtained by deanonymizing in the order of anonymization. At the same time, the identities of each pair of intermediate images should also remain the same. For properties (f), (g), and (h), we use two pairs of encryption and decryption passwords to illustrate the corresponding effect.

$$f(I_{e_1,e_2,d_2}) = f(I_{e_1})$$
$$f(I_{e_1,e_2,d_2,d_1}) = f(I)$$

(g)Passwords interchangeability when deanonymization. When deanonymizing an image that has been anonymized multiple times, the identity of the original image will be recovered regardless of the deanonymization order. Meanwhile, paired anonymization and deanonymization operations are eliminated as if they were never performed.

$$f(I_{e_1,e_2,d_1}) = f(I_{e_2})$$
$$f(I_{e_1,e_2,d_1,d_2}) = f(I)$$

(h)Passwords associativity. In multi-step anonymization and deanonymization operations, multi-step operations performed by multiple passwords can be equivalent to one operation performed by an associated password. Here '$+$' means the summation of different passwords.

$$f(I_{e_1,e_2}) = f(I_{e_1+e_2})$$

Method

Overview

Our purpose is to design an RSA-aware cryptography system based on passwords without compromising image quality. As mentioned in E2Style (Wei et al. 2022b), the negligible effect of rounding the floating-point latent code suggested that the StyleGAN latent space contains a high degree of redundancy. Based on this property, instead of using conditional GAN at image level, we proposed using the StyleGAN latent space to embed password information, realizing image anonymization and deanonymization through latent code manipulation.

Next, we design some modules and mechanisms to better control the editing of latent code. We first locate the identity-relevant layers in StyleGAN to minimize the impact on non-identity attributes. Then we use a password converter to convert password into a vector, which will be used in a modulation model to modulate the latent code. To enhance the security of the anonymized image and align with the full properties of RSA, it is essential to impose constraints not only on the single pair but also on scenarios involving multiple pairs of encryption and decryption passwords, so we introduce three types of loss functions: single anonymization loss, sequential anonymization loss and associated anonymization loss.

Before introducing the specific framework and loss functions, we briefly introduce the latent space of StyleGAN. The image synthesis process of StyleGAN involves its multiple latent spaces. It first randomly samples a vector $z \in$

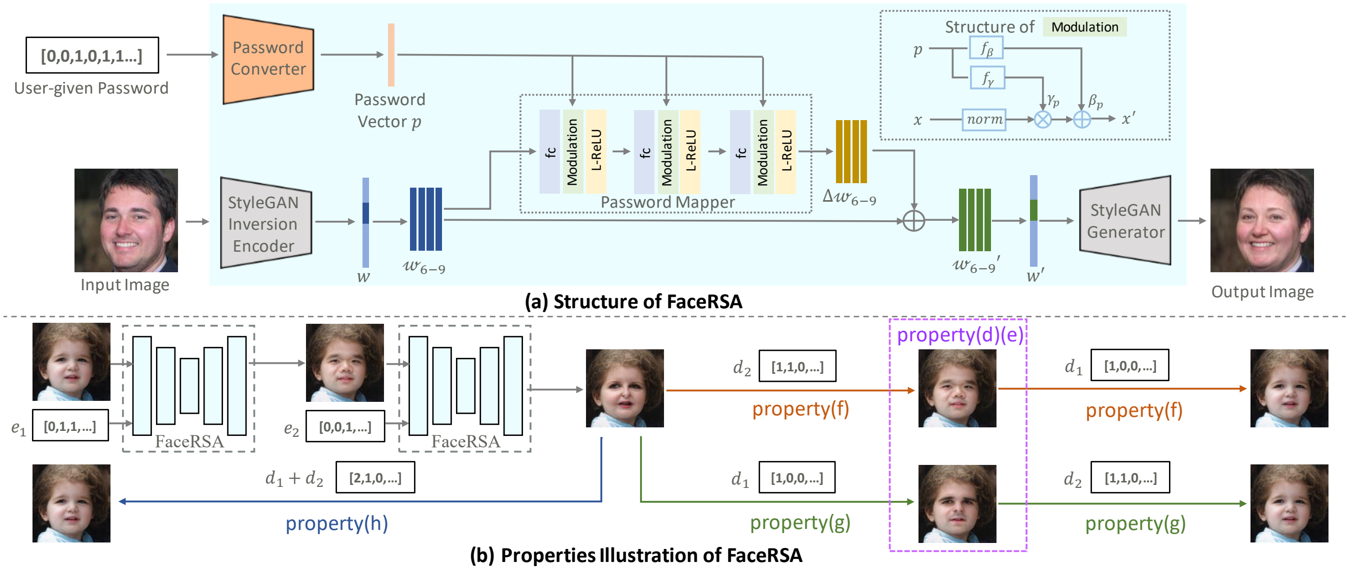

Figure 1: (a)The overall pipeline of our method, models with gray color are frozen. (b)Properties demonstration of our framework in the case of two pairs of encryption and decryption passwords. Property (a)-(c) can be directly observed from the figure, and we distinguish properties (d)-(h) with different colored lines and boxes. Our model satisfies the RSA properties mentioned in Preliminaries Section. For the sake of brevity, we only show the complete input and output with our FaceRSA framework twice, and the rest of the FaceRSA framework is omitted.

R^{512} and transforms it to a style code $w \in R^{512}$ after 8 fully connected layers. This space is called W space, and many studies have shown that it is rich in semantic information. Some studies (Collins et al. 2020; Goetschalckx et al. 2019; Jahanian, Chai, and Isola 2019) have extended the W space to $W+$ space, which consists of different w vectors corresponding to different layers of the StyleGAN structure. These different layers of $W+$ Space controls various semantic information from coarse to fine. In the case of the 18-layer StyleGAN network, the vector in the $W+$ space can be expressed as: $w = [w_1, w_2, \cdots, w_{18}]$.

FaceRSA

Thanks to the powerful synthesis and semantic decoupling capabilities of StyleGAN, we choose to accomplish the identity transformations in the StyleGAN latent space. Specifically, to anonymize a real image, we first obtain its latent code w in the $W+$ space using e4e (Tov et al. 2021), which is an encoder-based GAN inversion method with better editing capability. Then, we utilize the well-designed mapping network to predict the bias Δw in latent space based on the user-given password, a N-bit binary code. The modified latent code, $w' = w + \Delta w$, is subsequently fed back into the pre-trained StyleGAN to obtain the output result. The overall pipeline is illustrated in Figure 1.

Locating the Identity-relevant Layers. As observed by many researches (Xia et al. 2021; Yang, Shen, and Zhou 2021; Patashnik et al. 2021), various layers in the latent space of StyleGAN correspond to different semantic features. In order to preserve identity-irrelevant attributes of an image, such as pose and expression, we localized the identity-relevant layers by modifying the original latent code step by step. Finally, we opted to change only specific layers - detailly, layers 6-9 in the $W+$ latent space - while keeping all other layers unchanged. Specific implementation details and ablation experiments for this setting are provided in the supplementary material. By adopting this approach, we can change identity information while minimizing the effect on other image attributes.

Password Converter. To better utilize the password for controlling the mapping of latent codes, we convert the N-bit discrete password into a 512-dimensional real password vector using a password converter, a simple 2-layer MLP. This is done to align the dimensions of the StyleGAN latent space and facilitate the use of the modulation module in subsequent steps.

Modulation Model. To implement identity transformations that are controlled through passwords, we propose the usage of a modulation module. This module enables the password vectors to explicitly control the changes of latent code. We borrow the structure of modulation model used in HairCLIP (Wei et al. 2022a), which has the following form:

$$x' = \gamma_p \left(\frac{x - \mu_x}{\sigma_x} \right) + \beta_p$$

where μ_x and σ_x denote the mean and standard deviation of x. γ_p and β_p are calculated from the password vector p with simple fully connected networks.

Loss Functions

Our objective is to enable adaptation to changes in image identity information across various scenarios while preserving other identity-irrelevant information. Here all the original images mentioned below refer to the inverted images.

Our total loss function for different cryptography scenarios comprises three parts: single anonymization loss, sequential anonymization loss, and associated anonymization loss.

Single Anonymization Loss $\mathcal{L}_{single}$. The main purpose of this part of the loss function is to constrain the function of the cryptography system in the case of only single encryption and decryption operation is used. We consider the simplest case of this scenario: two different encryption passwords e_1 and e_2, one correct decryption password d_1 and one wrong decryption password d_1'.

According to the RSA properties (a)-(e) mentioned in Preliminaries Section, to realize the anonymization and deanonymization function, we introduce facial identity difference loss on image pairs (I, I_{e_1}),(I, I_{e_2}),$(I, I_{e_1,d_1'})$,(I_{e_1}, I_{e_2}) with the formulation $\mathcal{L}_{change} = cos(\mathcal{F}(I_1), \mathcal{F}(I_2))$ where $cos(\cdot)$ denotes the cosine similarity and $\mathcal{F}$ is a pre-trained Arcface (Deng et al. 2019) network to extract facial identity embeddings. I_1 and I_2 represent the different images in the image pairs.

Also, to ensure the original image is recovered correctly, pixel L_2 loss $\mathcal{L}_{pix} = \|I - I_{e_1,d_1}\|_2$ and cosine identity similarity loss $\mathcal{L}_{recon} = 1 - cos(\mathcal{F}(I), \mathcal{F}(I_{e_1,d_1}))$ are used on image pair (I, I_{e_1,d_1}).

During the training process, the generated latent codes must be constrained to remain on the well-defined manifold of the StyleGAN to prevent artifacts. Therefore, for each generated latent code w_*, we introduce a regularization loss $\mathcal{L}_{reg} = \|w_* - w\|_2$ where w denotes the inverted latent code of the original image I.

Additionally, to minimize the impact on downstream tasks and prevent changes in attributes such as expressions while changing the identity, we also use face parsing loss $\mathcal{L}_{parsing}$ as mentioned in E2Style (Wei et al. 2022b) and facial landmark loss $\mathcal{L}_{lmk}$ for all generated images I_*. LPIPS loss (Zhang et al. 2018) is also applied to all generated images I_* to preserve image quality and improve similarity in feature level.

Sequential Anonymization Loss $\mathcal{L}_{seq}$. When facing the situation of sequential anonymization and deanonymization, we define an image sequence as $\left\{I^k\right\}_{k=0}^{2m}$ with m-pairs of encryption and decryption passwords. Here image I^0 denotes the original image and each image I^n in the sequence is generated by applying cryptography algorithm on image I^{n-1} with a single password from the sequence $e_1, \cdots, e_m, d_m, \cdots, d_1$ in order. It is crucial to ensure that the identity information is maintained between the corresponding intermediate image pairs $\left\{(I^k, I^{2m-k})\right\}_{k=0}^{m-1}$. We also add loss functions to ensure the image quality and the identity diversity of each anonymized image $\left\{I^k\right\}_{k=1}^{m}$.

Associated Anonymization Loss $\mathcal{L}_{asso}$. In this context, we require that multiple anonymization operations can be regarded as obtaining an equivalent password through one anonymization operation. We randomly selected images I^{i-1} and I^j in the sequence $\left\{I^k\right\}_{k=0}^{m}$ where $j > i$, and the equivalent password is expressed as the sum of intermediate consecutive encryption passwords, that is $\tilde{e} = \sum_{k=i}^{j} e_k$. Associated identity loss $\mathcal{L}_{asso-id}$ and pixel L_2 loss $\mathcal{L}_{asso-pix}$ are computed between the image $(I^{i-1})_{\tilde{e}}$ and I^j to satisfy the associated anonymization property. Note that although this loss is only imposed on anonymization process, it will be shown in the experiment section that our framework can generalize associated password to de-anonymization process.

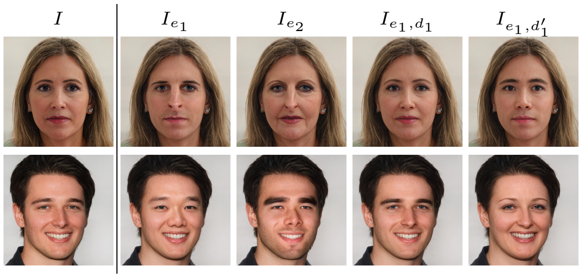

Figure 2: Qualitative result for single encryption and decryption password pair of our framework. I refers to the original image, (e_1, d_1) is a pair of encryption and decryption password and $d_1' \neq d_1, e_1 \neq e_2$, each column shares the same password. Our framework satisfies the basic anonymization and deanonymization requirements.

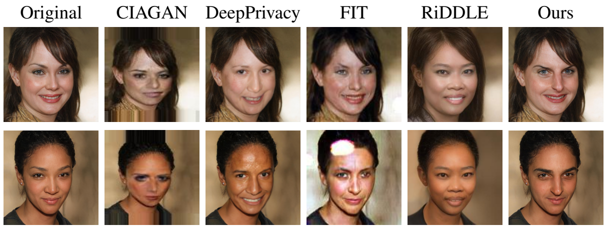

Figure 3: Qualitative comparison of anonymization ability. Our method shows the best quality and preserves the most identity-irrelevant attributes, which is fully compliant with the requirements of anonymization.

The detailed designs and the hyperparameter settings of $\mathcal{L}_{single}$, $\mathcal{L}_{seq}$ and $\mathcal{L}_{asso}$ are presented in the supplementary material.

Finally, the total loss of our training process is

$$\mathcal{L}_{total} = \mathcal{L}_{single} + \mathcal{L}_{seq} + \mathcal{L}_{asso}$$

Experiment

Implementation details of our approach are provided in the supplementary material. For all compared methods, we use the official pre-trained models.

We first show in Figure 2 the effect of our entire system while only a pair of encryption and decryption password is used. As we can see, using different encryption passwords for a single image result in different anonymized images, whereas using the same encryption password for different images also generates different anonymization images. Additionally, the original image can be correctly recovered with the correct decryption password, while using an incorrect decryption password will produce a new anonymized image

	ID↓	Detect↑	Lmk↓	Pose↑	Exp↑
CIAGAN	0.150	0.950	1.208	0.930	0.283
DeepPrivacy	0.118	0.998	0.023	0.966	0.289
FIT	0.296	0.992	0.011	0.971	0.341
RiDDLE	**0.115**	0.999	0.013	0.956	0.314
Ours	0.247	**1.000**	**0.009**	**0.996**	**0.546**

Table 2: Quantitative comparison of face anonymization ability. Our method outperforms others in most metrics except identity similarity since our method preserves more identity-irrelevant attributes, which will undesirably increase the identity similarity measured by Arcface.

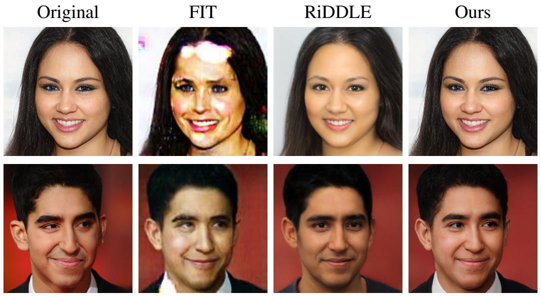

Figure 4: Qualitative comparison of recovery ability. Our method shows the best recovery quality compared with the existing methods, thus having better application potential.

that differs from the original identity. Figure 2 shows that our framework satisfies the properties (a)-(e) mentioned in Preliminaries Section which related to the scenario of single pair of encryption and decryption password.

We compare our method with two password-aware anonymization methods, FIT (Gu et al. 2020) and RiDDLE (Li et al. 2023), one identity vector controlled conditional GAN based method CIAGAN (Maximov, Elezi, and Leal-Taixé 2020), and one inpainting-based method DeepPrivacy (Hukkelås, Mester, and Lindseth 2019). For the specific implementation details, RiDDLE and our method are based on GAN inversion while others are based on conditional GAN. In our experiment, we ignore the minor distortion caused by GAN inversion, and the test images are after inversion. We separately compare anonymization ability with each method and deanonymization ability with FIT and RiDDLE, since only FIT and RiDDLE enable the recovery of the original image. Here we use e_i to represent the i-th encryption password and d_i to represent the i-th decryption password, but the relationships between the encryption and decryption passwords used by each methods are actually different.

Anonymization Ability. Qualitative result of anonymization could be seen in Figure 3. We observe that CIAGAN suffers from severe image distortion, which affects the usability of the image. FIT always generates an anonymized image with some white dots and changes the lighting of the overall image. DeepPrivacy successfully generates photo-realistic images, but sometimes modify the expression of the original image. RiDDLE also utilizes StyleGAN2 as a generator, but remains less identity-irrelevant information, such as pose and hairstyle since they use all of the latent code to transform identity. Our method, on the contrary, remains more identity-irrelevant information thanks to changing only some specific layers related to identity.

	LPIPS↓	MSE↓	PSNR↑	SSIM↑	ID↑
FIT	0.1780	0.0064	61.21	0.9928	0.6767
RiDDLE	0.0324	0.0124	67.61	0.9987	0.8315
Ours	**0.0099**	**0.0019**	**75.51**	**0.9996**	**0.9206**

Table 3: Face recovery ability. Our method outperforms on all metrics, demonstrating the best original image recovery ability.

Quantitative evaluation is conducted on several aspects with anonymized images: 1) identity similarity with the original image using pre-trained Arcface (Deng et al. 2019) network, 2) face detection rate using dlib (Kazemi and Sullivan 2014) to ensure that the results are still faces, and 3) performance on different computer vision tasks, such as landmark and pose detection. We evaluate normalized L_2 landmark distance using face_alignment (Bulat and Tzimiropoulos 2017), cosine pose similarity using 6DRepNet (Hempel, Abdelrahman, and Al-Hamadi 2022), and cosine expression similarity using DECA (Feng et al. 2021). Quantitative results are shown in Table 2, our method perform the best in most of the metrics except identity similarity. Although our method does not reach the lowest identity similarity, we could observe from Figure 3 that the anonymization effect is obvious enough. Meanwhile, our method retains more identity-irrelevant attributes in the original image while Arcface uses some identity-irrelevant information for its identity embedding, which will undesirably increase the measured identity similarity.

Recovery Ability. Qualitative results for face recovery are shown in Figure 4, our method realizes the most faithful recovery of the original image while FIT appears artifacts and RiDDLE comes up with inconsistent expressions. Moreover, we use the following metrics to measure the performance of image recovery : LPIPS, MSE, PSNR, SSIM and ID similarity, quantitative results are listed in Table 3. Our method shows the best performance on all metrics, holding the best recovery ability of the original image.

To the best of our knowledge, our work is the first to consider an anonymization algorithm with multiple passwords involved. We separately demonstrate the performance of our framework in multiple scenarios below.

Sequential Anonymization and Deanonymization. Figure 5 demonstrate the result of sequential anonymization and deanonymization compared with recovery-aware method FIT and RiDDLE. Although these methods take into account the recovery of images, when multiple anonymization algorithms are applied, these methods cannot sequentially recover the corresponding images in the anonymization process. This brings a security risk that an attacker could easily tell whether an image has been anonymized through apply-

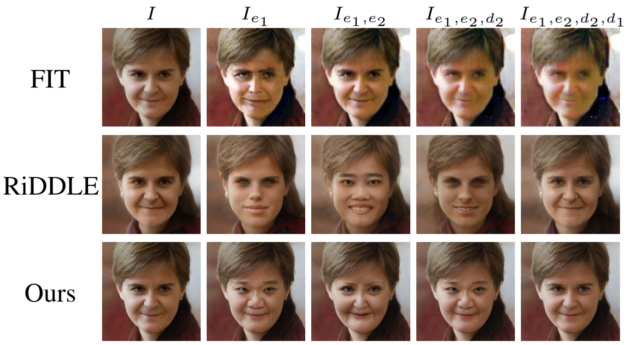

Figure 5: Sequential anonymization and deanonymization comparison. Our method could correctly recover the corresponding images while other either fails or appears artifacts.

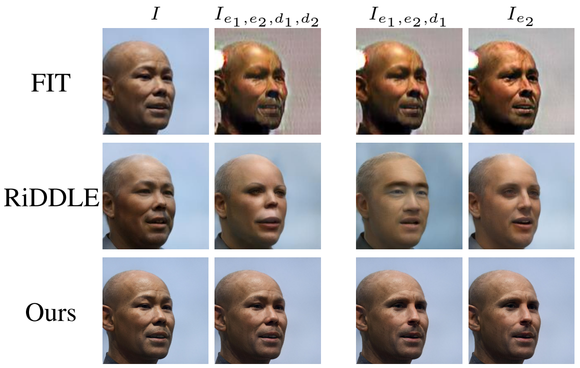

Figure 6: Commutative decryption result. The left two columns show the effect of deanonymization in another order, the right two columns show the elimination effect of pairwise used encryption and decryption password.

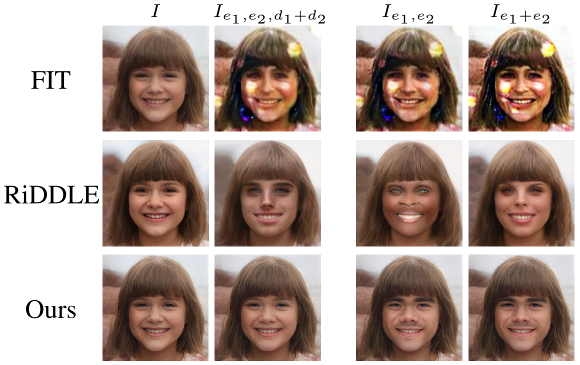

Figure 7: Result of encryption and decryption with associated password. The left two columns show the effect of equivalent password during anonymization and the right two columns show the effect during deanonymization.

ing anonymization and deanonymization method on it, and further takes DoS(Deny of Service) attack on an anonymized image. FIT fails to accurately recover the corresponding encrypted image and the original image, while RiDDLE demonstrates some recovery capability but shows inconsistencies in some facial details, such as the opening and closing of lips. Our approach, however, can precisely recover the original image and the encrypted image, which satisfies the property (f) as mentioned in Preliminaries Section.

Password Interchangeability. We then investigate the scenario where the decryption passwords are not used in the same order as the encryption passwords during the deanonymization process. In the case of RSA, when encryption and decryption passwords are used in pairs, the encryption effect will be eliminated. We present the results of using two pairs of encryption and decryption passwords in Figure 6, we could observe from the left two columns that only our method could recover the original image even if the decryption passwords are not used in the correct order. Also, the right two columns show that although (e_1, d_1) are not used continuously, it comes out that the effect of anonymization is eliminated, which satisfies the property (g).

Password Associativity. We show qualitative results for associated passwords in Figure 7. The left two columns demonstrate that only our method achieves almost the same result via an equivalent password as the anonymization algorithm sequentially used on an image with two encryption passwords. Although the case is not covered during the training phase, the right two columns show that an equivalent password also exists during the decryption process, which satisfies the properties (h).

In addition, we also performed detailed ablation experiments in the supplemental material to verify the effectiveness of our framework and loss function design.

Conclusion

Our paper introduces FaceRSA, an RSA-aware facial identity cryptography framework. The framework is built upon the latent space of StyleGAN, thanks to its good editability and redundancy nature which fits well for our task. With our well-designed password mapper, the generation process of facial identity is controlled by the user-given discrete passwords, which makes our framework a cryptosystem for anonymization and deanonymization. The mechanism to locate identity-related layers helps us to minimize the impact on other unrelated attributes while completing the editing of the identity. In addition, the customized three types of losses enable our cryptography framework to possess all the properties analogous to RSA. Extensive qualitative and quantitative comparisons demonstrate that our framework outperforms existing methods in terms of the quality of the synthesis images, preservation of identity-irrelevant information, deanonymization accuracy and properties analogous to RSA. In the future, we consider introducing some other cryptographic concepts into image anonymization and deanonymization.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62372423, U2336206, 62121002, U20B2047, 62072421, and 62002334, Key Research and Development program of Anhui Province under Grant 2022k07020008.

References

Abdal, R.; Zhu, P.; Mitra, N. J.; and Wonka, P. 2021. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (ToG)*, 40(3): 1–21.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Cao, J.; Liu, B.; Wen, Y.; Xie, R.; and Song, L. 2021. Personalized and invertible face de-identification by disentangled identity information manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3334–3342.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29.

Collins, E.; Bala, R.; Price, B.; and Susstrunk, S. 2020. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5771–5780.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Feng, Y.; Feng, H.; Black, M. J.; and Bolkart, T. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. volume 40.

Goetschalckx, L.; Andonian, A.; Oliva, A.; and Isola, P. 2019. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the ieee/cvf international conference on computer vision*, 5744–5753.

Gu, X.; Luo, W.; Ryoo, M. S.; and Lee, Y. J. 2020. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. In *European Conference on Computer Vision*.

Hempel, T.; Abdelrahman, A. A.; and Al-Hamadi, A. 2022. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2496–2500. IEEE.

Hukkelås, H.; Mester, R.; and Lindseth, F. 2019. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing: 14th International Symposium on Visual Computing, ISVC 2019, Lake Tahoe, NV, USA, October 7–9, 2019, Proceedings, Part I 14*, 565–578. Springer.

Jahanian, A.; Chai, L.; and Isola, P. 2019. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171*.

Jiang, Y.; Huang, Z.; Pan, X.; Loy, C. C.; and Liu, Z. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13799–13808.

Karras, T.; Laine, S.; and Aila, T. 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.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Kazemi, V.; and Sullivan, J. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1867–1874.

Li, D.; Wang, W.; Zhao, K.; Dong, J.; and Tan, T. 2023. RiDDLE: Reversible and Diversified De-identification with Latent Encryptor. *arXiv preprint arXiv:2303.05171*.

Li, J.; Han, L.; Chen, R.; Zhang, H.; Han, B.; Wang, L.; and Cao, X. 2021. Identity-preserving face anonymization via adaptively facial attributes obfuscation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3891–3899.

Maximov, M.; Elezi, I.; and Leal-Taixé, L. 2020. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5447–5456.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2085–2094.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rivest, R. L.; Shamir, A.; and Adleman, L. 1978. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21(2): 120–126.

Sun, J.; Deng, Q.; Li, Q.; Sun, M.; Ren, M.; and Sun, Z. 2022. Anyface: Free-style text-to-face synthesis and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18687–18696.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an Encoder for StyleGAN Image Manipulation. *arXiv preprint arXiv:2102.02766*.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Tan, Z.; Yuan, L.; Zhang, W.; and Yu, N. 2022a. Hairclip: Design your hair by text and reference image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhang, W.; Hua, G.; and Yu, N. 2023. HairCLIPv2: Unifying Hair Editing via Proxy Feature Blending. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV), 23589–23599.

Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhang, W.; Yuan, L.; Hua, G.; and Yu, N. 2022b. E2Style: Improve the efficiency and effectiveness of StyleGAN inversion. *IEEE Transactions on Image Processing*, 31: 3267–3280.

Wen, Y.; Liu, B.; Ding, M.; Xie, R.; and Song, L. 2022. Identitydp: Differential private identification protection for face images. *Neurocomputing*, 501: 197–211.

Xia, W.; Yang, Y.; Xue, J.-H.; and Wu, B. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2256–2265.

Yang, C.; Shen, Y.; and Zhou, B. 2021. Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, 129: 1451–1466.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.