Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Private-Encoder: Enforcing Privacy in Latent Space for Human Face Images

Yuan Zhao[1] | Bo Liu*[1] | Tianqing Zhu[1] | Ming Ding[2] | Wanlei Zhou[3]

[1]School of Computer Science, University of Technology Sydney, New South Wales, Australia
[2]Data61, CSIRO, New South Wales, Australia
[3]City University of Macau, Macau

Correspondence
*Bo Liu is the corresponding author. Email: bo.liu@uts.edu.au

Present Address
15 Broadway, Ultimo NSW 2007, Australia.

Summary

The explosive growth of various computer vision technologies generates a tremendous amount of visual data online every day. In addition to bringing convenience and revolutionizing our daily life, image data also reveals a wide range of sensitive information and poses unprecedented privacy leakage risks. Particular, in the case of photos contain human faces, people can easily access those face images on social media without any consent, and the misuse of personal information could cause serious privacy violation to individuals. Therefore, it is essential to consider sanitizing people's identity information when using images containing human faces. As a result, there has been rapid development in the area of facial anonymization, also called image de-identification. However, due to the emergence of numerous Deep-Learning based attacks, traditional anonymization methods such as blurring and Mosaic are weak and ineffective to protect individual's privacy in face images. To respond to this challenge, this paper proposes a novel de-identification method that utilizes a deep neural network. The proposed framework encompasses two modules: Encoder network and Generator network. The Encoder transforms a face image into a high-semantic latent vector of codes, which will be de-identified according to the differential privacy criterion. The Generator leverages the unconditional Generative Adversarial Network (GAN) to synthesize high-quality images based on the modified latent codes from the Encoder. Extensive experimental results indicate that our proposed model can protect image privacy while keeping the processed image visual realistic.

KEYWORDS:
Image Privacy, Face De-Identification, Face Anonymization, Generative Adversarial Network, Latent Space Manipulation

1 | INTRODUCTION

With the wide deployment of devices equipped with cameras, our society has witnessed a rapid increase in using and generating visual data. These data are used by people as a new form of daily communication and play a crucial role in developing advanced computer vision technologies, such as face recognition, image detection, etc. However, a great amount of sensitive information, such as human faces and/or plate numbers, are contained in the visual data. Directly sharing and using these images inadvertently pose a serious risk of privacy violation.

Government regulations such as the General Data Protection Regulations (GDPR) has went into effect by the European Union. According to GDPR, every person in the images dataset needs to consent to the use of his/her images. This regulation challenges the conventional way of research in computer vision because obtaining everyone's permission in a large-scale images dataset is nearly impossible. Fortunately, according to GDPR, if the image data does not reveal any specific person's identity information, it will be free to use without any consent. Moreover, most computer vision applications do not rely on images' identity features. For instance, image segmentation and object detection only need to detect, instead of identifying certain people in an image.

Therefore, to achieve a balanced trade-off between privacy protection and practical application, it is necessary to sanitize images' identity information while keeping the processed images real-looking.

However, for face images, anonymizing identity information to satisfy the requirement of GDPR while retaining its utility is a challenging task. Traditional anonymization techniques are mainly based on obfuscation, such as Mosaic or Blur, which are inadequate for removing privacy-sensitive information but substantially alter/destroy the original face [15]. Given a face image, an ideal de-identification method should be able to preserve the appearance features of the original image and just remove its identity characteristics. Consequently, the processed images would still look realistic to human observers and AI-based computer vision tools, such as face detectors, emotion classifiers, but people in those images cannot be identified. To be more specific, we formulate the following criterion to regulate the de-identification methods:

- **Anonymization:** The anonymization techniques should have the ability to remove privacy-sensitive information in the input images and reduce the identification possibility of the processed images by vision methods or human observers;
- **Realistic:** The processed image dataset should keep similar distribution with the original, and each image among the dataset should keep high visual quality;
- **Usability:** The complexity of the Anonymization process should be kept as low as possible;
- **Configurable:** The method should support an adjustable protection mechanism, which offers various levels of Anonymization according to users' requirement.

To satisfy the above-mentioned properties, we propose a novel privacy protection framework enforcing de-identification in latent space. Our network builds upon the unconditional GAN to produce realistic images. Unlike the conventional GAN-based image generation controlled by a random noise vector, we adopt an encoder-decoder architecture to create an operable and high-semantic latent space to implement the anonymization processing step. Besides, an Identity-Level loss function is introduced during the network's training process to regularize the network in latent space so as to provide different de-identification effects from less private to more private. Therefore, the proposed method provides configurable image anonymization.

More specifically, the anonymization process of the proposed method first encodes input image into latent space as latent codes, and then generates a de-identified version of the latent codes according to the privacy requirement. Finally, the Decoder uses the modified latent codes to generate the anonymized image. Different from manipulation in pixel space, the proposed image processing in latent space has the following advantgaes: (1) manipulation in the latent space are more accurate so it can appropriately alter original images' characteristics and features, thus preserving output image's quality and utility; (2) the entire anonymization process is unsupervised, which does not require complicated pre-processing and annotations of face areas; (3) unlike de-identification by directly altering pixels, latent space manipulation can provide rigorous privacy protection because the face information is compressed in the tractable latent vectors.

In summary, the major contributions of our works in this paper are summarized as follows:

- We present a novel face images privacy protection framework that implements de-identification of face images via editing images' identity-related features in the latent space;
- We design a dedicated and adjustable privacy-related loss function to regularize the network's training process;
- We validate that our framework outperforms both traditional protection techniques, such as blur and Mosaic, and the state-of-the-art methods, such as CIAGAN [17] and DeepPrivacy [11], regarding privacy protection as well as visual quality and utility preservation. In addition, we evaluate the impact of the privacy regularization parameter on the performance of our proposed method.

2 | RELATED WORKS

2.1 | Generative Adversarial Networks (GANs)

GAN is one of the most popular techniques in computer vision communities, which have significantly advanced image synthesis. Proposed in Goodfellow et al. [8]'s breakthrough work, the fundamental architecture of GANs is known as a combination of Discriminator and Generator. The Generator learns how to synthesize images similar to training images, while the Discriminator tries to distinguish real training images from fake synthesized images from Generator. Two networks are trained simultaneously by an adversarial process. The Generator becomes better at producing realistic images while the Discriminator becomes better at distinguishing fake images. The training achieves an equilibrium when the Discriminator can no longer detect the real images from the fake ones.

With the rapid evolution of GANs, it now has a broad diversity of application cases, from general image generation[13, 27, 3] to text-to-photo generation[28]. Benefits from those numerous contributions, GAN has become a powerful tool in computer vision. Ren et al.[20] first employ GANs to anonymization task by altering pixels value in the original image to hide the identity of the individuals. This work comes with a significant limitation that is the generated faces are still, in general, identifiable by humans observer. Hukkelas et al.[11] and Maximov et al. [17] employ Conditional GAN to perform de-identification. By using pre-annotations, their models could locate privacy-sensitive areas then replace these areas' styles with different identification to generate anonymization images. Although their works provide appealing results, both of them require complicated pre-processing and annotations. Besides, they cannot also implement different levels anonymization.

In contrast, this paper leverages unconditional GAN to remove privacy information while preserving output's visual quality and utility. One of the majority motivations is that the unconditional GANs' latent space can be directly editted in the image generation process. Especially, the state-of-the-art GAN: StyleGAN [12], which is equipped with the styles-based Generator, has demonstrated a high semantic latent space $\mathbb{W}$, carrying useful information such as gender, age group, color of hair/eyes, etc. Therefore, we can encode the image into the StyleGAN's latent space, perform privacy alternation, and then reconstruct images using the modified latent codes to obtain high-fidelity images.

2.2 | Image Privacy Protection

Until recently, there exists a limited number of research works on anonymize face images. Typically, the current standard techniques of image privacy protection, such as Mosaic and Blur, are approved ineffective and inapplicable to satisfy the emerging protection requirement. These methods protect images' privacy by directly perturbing images' Regions of Interest (ROIs) pixel values. This type of method can obfuscate corresponding sensitive information, which incurs conspicuous haziness in processed images, leading to significant utility loss [23]. Moreover, the techniques mentioned above have demonstrated significant vulnerabilities in face of the advanced convolution-based re-identification attacks [19]. MacPherson et al. [18] presented faces obfuscated by the aforementioned techniques can be re-identified up to 96% by utilizing body or scenes features from images.

Consequently, more sophisticated and novel concepts have adopted to enhance processed images' privacy and utility. For instance, Hui-Po et al. [24], and Tao et al. [14] obfuscated images' sensitive information by manipulating face attributes. The rationale of those methods is that facial attributes, such as hairstyle or eyes' colour, could be an essential reference for faces' identities. Therefore, changing these features, i.e., transform black eyes to blue eyes, seems reasonable to anonymization. Although such approaches render faithful processed images, they heavily rely on pre-defined attributes, which are impractical in the general situation. Fan [6] imposed calibrated Differential Privacy (DP) noises into the image's SVD features to achieve provable protection, which guarantees indistinguishability among visually similar images. Nevertheless, this method applies overly strong obfuscation in processed images which result in heave utility loss. Wen et al. [26] also adopts disturbance to obfuscate privacy-sensitive information, but separate images features into identity and non-identity, and then added noise to the identity vectors to protect image privacy.

3 | IMAGE PRIVACY EMBEDDING FRAMEWORK

3.1 | Network Architecture

The complete architecture of our image privacy embedding framework is illustrated in Fig. 1. It adopts an encoder-decoder architecture. We build our model on the one proposed by Richardson *et al.* [21], which aims to reconstruct input images. Nevertheless,

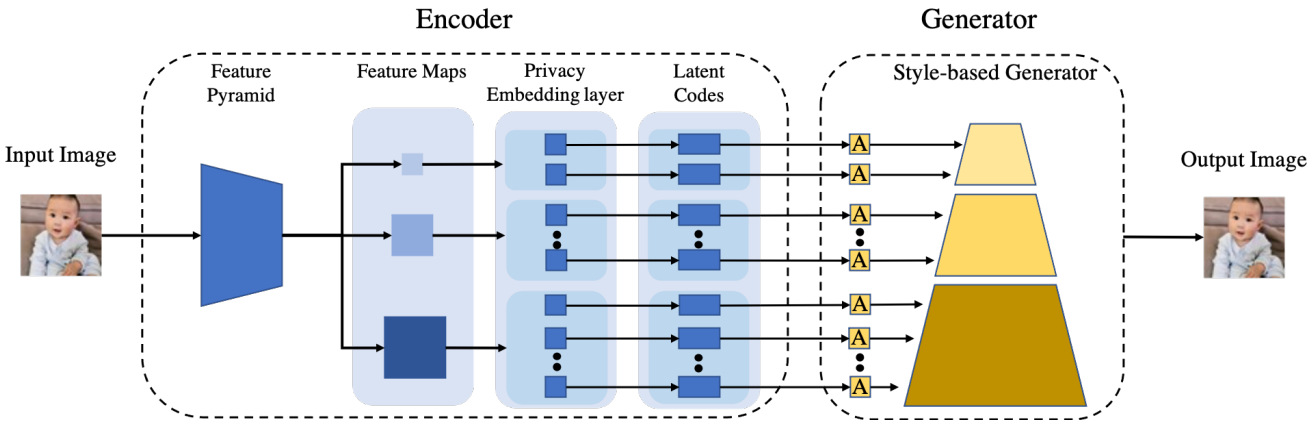

FIGURE 1 The framework of our proposed method. The Encoder consists of a Feature Pyramid network and Privacy Embedding Layers. The Generator utilizes StyleGAN2's generator network. Feature Pyramid network first converts input image into three levels feature maps in latent space. Then, Privacy Embedding layers implement manipulation on feature maps to generate privacy enforced latent code. Finally, Generator employs latent code to synthesize the output image.

the objectives of our works are not only generating images that resemble the original ones, but also limiting the amount of private information revealed in the generated images. Therefore, we perform several alterations. First, the Encoder leverages the Feature Pyramid network to project input images to spatial feature maps in the latent space. Second, it employs Privacy Embedding layers to implement semantic manipulation on feature maps to produce the privacy-anonymization latent codes. Third, the affine transform generates parameters for the fixed and pre-trained Generator network regarding these latent codes to synthesize the de-identification version of input images. The entire image-to-image translation is an end-to-end style that starts from input pixels to latent space feature maps, followed by modified latent codes, then end at output pixels. Hence, different from the state-of-the-art anonymization techniques: CIAGAN and DeepPrivacy, the proposed framework achieves image de-identification in the latent space instead of the pixel space.

3.1.1 | Encoder

The primary objective of Encoder is to generate latent vectors with respect to the input images and to perform de-identification editing on such vectors. There are two challenges to realize the goal: (1) How to project the image into the latent space accurately; and (2) How to anonymize image in the latent space semantically.

For the first challenge, a simple solution is to directly extract the same dimension vectors with respect to the Generator from the last layer of the Encoder network. However, such an approach presents a substantial bottleneck limiting the reconstruction fidelity and latent space's semantic richness [1, 2]. We attribute this limitation to the absence of original image's spatial information in the latent spaces. This is mainly because low dimension style vectors can not full reflect the original image's high-level features especially the pixels' relation in images. Without spatial information, the input image's semantics are compressed in an entangled manner, making it difficult for further manipulation and reconstruction. Therefore, our Encoder adopts a Feature Pyramid Network (FPN) as the mapping network to produce latent space with spatial dimensions. FPN projects the input images into three levels of feature maps, representing coarse, medium and fine details of the input image [21]. This property allows Encoder produces high semantic and fidelity latent space, which enable further manipulation and reconstruction.

For the second challenge, we employ a trainable Privacy Embedding Network (PEN) to transform the feature maps into latent codes for future de-identification manipulation. The PEN adopts fully convolutional layers' architecture followed by LeakyReLU activations to best comprehend and interpolate the spatial information of feature maps. Each PEN corresponds to one Latent code vector. The specific layer number of each PEN is aligned with the feature maps' hierarchical scales to guarantee to generate the same dimension latent codes. Feature Pyramid Network and Privacy Embedding Network are jointly trained to protect sensitive information in latent space.

3.1.2 | Generator

The Generator generates an output image utilizing latent codes extracted by Encoder. Motivated by the state-of-the-art visual synthesis quality and high semantic latent space, we employ a pre-trained StyleGAN2's generator network as our Generator. StyleGAN2 is equipped with re-designed generator architecture, which provides disentangled latent space $\mathbb{W}$ and editing capabilities to synthesize images. To better utilize the representative power of StyleGAN2, followed by common practice [21], we use the extended latent space $\mathbb{W}+$, which composed of the concatenation of 18 vectors $\mathtt{w}$, each with a dimension of 512 for each input layer of StyleGAN2, to control image generation.

Consequently, the latent codes, aligned with the hierarchical representation, are fed to Generated through an affine transform to generate the output image. The complete data translation of our framework is an end-to-end image-to-image translation. More specifically, we denote the Encoder's latent space encoding and manipulation process as $\mathbb{F} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{18 \times 512}$, where the input image x maps to a 18×512-dimension codes. The Generator's reconstruction transform is denoted as $\mathbb{F}^{-1} : \mathbb{R}^{18 \times 512} \rightarrow \mathbb{R}^{m \times n}$.

3.2 | Training and Losses

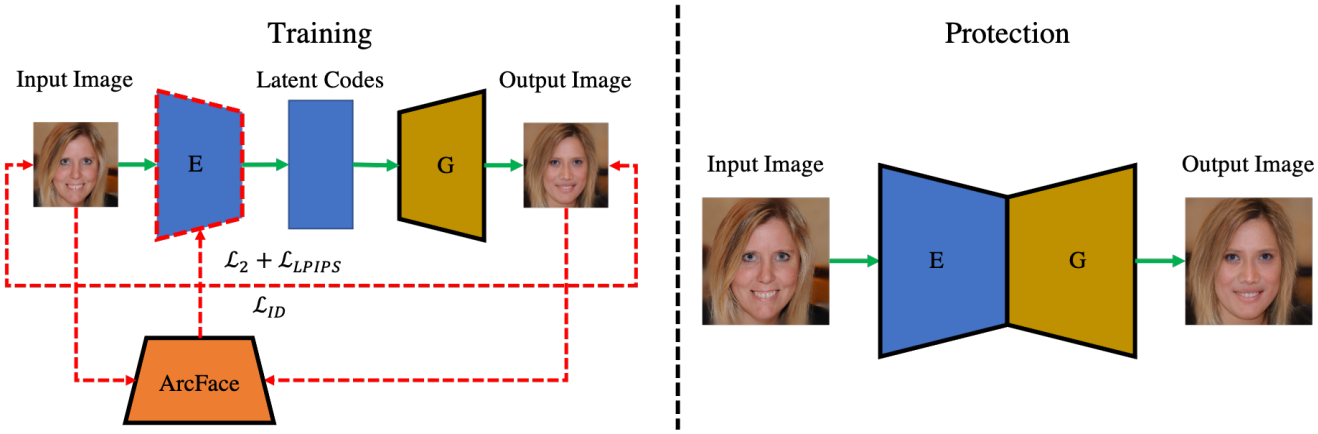

FIGURE 2 The training and protection scheme of our framework. Green arrows refer to data flows from the input image to the generated image. Dashed red lines indicate loss functions. Besides, the trapezoid with a red dash outline indicates a trainable network, while black full line trapezoids represent fixed and pre-trained networks.

Fig. 2 left part illustrates the training scheme of our framework. We use **E** and **G** to denote our Encoder and Generator. Since the Generator network is built upon the representative power of pre-trained StyleGAN2's generator [12]; therefore, only the Encoder is updated during the training to achieve image anonymization. Besides, the entire training scheme does not require any pre-annotations. Encoder implements all image manipulation operations on images' latent space instead of Generator on the pixel level. To semantically guided the training, we utilize a weighted combined loss function, which consists of three dedicated sub-loss functions for different objectives:

Pixel-Level Loss: $\mathcal{L}_2$ loss is adopted to enforce the reconstructed images $\hat{x} = G(E(x))$ to pixel-wise resemble input images x,

$$\mathcal{L}_2(x) = \|x - \hat{x}\|_2, \tag{1}$$

where $E(\cdot)$ denotes Encoder network, $G(\cdot)$ denotes Generator network, x and $\hat{x}$ are original and corresponded processed image.

Perceptual-Level Loss: In addition to preserving perceptual quality, we leverage the Learned Perceptual Image Patch Similarity (LPIPS) [29] loss to encourage the reconstructed images perceptually similar with the originals,

$$\mathcal{L}_{\mathbf{LPIPS}}(x) = \|L(x) - L(\hat{x})\|_2, \tag{2}$$

where $L(\cdot)$ represents the perceptual features extractor.

Identity-Level Loss: To limit the amount of private information presented in the reconstructed images, we regularize the cosine similarity between the input and reconstructed images' identity feature vectors. Specifically, by employing the pre-trained

ArcFace network [5], we obtain the identity features vector of images. Then, we set up a privacy regularizer $\beta \in [0, 1]$ to restrict the similarity between input and reconstructed images' identity features to reduce the privacy information exposed in the reconstructed images. Formally, the identity loss function is written by:

$$\mathcal{L}_{ID}(x) = |\beta - Cos(Arc(x) - Arc(\hat{x}))|, \tag{3}$$

where $Cos(\cdot)$ denotes cosine similarity and $Arc(\cdot)$ denotes pre-trained ArcFace network. Besides, accompany with the privacy regularizer β's, the identity-level loss will impose a different level of privacy protection effects. As $Cos(Arc(x) - Arc(\hat{x})) = 1$ indicates highest similarity between x and $\hat{x}$, a smaller β will enforce larger distance in identities and therefore better privacy protection.

The overall weighted sum loss function is defined as:

$$\mathcal{L}(x) = \lambda_1 \mathcal{L}_2(x) + \lambda_2 \mathcal{L}_{LPIPS}(x) + \lambda_3 \mathcal{L}_{ID}(x), \tag{4}$$

where $\lambda_1, \lambda_2, \lambda_3$ are constant weighting corresponded loss.

3.3 | Protection Stage

The right-hand-side part of Fig. 2 illustrates the protection scheme of our framework. With our model trained to minimize loss function Eq. (4), the network enables de-identification of the input images. During this stage, both Encoder and Generator are fixed. Therefore, the input image is encoded into latent space, and then processed by the proposed privacy-enhancement mechanism, resulting in an output a privacy-preserving latent code. The Generator will then synthesize a de-identification image according to the privacy-preserving latent code. The Latent Codes part is omitted in Fig. 2 for brevity.

3.4 | Attack Model

We consider a robust threat model to validate our framework's privacy protection capability in a worst-case scenario. The adversary's objective is to learn personal identity by accessing images and then using the extracted identity information to match other people's images illegally. For example, an adversary can utilize the face on Google street view to search corresponding individual social network accounts or other personal images published on the Internet to further illegally surveil people. We assume that the adversary can acquire all processed images shared in online social networks but have no access to the original images (which represent corresponding personal images without processed by privacy-enhancement methods). Besides, the adversary is capable of utilizing state-of-the-art face recognition methods to launch identification attacks.

To quantify this risk, we calculate the **Identity Similarity** between the original and processed images,

$$Id\_Similarity = Cos(F(x) - F(\hat{x})), \tag{5}$$

where $F(\cdot)$ represents identity features extractor which based on pre-trained facial recognition networks.

Specifically, the higher Identity Similarity between the original and processed images indicates a higher possibility of success illegal identification by the adversary, and hence a lower privacy-level, and vice versa. Therefore, the objective of image privacy protection techniques is to reduce their output image's Identity Similarity compared with that of the input one. Given by the dedicated Identity-Level loss, our framework provides adjustable control over the processed images' Identity Similarity with the original images. Hence, our framework could effectively defence the re-identify attack, despite the state-of-the-art facial recognition model.

4 | EXPERIMENT

In this section, we implement extensive and comprehensive experiments to evaluate our framework's effectiveness of identity anonymization. The proposed method is compared with both classic and state-of-the-art anonymization methods on various faces image datasets. The experiment results indicate that the proposed method acquires the best performance regarding various qualitative and quantitative evaluation metrics. Besides, we also present a set of comparisons to reflect how privacy regularizer β affects the anonymization performance of our method. The datasets, baselines and evaluation metrics will introduce in the following:

Datasets. The experiments are conducted on two public well-known faces image datasets to exhibit the performance of the proposed image de-identification framework.

- **CelebA [16]:** the dataset consists of 202599 face images with various features, such as age, gender and race. For a fair comparison, we use the aligned version where each image centred on a point in-between person's eyes and then resized to 256×256 resolution. Only 20k images are randomly selected to train the proposed model for saving time.
- **Flickr-Faces-HQ [7]:** This dataset is composed of 70k high-quality PNG images with 1024×1024 resolution and also provides considerable coverage in terms of personal age, ethnicity, image background, accessories, etc. To reduce training complexity and time, we randomly selected 10k image from this dataset. Every selected image are aligned and cropped to the cetral point, then resized to a resolution of 256×256.

Comparative studies. We compare two classic anonymization methods and two state-of-the-art learning-based techniques.

- **Classic Methods:** We use Mosaic and Blur to compare them with our method. Both of them are current mainstream and most commonly used image privacy-enhanced techniques which well represent the traditional methods.
- **Learning-Based Method:** We select DeepPrivacy [11] and CIAGAN [17] as benchmark schemes. We adopt the official codes and pre-trained models given by the authors. These two methods are selected because they satisfy our proposed de-identification criterion and achieved better performance compared with the other existing learning-based methods.

4.1 | Evaluation Metrics

To quantitatively evaluate and compare the interested schemes, we employ the following metrics to assess their performance in the aspects of visual quality, privacy protection and utility.

4.1.1 | Visual Quality Metrics

Three different evaluation metrics are employed to measure the visual quality of the de-identification images:

- **MSE:** This metric calculates the pixel-wise Mean Square Error (MSE) between the input and anonymous images to compare different outputs visual quality at the pixel-level. A lower MSE value indicates a higher similarity between the original and the de-identification images, implying better visual quality preservation.
- **SSIM [25]:** Rather than directly comparing the images pixel by pixel, we use Structural Similarity (SSIM) to measure the perceptual difference between the input and processed images incorporating Luminance, Contrast and Structure. Therefore, a lower SSIM indicates better images' visual quality preservation from the human perceptual perspective.
- **FID [9]:** Different from the previous metrics which measure pair-wise image similarity, Frechet Inception Distance (FID) calculates the Fréchet distance between the input and processed image datasets' multidimensional Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ using the Inception v3 [22] features to quantify their quality similarity. A lower FID represents better quality preservation.

4.1.2 | Privacy Metrics

The objective of the privacy metrics is to evaluate the performance of privacy protection. There are two different privacy metrics used in our experiments.

- **Identity Similarity:** According to Eq. (5), we define **Identity Similarity** as below. This metric calculates the cosine similarity of the original and anonymized images' identity feature vectors to quantify the effectiveness of privacy protection. As the ArcFace model is employed in our encoder's loss function, we leverage another state-of-the-art Facial Recognition network's pre-trained model, CurricularFace [10], to extract the images' identity features.

$$Id\_Similarity = Cos\left(CF(x) - CF(\hat{x})\right),$$

where $CF(\cdot)$ denotes pre-trained CurricularFace-based identity feature extractor. A lower Identity Similarity value between the original and processed images indicates a higher level of de-identification. Averaging the Identity Similarity among 10k random identities from the FFHQ dataset, we obtain an empirical threshold δ value: 0.19. Hence, in the following experiments, an image pair with an Identity Similarity lower than 0.19 will be regarded as different identities.

- **De-Identity Rate:** Besides the **Id_Similarity**, we present another evaluation metric**: De-Identify Rate** $= \hat{y}/y$, where $\hat{y}$ is the number of image pairs that can be recognized by the pre-trained CurricularFace as different identities, and y is the total number of images pairs in the experiment. This metric is used to measure the ratio of the anonymized images that have completely removed the original identity characteristics.

4.1.3 | Utility Metric

The processed images using the anonymization methods should maintain a high utility in practical identity-agnostic computer vision tasks, such as face detection. To quantitatively compare the studied methods in terms of utility preservation, we perform face detection using the standard Dlib-ml library's HOG-based face detector [4] on their processed images. We measure the percentage of detected faces to evaluate the performance of each anonymization method, with 1.0 representing perfect utility preservation.

4.2 | Impact of Privacy Regularizer

We now discuss the impact of the privacy regularizer β on the visual quality, utility and privacy. Recall that β is incorporated in the identity loss function to regulate the input and processed images' identity similarity. A lower β leads to a higher variation, and hence more potent privacy protection on the processed images, and vice versa. We train our framework by varying β from 0.0 to 1.0 with an interval of 0.1 to construct different models. Then, we calculate the defined metrics over different models to evaluate the impact of β on the performance.

4.2.1 | Visual Quality Evaluation

First, we show the experiment results of visual quality. As illustrated in Fig. 3, the trend of the quality metrics is consistent with each other. The processed images' quality increases with the decrease of the required privacy protection level. This phenomenon indicates that altering the image's identity features will also reduce the quality of the reconstructed images. However, according to quantitative results, the quality reduction is not obvious, which verifies that our method can generate sufficiently high-quality images while providing privacy protection.

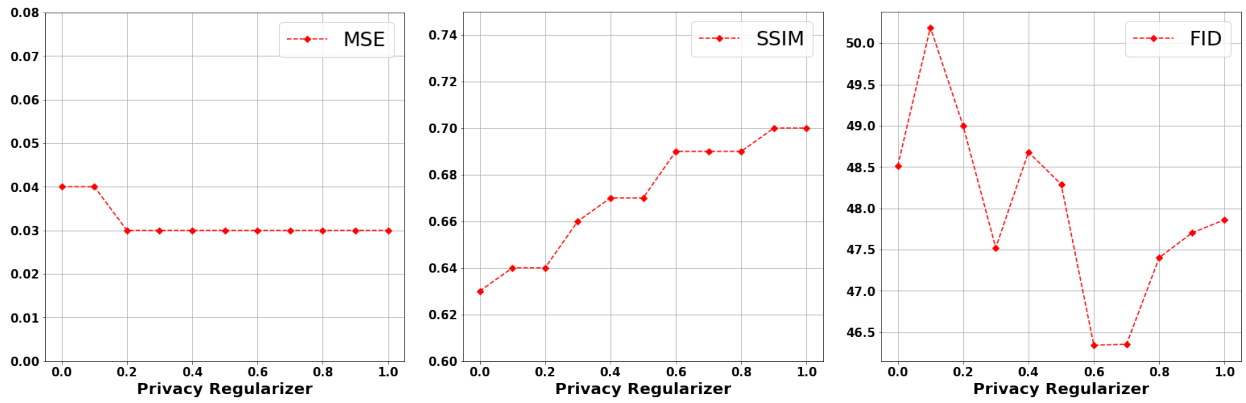

FIGURE 3 The utility metrics corresponding to β from 0 to 1.

.

4.2.2 | Privacy Protection Evaluation

The quantitative results of privacy protection are shown in Fig. 4. With the relaxation of privacy regularizer, the average of Id_Similarity continues to rise, and the De_Identify Rate declines, which shows that a smaller privacy regularizer provides a higher privacy protection level, and vice versa. Moreover, there is an "elbow" point appearing at around $\beta = 0.2$ on the De_Identify Rate curve, where the privacy protection level on the processed images starts drops rapidly. Besides, the privacy

protection becomes negligible at $\beta = 0.4$ when facing the re-identification attacks using the state-of-the-art facial recognition models.

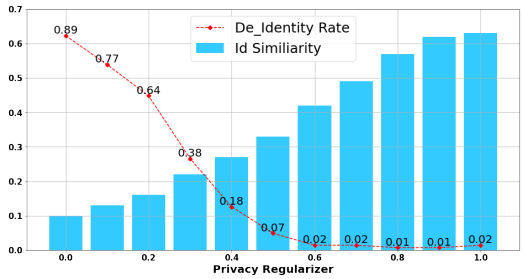

FIGURE 4 Average Id_Similarity and De_Identify rate along with privacy regularizer β increase from 0 to 1.

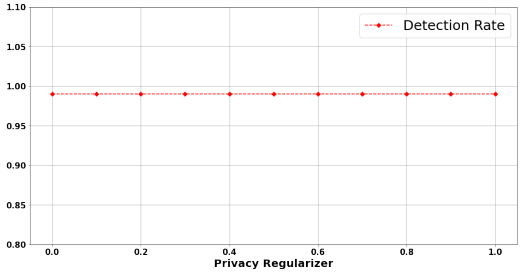

FIGURE 5 Detection Rate along with privacy regularizer β increase from 0 to 1.

4.2.3 | Utility Performance Evaluation

We show in Fig. 5 the results of utility with respect to the range of privacy regularizer. The detection rates remain at almost 0.99 with various privacy regularizers, demonstrating that our method achieves a nearly perfect score in preserving utility. Besides, it also proves that some computer vision tasks, such as face detection, are identity-agnostic, which do not rely on people's identity information. Therefore, the proposed anonymization techniques could be employed to protect the privacy in the publicly available large-scale face image datasets while preserving their utility in computer vision tasks.

4.2.4 | Qualitative Comparison:

Furthermore, we visualize several samples with different β in Fig. 6 to conduct a qualitative comparison. As β decreases, the visual identity of the processed image significantly changes comparing with that of the original one, while most of the non-identity features are retained to generate a high fidelity for the processed images.

4.3 | Comparison with Classic Methods

In this subsection, We present comparison experiments between our method and the mainstream image anonymization techniques, i.e., blurring and mosaic. For the sake of fairness, all methods will be calibrated to reach a comparable value in terms of a performance metric value, and then we will apply the other performance metrics to evaluate their performance. The experiments in this section are conducted using the **FFHQ** dataset.

4.3.1 | Visual Quality Evaluation

We first evaluate the visual quality of the anonymized images. Our model and two benchmark methods are fine-tuned to make their privacy metric values reach the following numerical range [0.1, 0.2, 0.4], which represents a variety of privacy protection levels in the order of strength. Then, we evaluate the aforementioned visual quality metrics. The results are summarized in Tables 1, 2 and 3. As shown in these tables, our framework outperforms the blur and mosaic techniques in every category of performance metrics at all of the investigated privacy protection levels. These results show that for a given privacy protection level, our method can generate a higher utility compared with the conventional techqniues.

4.3.2 | Privacy Protection Evaluation

In this subsection, we evaluate the privacy protection performance of our method. Similar to the evaluation of visual quality, we calibrate the interested methods to achieve a similar SSIM value for fair comparison. From the experimental results, we find that our framework can achieve a relatively stable SSIM value at around 0.65, with different sets of parameters (more details will be discussed in the latter part of this section). In the following, we only evaluate the privacy protection level under an SSIM of 0.65.

FIGURE 6 Qualitative comparison of different privacy regularizer β. With a lower regularization parameter, the processed image's identity similarity significantly different from the original. Besides, corresponding to discover in De-identify Rate curve, we note that images output by models whose β higher than 0.4 still reserve very similar visual identity with the original. Only images processed by models whose β lower than 0.2 have relatively large difference with the original.

TABLE 1 Quality metrics at identity similarity around 0.1.

	MSE↓	SSIM↑	FID↓
Blur	0.12	0.50	158.23
Mosaic	0.42	0.40	130.71
Ours	**0.04**	**0.63**	**48.51**

TABLE 2 Quality metrics at identity similarity around 0.2.

	MSE↓	SSIM↑	FID↓
Blur	0.05	0.60	98.65
Mosaic	0.10+	0.46	119.99
Ours	**0.03**	**0.66**	**47.52**

TABLE 3 Quality metrics at identity similarity around 0.4.

	MSE↓	SSIM↑	FID↓
Blur	**0.03**	0.65	65.82
Mosaic	0.05	0.54	108.03
Ours	**0.03**	**0.69**	**46.33**

TABLE 4 Identity Similarity at same SSIM value (0.65).

MethodsPrivacy	Identity Similarity↓
Blur	0.38+-0.11
Mosaic	0.87+-0.04
Ours	**0.16+-0.08**

TABLE 5 Detection Rate at same SSIM value (0.65).

MethodsUtility	Detection Rate↑
Blur	0.4575
Mosaic	0.9818
Ours	**0.9999**

Table 4 shows that our method can significantly reduce the identity similarity between the input and processed images comparing with the benchmark techniques, which indicates that our method can provide a higher privacy protection level.

4.3.3 | Utility Performance Evaluation

Next, we present the evaluation results of utility. We calculate the detection rate of each method when the Identity Similarity reaches [0.1, 0.2, 0.4] and the SSIM is around 0.65, respectively. The results are reported in Fig.7 and Table 5.

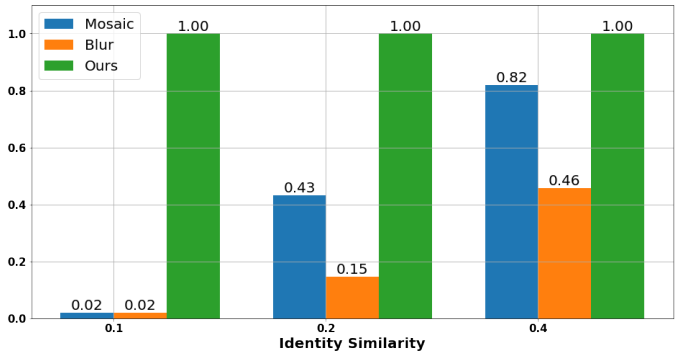

FIGURE 7 Detection Rate at Identity Similarity on [0.1, 0.2, 0.4].

As shown in Fig. 7, our anonymized images consistently achieve 100% detection rates under various privacy metric values. This result indicates that the proposed method could perfectly maintain image utility in the face detection task. On the contrary, the mosaiced and blurred images have much lower detection rates, indicating that these anonymization techniques incur heavy utility loss in detection tasks. Table 5 shows that the mosaic and blur techniques will inevitably cause utility loss even under the same visual quality, while our method shows perfect utility preservation.

4.3.4 | Qualitative Comparison

Apart from the above quantitative comparison, we also illustrate several original and processed images in Fig. 8 with SSIM=0.65 to qualitatively exhibit the privacy protection. Regardless of the relatively high SSIM value, the blur and mosaic technique lead to noticeable perturbation on images, which will significantly compromise their applications in practice. In contrast, our method semantically modifies the ROIs of the images, while maintain the fidelity in the processed images.

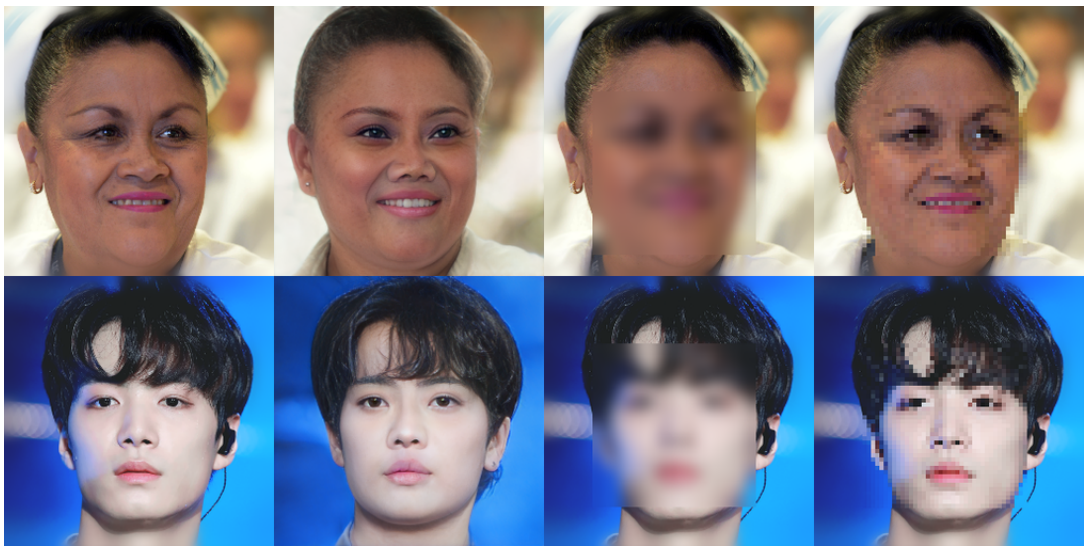

FIGURE 8 Qualitative comparison between our method and classic anonymization techniques under SSIM values: 0.65, from left to right: **Original**, **Ours**, **Blur** and **Mosaic**.

4.4 | Comparison with the state-of-the-art Methods

This section compares our method with the state-of-the-art de-identification methods, i.e., DeepPrivacy [11] and CIAGAN [17], which are both trained and tested on the CelebA dataset. Thus, we also apply our framework to the CelebA dataset for fairness comparison.

Both the reference works cannot adjust the privacy protection levels. Hence, we calculate their outputs' Identity Similarity to be 0.1 and 0.2, respectively. Then we fine-tune our model to achieve the same Identity Similarity and conduct a comparative experiments. Table 6 lists the quantitative comparison results. In terms of the visual quality metrics, CIAGAN achieves an impressive performance on FID by obtaining a score of 12.72. Our method is slightly inferior to CIAGAN by achieving an FID score of 31.11. Although FID is usually employed as an important metric to evaluate the output quality of GANs, it is calculated based on the distribution of generated images, which cannot fully capture the quality of a single image. Besides, our method outperforms CIAGAN and DeepPrivacy in the other Visual Quality metrics of MSE and SSIM. It shows that our network could generate anonymous images with comparable visual quality.

Fig. 9 illustrates more perceptual comparison results. From this figure, we can see that our model produces more visually-realistic anonymous faces that preserve more characteristics of the original identity. In contrast, the process images from CIAGAN look different to the source images, because of the direct change of the original ID. However, when the fake Identification does not share the same gender, age or makeup, CIAGAN tends to produce extremely unrealistic images (e.g., row 3, column 5 in Fig. 9). Besides, distortions and artifacts often occur on their processed images. The processed images from DeepPrivacy could relatively well keep the facial pose and outline, nevertheless it adds fuzziness on the face area. In addition, both CIAGAN and DeepPrivacy share another significant flaw, i.e., these two techniques rely on facial landmark detection to provide pre-annotation and require to feed their networks with face-removing images, making it difficult to deploy them in real-world applications. On the contrary, our approach does not have these issues and can provide adjustable privacy protection.

TABLE 6 Quality metrics at Identity Similarity around 0.2.

	MSE↓	SSIM↑	FID↓	DR↑
CIAGAN	0.07+-0.02	0.65+-0.07	**12.72**	0.9939
Ours(Id=0.2)	**0.02+-0.00**	**0.73+-0.08**	31.11	0.9989
DeepPrivacy	0.09+-0.04	0.61+-0.09	25.94	0.9976
Ours(Id=0.1)	**0.02+-0.00**	0.72+-0.08	33.41	0.9976

4.5 | Discussions

In summary, the experimental results demonstrate that our method could provide adjustable privacy protection, while generating sufficiently high-quality images. This makes our method capable of satisfying different application requirements in practice. From the presented results, it is obvious that our approach significantly outperforms the classic obfuscated-based methods in anonymization task to achieve a balanced trade-off among visual quality, privacy protection and utility preservation. Compared with the deep learning based methods, i.e., CIAGAN and DeepPrivacy, our method can provide more semantic and accurate anonymization. The qualitative results show that the generated images from the propsoed method can retain more original characteristics. In contrast, both CIAGAN and DeepPrivacy fail to preserve enough original features.

However, according to our extensive experiments, we find several weaknesses of the current deep learning based de-identification methods. First, these methods rely on face detection. Any face that is not detected by the deep learning based methods cannot be anonymized. Our method suffers from a similar issue as it depends on the pre-trained StyleGAN. Thus, it is challenging to anonymize face images that are not facing forward because such examples are not available during the StyleGAN training process. In addition, faces covered with objects, such as earrings, are extremely hard to process. To generate a face with such features requires a more careful design, which is still an open problem for the deep learning based methods.

FIGURE 9 Qualitative comparison between our method and SOTA anonymization techniques DeepPrivacy and CIAGAN. From top to bottom show the outputs of: **Original**, **DeepPrivacy**, **CIAGAN**, **Ours** ($\beta = 0.0$), and **Ours** ($\beta = 0.2$).

5 | CONCLUSION

This paper presents a novel image privacy protection framework that could protect the image's privacy in the latent space and achieve a balanced trade-off between the image's privacy, utility and quality. The proposed framework consists of an Encoder and a Generator. Input images are translated by Encoder into the latent space and then subject to semantic manipulations to protect privacy of faces. Using the Encoder's output, the Generator is built upon a pre-trained unconditional GAN to reconstruct a high-fidelity and anonymous image. The advantages of our framework are two-fold: i) it can remove the identity information in the target image while retaining the other information that has nothing to do with identity (such as image structures), thereby providing a visually realistic image, and ii) the degree of de-identification can be controlled via a parameter to provide adjustable protection so that users can flexibly tune their requirements of privacy and utility. Our experimental results demonstrate the effectiveness of our framework in real-world image datasets, thanks to its ability to generate comparable performance metrics with the classic techniques as well as the state-of-the-art methods. In the future, we will further explore the disentanglement of sensitive and non-sensitive attributes in images as well as videos.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2stylegan: How to embed images into the stylegan latent space?" In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4432–4441.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2stylegan++: How to edit the embedded images?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8296–8305.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: (2019). arXiv: 1809.11096 [cs.LG].

[4] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)* 1 (2005), pp. 886–893.

[5] Jiankang Deng et al. “Arcface: Additive angular margin loss for deep face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4690–4699.

[6] FanLiyue. “Practical image obfuscation with provable privacy”. In: *2019 IEEE International Conference on Multimedia and Expo (ICME)* (2019).

[7] *Flickr-Faces-HQ (FFHQ) Dataset*. Nov. 2020. URL: https://github.com/NVlabs/ffhq-dataset.

[8] Ian J Goodfellow et al. “Generative Adversarial Networks”. In: (2014). arXiv: 1406.2661 [stat.ML].

[9] Martin Heusel et al. “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”. In: (2018). arXiv: 1706.08500 [cs.LG].

[10] Yuge Huang et al. “Curricularface: adaptive curriculum learning loss for deep face recognition”. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5901–5910.

[11] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. “Deepprivacy: A generative adversarial network for face anonymization”. In: *International Symposium on Visual Computing* (2019), pp. 565–578.

[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* (2019), pp. 4401–4410.

[13] Tero Karras et al. “Progressive Growing of GANs for Improved Quality, Stability, and Variation”. In: (2018). arXiv: 1710.10196 [cs.NE].

[14] Tao Li and Lei Lin. “Anonymousnet: Natural face de-identification with measurable privacy”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 0–0.

[15] Bo Liu et al. “When Machine Learning Meets Privacy: A Survey and Outlook”. In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36.

[16] Ziwei Liu et al. “Deep learning face attributes in the wild”. In: *Proceedings of the IEEE international conference on computer vision* (2015), pp. 3730–3738.

[17] 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* (2020), pp. 5447–5456.

[18] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. *Defeating image obfuscation with deep learning*. arXiv eprint:1609.00408 cs.CR. 2016. arXiv: 1609.00408 [cs.CR].

[19] Seong Joon Oh et al. “Faceless person recognition: Privacy implications in social media”. In: *European Conference on Computer Vision(ECCV)* (2016), pp. 19–35.

[20] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. “Learning to anonymize faces for privacy preserving action detection”. In: *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 620–636.

[21] Elad Richardson et al. “Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation”. In: (2021). arXiv: 2008.00951 [cs.CV].

[22] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2818–2826.

[23] Nishant Vishwamitra 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* (2017), pp. 39–47.

[24] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. “InfoScrub: Towards Attribute Privacy by Targeted Obfuscation”. In: *arXiv preprint arXiv:2005.10329* (2020).

[25] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[26] Yunqian Wen et al. “A Hybrid Model for Natural Face De-Identiation with Adjustable Privacy”. In: *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)* (2020), pp. 269–272.

[27] Han Zhang et al. “Self-attention generative adversarial networks”. In: *International conference on machine learning* (2019), pp. 7354–7363.

[28] Han Zhang et al. “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision* (2017), pp. 5907–5915.

[29] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 586–595.