Received: 18 August 2022 | Revised: 15 November 2022 | Accepted: 21 November 2022

DOI: 10.1002/cpe.7554

RESEARCH ARTICLE

WILEY

Face image de-identification by feature space adversarial perturbation

Hanyu Xue[1] | Bo Liu[1] | Xin Yuan[2] | Ming Ding[2] | Tianqing Zhu[1]

[1]Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Ultimo, New South Wales, Australia

[2]Data61, CSIRO, New South Wales, Australia

Correspondence
Bo Liu, Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Ultimo, NSW 2007, Australia.
Email: bo.liu@uts.edu.au

Funding information
ARC Linkage Project, Grant/Award Number: LP180101150

Summary

Privacy leakage in images attracts increasing concerns these days, as photos uploaded to large social platforms are usually not processed by proper privacy protection mechanisms. Moreover, with advanced artificial intelligence (AI) tools such as deep neural network (DNN), an adversary can detect people's identities and collect other sensitive personal information from images at an unprecedented scale. In this paper, we introduce a novel face image de-identification framework using adversarial perturbations in the feature space. Manipulating the feature space vector ensures the good transferability of our framework. Moreover, the proposed feature space adversarial perturbation generation algorithm can successfully protect the identity-related information while ensuring the other attributes remain similar. Finally, we conduct extensive experiments on two face image datasets to evaluate the performance of the proposed method. Our results show that the proposed method can generate real-looking privacy-preserving images efficiently. Although our framework has only been tested on two real-life face image datasets, it can be easily extended to other types of images.

KEYWORDS

adversarial perturbation, feature space, image, privacy

1 | INTRODUCTION

The rapid development of computer vision (CV) technology has recently made the automatic processing of large-scale visual data prevalent. However, those visual data contain a large amount of personal information, leading to inadvertent disclosure of an individual's privacy. While we enjoy the benefits of advanced CV technology, including camera surveillance and video conferencing, we are reluctant to surrender our privacy and in-dignify ourselves as manipulable data records. In addition, the information that people share on the Internet is facing more powerful malicious attackers than ever before. The traditional privacy-preserving methods are less effective against deep learning based attacking models. Therefore, new privacy preservation methods are urgently needed.

Various defense techniques and mechanisms have been proposed to enhance privacy by de-identifying face images.[1] Traditional obfuscation-based methods usually obfuscate the sensitive information by blurring, pixelating, or masking an image, which are illustrated in Figure 1. However, the image processed by these traditional methods can be easily and accurately detected by deep learning models. Therefore, new defence techniques and mechanisms have been designed to protect image privacy from the deep learning models, including face identity transformer, differential privacy (DP), GAN-based inpainting, and adversarial examples (AEs), and so forth.[2-9] The face identity transformer was proposed in Reference 2, which can perform automatic photo-realistic password-based anonymization and deanonymization of human faces. DP-based methods can provide provable privacy guarantees, but produce lower-quality images by including DP noises to the original image or in the transformed domain of the image.[3] GAN-based inpainting was proposed to generate content to cover the sensitive information or identity of the image without

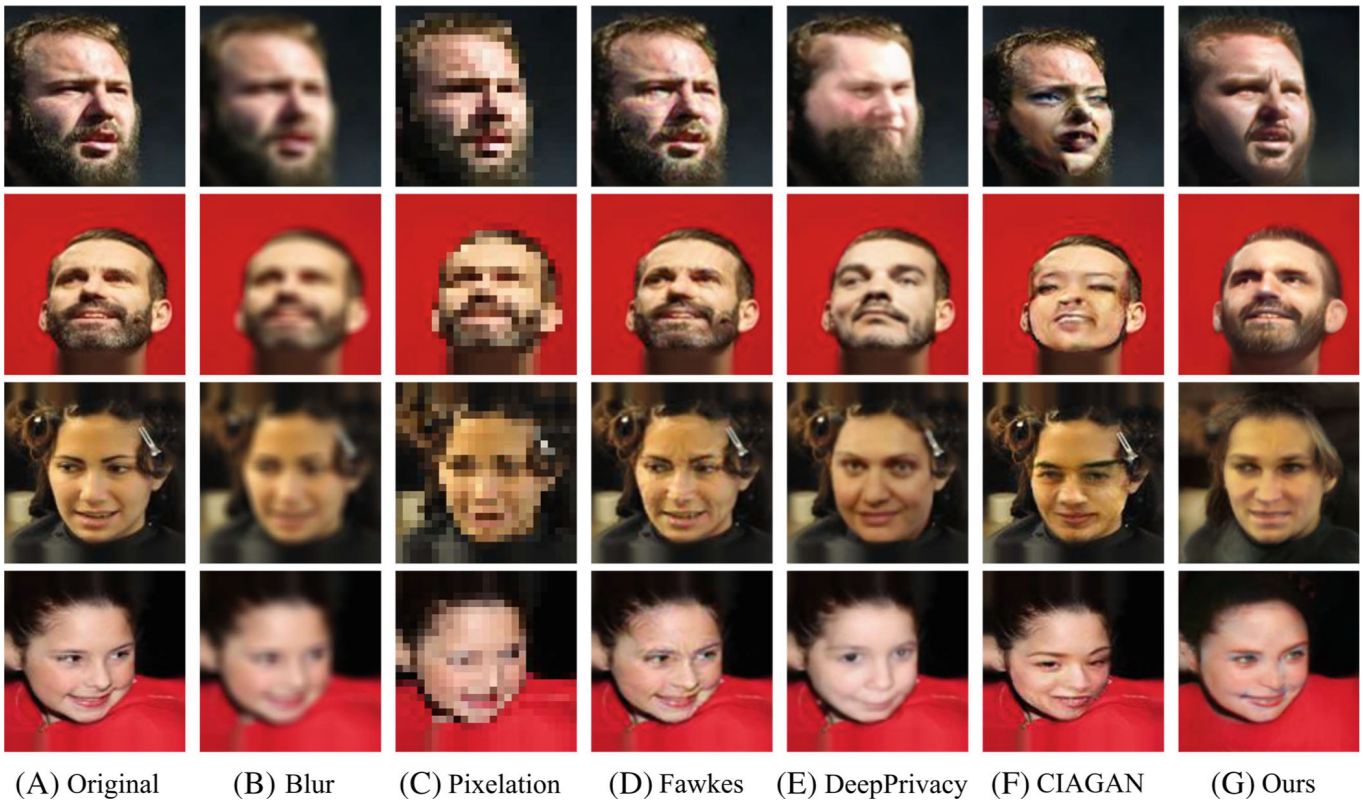


FIGURE 1 Face de-identification results. From left to right, (A) Original image; (B) Blur noise; (C) Pixelation noise; (D) Fawkes;[21] (E) DeepPrivacy;[11] (F) CIAGAN;[10] (G) feature space adversarial perturbation (**Ours**).

degrading the quality of the original image.[4] The conditional GAN (CGAN)-based method was designed by adding labels to the generator and the discriminator for better network training.[5] Conditional Identity Anonymization Generative Adversarial Network (CIAGAN)[10] and DeepPrivacy,[11] both based on CGAN, proposed to add AE noises in the feature space of the images. CIAGAN can de-identify faces and bodies while generating high-quality images and videos. DeepPrivacy can generate an image with considerations of both pose and background. Nevertheless, both CIAGAN[10] and DeepPrivacy[11] add the latent noise in a vague direction, without considering the specific features of the image.

Considering deep learning-based privacy attacks, adversarial examples (AE)-based protection methods have great potential. The earliest research on AE was proposed by Szegedy et al.[12] It was shown that a small perturbation could have a considerable and negative impact on the accuracy of deep neural networks. In a recent paper,[13] the author designed an adversarial example attack against deep neural networks to mislead the classifier. It was revealed that the deep neural network is different from humans in the task of image classification, and AE is an efficient method to generate noise on images that affects the deep neural networks. Many subsequent studies have focused on adversarial noise in different settings, such as adversarial noise for the convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), and more robust adversarial noise.[14] There is an arms race-like relationship between attack and defence technologies in such circumstances. One of the major issues of AE is its transferability, that is, its effectiveness on alternative black-box or unknown models. To improve AE noise's transferability, some papers[15,16] have transferred the calculation of noise direction from the output layer to the intermediate layer of the model. This can avoid the differences between models, thereby increasing transferability. Pidhorskyi et al.[17] studied the potential of adding adversarial perturbations on the feature level of images.

Most of the above existing work treated AE as a threat to system security. Only very recently, researchers started to use adversarial examples (also called adversarial perturbations) as a method to protect image privacy.[18-23] However, these methods either focus on the utility[18-22] or focus on the privacy protection,[23] which is hard for users to choose a good trade-off.

To overcome the problems mentioned above, we propose a novel face image de-identification framework, where latent noise is generated based on the gradient directions concerning the identity and the attributes of an individual face image, which can accurately de-identify the image. Moreover, we use a pre-trained model as a decoder that can map the perturbed feature vector back to an image (i.e., the generated AE). The main contributions of our work are summarized as follows:

1. We propose a novel face image de-identification framework based on feature space adversarial perturbations referred to as the FSAP framework for short. This framework can preserve face identity information against automated recognition by DNNs while keeping a high utility of the image.

2. We propose a feature space adversarial perturbation generation algorithm. By alternative updating the perturbation according to the specially designed ID loss and attribute loss items, we can successfully direct the noise to identity-related information while ensuring the other attributes remain similar. In addition, feature space manipulation can provide good transferability of the generated perturbation. Furthermore, users can select a trade-off between privacy and utility according to their own needs by adjusting the parameters.
3. We implement the proposed image protection framework and methods on a real-life image dataset and show its effectiveness in safeguarding people's privacy.

2 | RELATED WORK

Recent image privacy researches[24-26] have focused on altering identity-related information in images via different methods, including obfuscation, GAN-based in-painting, differential privacy (DP) and adversarial examples (AEs).

The main techniques under investigation are obfuscation and in-painting. Simple obfuscation has been shown to be ineffective against DNN-based recognizers.[27,28] Therefore, some researchers have started to use GAN and AE to generate content to replace the sensitive information in the images.[29-34] For example, Sun et al.[29] proposed GAN-based head in-painting to remove the original identities. Hukkelås et al.[11] proposed a CGAN-based architecture to anonymize faces without destroying the data distribution of the original image. Deb[35] proposed a framework to generate face masks based on GAN and AE techniques. Valeriia[36] proposed a method to optimize the AE method in privacy protection. Zhang[37] proposed a face protection framework against DNN by filtering the AE methods.

Furthermore, there have been some recent attempts to combine the DP with image privacy.[38] Fan[39] proposed an ε-differential private method at the pixel level of the image. However, making image pixels indistinguishable does not make much sense in practice, and the generated images are of low quality. In another work from the same author,[40] metric privacy is defined in the image transformation domain, but the obfuscated images are still of low quality. Lecuyer et al.[41] proposed the PixelDP framework, which includes a DP noise layer in the DNN. The PixelDP scheme forces the output prediction function to be DP, provided that the input changes on a small number of pixels (when the input is an image). However, the purpose of PixelDP is to increase robustness to adversarial examples rather than image privacy. Chen et al.[3] proposed a variant of DP by considering a perceptual similarity of the facial images, named perceptual indistinguishability (PI)-Net, which can achieve image obfuscation while ensuring PI.

To achieve a good trade-off between privacy and accessibility for face de-identification, reversible privacy protection has been studied in the literature.[42-44] Pan et al.[42] proposed a Multi-factor Modifier (MfM) based on conditional encoder and decoder framework, which achieves multi-factor facial de/re-identification. Based on a deep generative model, a personalized and reversible de-identification method was designed in Reference 43 to control the direction and degree of identity change by introducing a user-specific password and an adjustable parameter. You et al.[44] proposed a reversible privacy protection framework with an encoder and decoder using U-Net architecture to generate high-quality protected images without visible facial features. The original images are encoded with embedded face information before uploading onto the cloud.

Video-related de-identification has also been investigated in References 45-47. Unlike the face image de-identity, it requires to be modified seamlessly without causing any visual distortions or artifacts. In Reference 45, a multi-task extension of GAN was formulated to eliminate privacy-sensitive information of a video and detect privacy-preserving actions. In Reference 46, a feed-forward encoder-decoder network architecture was proposed conditioned on the high-level representation of a facial image. By coupling the latent space of the auto-encoder with a trained classifier network, a rich latent space with embedded identity and attribute information can be achieved.

Compared to the existing method for face image de-identification, to the best of our knowledge, ours is the first method that generates feature space AE noise in an optimization style. And compared to the state-of-the-art techniques, ours achieves compelling results in privacy, utility, and transferability.

3 | FEATURE SPACE ADVERSARIAL PERTURBATION BASED FACE IMAGE DE-IDENTIFICATION FRAMEWORK

In this section, we elaborate on the design of the proposed feature space adversarial perturbation (FSAP) based image de-identification framework. Under this framework, we further propose privacy protection methods against CNN face recognition.

3.1 | Problem formulation

Let $x \in \mathbb{R}^{h\times w\times c}$ denote a face image with c channels, each having a size of height h and width w. A CNN encoder $f_E(x)$ can generate a latent vector W of the face image x and a decoder $f_D(W)$ can reconstruct the face feature W to the output face image $\hat{x} \in \mathbb{R}^{h\times w\times c}$. If both the encoder and decoder are ideal, we should have $\hat{x} = x$, that is, $f_D(f_E(x)) = x$.

Our ultimate goal is to find a noise ΔW in the latent space that can perturb the identity of the input face image such that the output face image $\hat{x}$ has the following characteristics:

1. **De-identification.** The perturbed image $\hat{x}$ is likely to be recognized as a different face from the input image x by an arbitrary CNN face recognizer $f_I(\cdot)$, that is, $\mathcal{L}_I(x, \hat{x}; f_I) > \tau$, where $\mathcal{L}_I(\cdot)$ indicates the identity loss. τ denotes the threshold of the two face images being recognized as different identities.
2. **Maintaining the utility.** While targeting de-identifying the face identity, the perturbed image $\hat{x}$ should suffer from the minimum attribute loss to the input image x, that is, min $\mathcal{L}_P(x, \hat{x})$, where $\mathcal{L}_P(\cdot)$ indicates the attribute loss. To the naked eye, the perturbed image $\hat{x}$ should have similar properties to the input image x, so that it would be difficult for humans to distinguish which image is real or an impostor.

To summarize this process into an optimization problem, we have

$$\begin{aligned} &\min \ \mathcal{L}_P(x, \hat{x}) \\ &\text{s.t. } \mathcal{L}_I(x, \hat{x}; f_I) > \tau. \end{aligned} \tag{1}$$

The optimization problem (3.1) tries to maximize the dissimilarity of the face image identities, while minimizing the similarity of the face image attributes. To achieve this goal, we design novel architecture, referred to as the feature space adversarial perturbation based face image de-identification framework (FSAP framework for short).

3.2 | FSAP framework

The framework for generating feature-level adversarial examples is shown in Figure 2, which consists of three stages: (1) Stage 1 - encodes the input image x into the latent vector W by using a pre-trained CNN, that is, f_E. (2) Stage 2 - updates the image latent vector W' by adding adversarial perturbation iteratively. (3) Stage 3 - the output image $\hat{x}$ is reconstructed from W' by a pretrained decoder, that is, f_D. The output image $\hat{x}$ is an image that does not contain the identity information of the original image.

3.2.1 | Feature extraction

In order to extract the face image features, we adopt the pixel2style2pixel (pSp)[48] encoding framework, which can be used to solve various image-to-image translation tasks and is compatible with StyleGAN2 architecture. We use the intermediate layer between the encoder and the

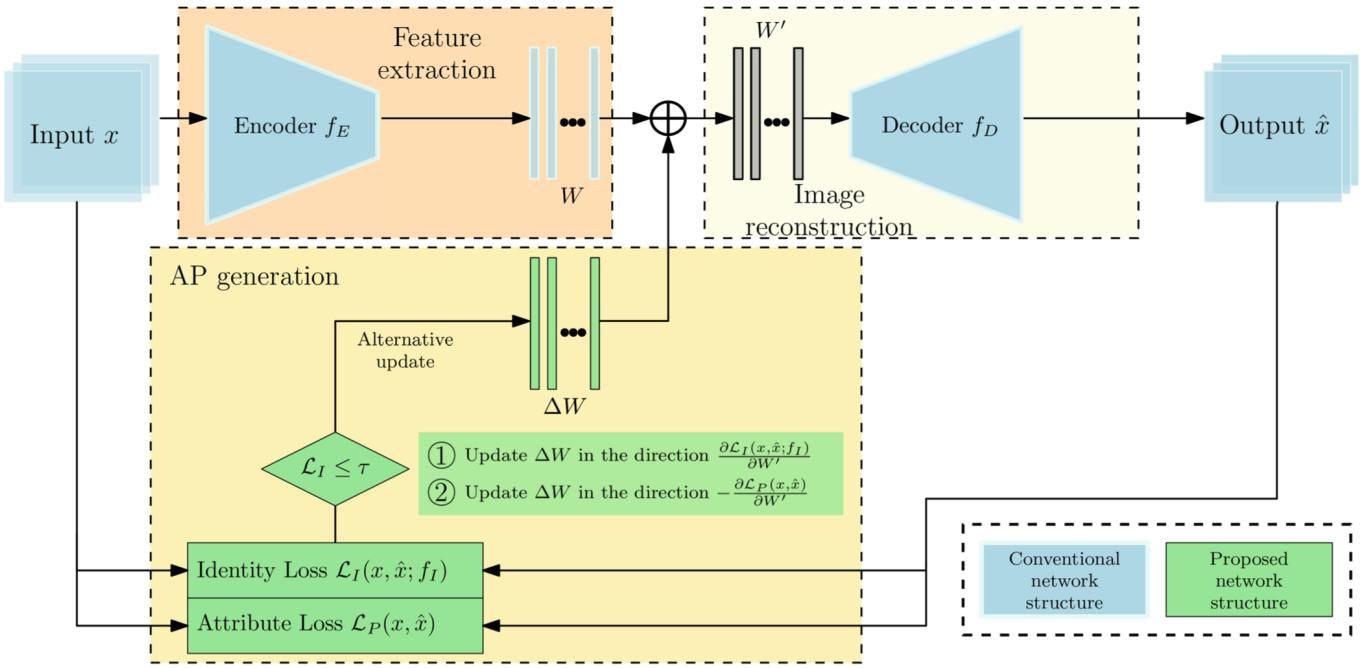


FIGURE 2 Feature space adversarial perturbation (FSAP) based privacy protection framework.

decoder as our latent code, denoted as W. Here, W contains 18 style vectors, with each vector of length 512. The encoder extracts the feature maps of the input image in three levels (low, medium, high) through a typical CNN (e.g., ResNet). These feature maps then were mapped to the latent vector. The process can be formulated as:

$$W = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_6 \\ z_7 \\ \vdots \\ z_{18} \end{bmatrix} = \begin{bmatrix} s_1(k_{high}(x)) \\ s_2(k_{high}(x)) \\ s_3(k_{medium}(x), k_{high}(x)) \\ \vdots \\ s_6(k_{medium}(x), k_{high}(x)) \\ s_7(k_{low}(x), k_{medium}(x), k_{high}(x)) \\ \vdots \\ s_{18}(k_{low}(x), k_{medium}(x), k_{high}(x)) \end{bmatrix} = f_E(x), \tag{2}$$

where $s_n(k), n \in [1, 2, \ldots, 18]$ is a fully convolutional network to map the feature maps k into the style vector z_n. The feature maps k have three different dimensions, that is, $\dim(k_{high}) < \dim(k_{medium}) < \dim(k_{low})$, and are built with a nested structure. $z_n \in \mathbb{R}^{512}, n \in [1, 2, \ldots, 18]$ are the style vectors corresponding to the 18 layers of the latent vector W. The input image x is an RGB face image, and f_E is the encoder.

3.2.2 | Adversarial perturbation generation

The latent vector W can be used to ideally reconstruct an output image $\hat{x}$ that is close to the original image x. We now start to train an AP to change the face identity, that is, train an AP (i.e., ΔW) that can generate a face ID loss above the threshold, while minimizing the face attribute loss.

The two loss items are calculated as described below:

1. *ID loss.* The ID loss $\mathcal{L}_I(x, \hat{x}; f_I)$ is to measure the identity similarity of the two faces. This loss function maps the input image x and the output image $\hat{x}$ into the face feature space, also known as face embedding. We adopt the most widely used method cosine similarity to compute the face embedding loss (i.e., ID loss), which is defined as:

$$\mathcal{L}_I(x, \hat{x}; f_I) = 1 - \frac{f_I(x) \cdot f_I(\hat{x})}{\|f_I(x)\|_2 \cdot \|f_I(\hat{x})\|_2}. \tag{3}$$

2. *Attribute loss.* The attribute loss $\mathcal{L}_P(x, \hat{x})$ measures the attribute similarity of the two faces. The loss function is a combination of three typical losses, including MSE ($\mathcal{L}_m$), LPIPS[49] ($\mathcal{L}_l$), and SSIM ($\mathcal{L}_s$), and is defined as follows:

$$\mathcal{L}_P(x, \hat{x}) := \{\lambda_1 \mathcal{L}_m(x, \hat{x}), \lambda_2 \mathcal{L}_l(x, \hat{x}), \lambda_3 \mathcal{L}_s(x, \hat{x})\}. \tag{4}$$

MSE ($\mathcal{L}_m$), Mean Square Error, takes the pixel loss of the input and output images, which controls the amount of noise added to the image. LPIPS[49] ($\mathcal{L}_l$), Learned Perceptual Image Patch Similarity, takes the perceptual loss from the perceptual latent distance of the input and output images and measures the perceptual similarity of the images, which controls the visual quality of the image. SSIM ($\mathcal{L}_s$), Structural Similarity Index Measure, controls the structural similarity of two images. The combination of these three losses can guarantee the utility of the image from different levels.

The details of the AP generation algorithm will be described in Subsection 3.3.

3.2.3 | Image reconstruction

We adopt the StyleGAN2[6] synthesis network as the generator. Unlike a traditional decoder, which uses the latent vector as the bottom layer of the network, StyleGAN2 generates the images from a constant vector. The latent vectors were fed to 18 layers of the network to affect the identity of the face. In order to de-identify the face image, the perturbed latent vector $W' = W + \Delta W$ is thus fed to each layer of the synthesis network as well. The process can be formulated as follows:

$$\hat{x} = f_D(W'), \tag{5}$$

where f_D is the reconstruction network that decodes the modified latent vector W' back to the RGB face image.

3.3 | Adversarial perturbation generation process

Traditional adversarial perturbations can be grouped into two major categories: target and non-target. Target attacks require that the model classifier misclassifies the AEs to a specific class for malicious purposes. The non-target attacks only require AEs to be misclassified with any wrong label to avoid detection. In the context of privacy protection, non-target AP just pushes the image away from the current identity. Therefore, it is a better option than target AP.

In addition, we want to minimize the AP so that the utility of the image can be kept as much as possible. While this is often a non-convex optimization problem. Some approximate methods have been developed. The fast gradient sign method (FGSM) proposed by Goodfellow[13] is a widely adopted method of generating AE/AP. The AP generated by FGSM and its variants is superior to other traditional methods when facing the white-box model. However, traditional FGSM has less transferability when facing the black-box model. Some studies[15,16] have found that adding noise to the feature space can improve the transferability of AP. The argument was that the existing recognition networks generally map the pictures to latent space vectors through the CNN to recognize images. Therefore, the noise added to the latent space vector will have better transferability. To sum up, our proposed method adds noise to the latent vector of the input image x in a non-target manner.

Also, our method differs from the conventional GAN-Based methods that add the latent noise in a vague direction. The conventional GAN-Based methods use a large dataset to train the network and one network to process all the data. Even with a large dataset and time-consuming training, the generalisation ability of the network is inversely proportional to the accuracy. Whereas our method can accurately add latent noise on the identity-related information of an individual image. The proposed latent noise is generated based on the gradient direction with regard to the two losses described before, that is, $\mathcal{L}_I$ and $\mathcal{L}_P$.

The perturbation ΔW takes the update from the following two losses alternately:

$$\Delta W_1 = \lambda_I \text{sign}\left(\frac{\partial \mathcal{L}_I(x, \hat{x}; f_I)}{\partial W'}\right); \tag{6}$$

$$\Delta W_2 = -\left(\lambda_1 \text{sign}\left(\frac{\partial \mathcal{L}_m(x, \hat{x})}{\partial W'}\right) + \lambda_2 \text{sign}\left(\frac{\partial \mathcal{L}_I(x, \hat{x})}{\partial W'}\right) + \lambda_3 \text{sign}\left(\frac{\partial \mathcal{L}_s(x, \hat{x})}{\partial W'}\right)\right), \tag{7}$$

where ΔW_1 and ΔW_2 are the latent noises regarding the identity and attributes of the input image, respectively. The first loss function, $\mathcal{L}_I(x, \hat{x}; f_I)$, measures the face identity dissimilarity between the input image x and the output image $\hat{x}$ with an arbitrary CNN face recognizer f_I. The second loss function, $\mathcal{L}_P(x, \hat{x})$, computes the face attribute loss between the input image x and the output image $\hat{x}$. When updating, the gradient is accumulated on the potential identity free face latent vector $W' = W + \Delta W$. It is worth noting that to simplify the experiments, we use the same attribute update rate λ_P to replace λ_1, λ_2, and λ_3. ΔW_2 can be rewritten to:

$$\Delta W_2 = -\lambda_P \left(\text{sign}\left(\frac{\partial \mathcal{L}_m(x, \hat{x})}{\partial W'}\right) + \text{sign}\left(\frac{\partial \mathcal{L}_I(x, \hat{x})}{\partial W'}\right) + \text{sign}\left(\frac{\partial \mathcal{L}_s(x, \hat{x})}{\partial W'}\right)\right), \tag{8}$$

The AP generation algorithm is shown in Algorithm 1.

4 | EXPERIMENTS

In this section, we carry out extensive experiments to verify the effectiveness of the proposed method. We also compare our method with the state-of-the-art face de-identification methods, that is, GAN-based: CIAGAN,[10] DeepPrivacy,[11] and AP-based: Fawkes.[21]

4.1 | Experiment setup

4.1.1 | Dataset

In this experiment, human faces are selected as the object of image privacy protection because they contain a large amount of identifiable information and have been the main concern in the field of image privacy protection. The face images for our experiments come from two well-known public face image datasets, that is, FFHQ[50] and CelebA.[51]

1. The FFHQ dataset contains 70,000 high-quality PNG images with a resolution of 1024 × 1024 and considerable variation in terms of age, ethnicity, and image background.
2. The CelebA dataset contains 202,599 face images covering large pose variations and background clutter.

Algorithm 1. AP generation algorithm based on the FSAP framework

Parameters: ID update rate: λ_I; Attribute update rate: λ_P; Maximum iteration number: N; Adjustable iteration number: k; ID distance threshold: τ.
Input: The original image x.
Output: The released privacy-preserving image $\hat{x}$.
Stage 1:
Obtain the latent vector $W = f_E(x)$.
Stage 2:
Initialization: latent noise $\Delta W = 0$; Perturbed latent vector $W'_0 = W$.
for $1 \leq n \leq N$ **do**
  $\hat{x}_n = f_D(W'_n)$;
  $W'_n = W'_n + \Delta W_1$;
  for $1 \leq i \leq k$ **do**
    $\hat{x}_i = f_D(W'_n)$;
    $W'_n = W'_n + \Delta W_2$;
    $i = i + 1$;
  end
  if $\mathcal{L}_I(x, \hat{x}_n; f_I) > \tau$ **then**
    Break
  end
  $n = n + 1$;
  $\hat{x} = \hat{x}_n$.
end

4.1.2 | Experimental settings

In this paper, we adopt a pSp encoder[48] pre-trained on the FFHQ dataset for feature extraction. The ID Discriminator f_I used in this paper is pre-trained on the state-of-the-art face recognition network ResNet[52] with arcface loss[53] on the real-life dataset. The synthesis network of StyleGAN2[6] is pre-trained on the FFHQ dataset.

Parameter settings: $N = 100$, $k = 6$, $\tau = 0.8$, $\lambda_I = 0.02$, $\lambda_P = 0.008$. τ is the threshold of the ID distance.

4.1.3 | Evaluation metrics

The following methods will be used to measure the proposed algorithm:

1. *De-identification*:
   Successful protection rate (**SPR**): We define successful protection as:

$$I(x, \hat{x}) > \delta, \tag{9}$$

where $I(x, \hat{x})$ is the identity distance calculated on the identifier I. δ is the threshold that recognizes the face as a different identity. SPR is then formulated as:

$$SPR = \frac{1}{m}\sum_{i=1}^{m} g_P(x_i), \tag{10}$$

where

$$g_P(x) = \begin{cases} 1, & \text{if } I(x, \hat{x}) > \delta, \\ 0, & \text{otherwise}, \end{cases} \tag{11}$$

with m being the number of tests.

2. *Utility*:
   1. (1) Successful detection rate (**SDR**): We defined SDR as the de-identification rate of face images that can be detected. It evaluates the utilities of computer vision tasks and is formulated as:

$$SDR = \frac{1}{m}\sum_{i=1}^{m} g_D(x_i). \tag{12}$$

   If the face can be detected, then $g_D = 1$. Otherwise, $g_D = 0$.
   2. (2) Landmarks distance. *chin/nose/eyes/mouth* indicates the mean distance of the key points corresponding to each facial area. It evaluates the utility of facial analysis.
3. *Distortion metrics*:

   Mean square error (**MSE**) is used to measure the distortion between two images at the pixel level.

4.2 | Performance evaluation

In this section, we display the results of our proposed method from three aspects. (i) Ablation Study; (ii) Results compared to other methods; (iii) Analysis of the Parameters.

4.2.1 | Ablation study

We conduct an ablation study on the framework to confirm the effectiveness of the proposed ID loss and attribute loss as introduced in Section 3.2. In particular, we consider the following cases: the framework is equipped with the ID loss module only.

It is worth noting that in order to protect the face identity with the ID loss module only, the perturbation ΔW adds on the gradient ascent direction of $\mathcal{L}_I$, that is, $\Delta W = \lambda_I \text{sign}(\frac{\partial \mathcal{L}_I(x,\hat{x};f_I)}{\partial W'})$, with $\text{sign}(\cdot)$ being a Sign function. In this case, the perturbation is added to the ID information without attribute constraints. The bi-loss mode images is generated based on Algorithm 1. Both the ID and attribute constraints are used to generate the perturbation.

Figure 3 gives the visual results of the proposed ablation study. The quantity result of the privacy evaluation method (SPR) and the utility evaluation methods (SDR and MSE) are reported in Table 1.

We use the framework of Face Recognition Library for the SPR, and dlib for the SDR. The ablation result shows that compared with the ID loss-only framework, the Bi-loss framework achieves a 0.4% increase in privacy performance and a 2.1% increase in face detection. Furthermore, the distortion of the Bi-loss framework has been reduced by 73.7%. In ablation experiments, the actual number of unsuccessfully protected samples is almost the same. However, since the Face Recognition Library first uses a face detection module to find faces in an image before figuring out the

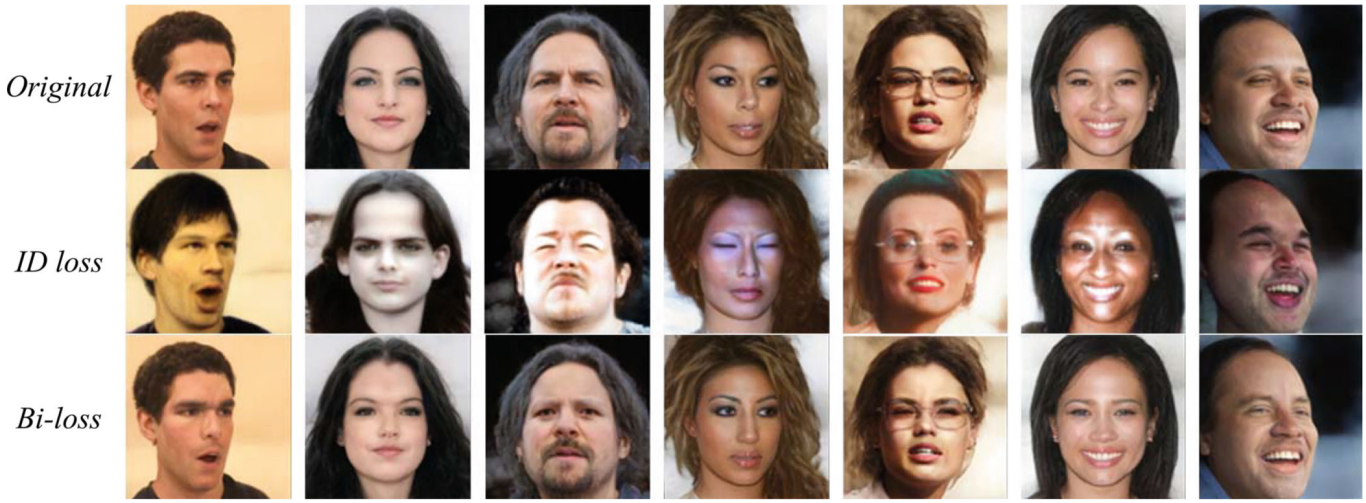


FIGURE 3 The visual results of the ablation study. The first row is the original image. The second row and third row are the de-identity images generated by ID loss and bi-loss framework, respectively.

TABLE 1 Ablation study

	SPR (↑)	SDR (↑)	MSE (↓)
ID loss solely	0.901	0.979	0.426
Bi-loss (ours)	**0.905**	**1**	**0.112**

Note: We use the same ID distance threshold, $\tau = 0.8$, for all settings in this table. The second column is the protection rate under Face Recognition Library. The third column is the detection rate by using *dlib*.[54] The hyperparameters are set to $\lambda_I = 0.02$, $\lambda_P = 0.008$, and the maximum iteration number $N = 100$.

TABLE 2 Privacy evaluation

	Face recognition (↑)	FaceNet (VGGFace2) (↑)
CIAGAN[10]	0.918	0.943
DeepPrivacy[11]	0.939	0.816
Fawkes[21]	0.704	0.564
Ours	**0.967**	**0.960**

Note: The values in this table are the successful protection rates (SPRs). The generation mode of Fawkes[21] is set to *high*, which is the highest privacy level authors recommended. The threshold of Face Recognition is $\delta = 0.6$ and the threshold of FaceNet is $\delta = 1.1$ according to Reference 55.

TABLE 3 Utility evaluation

		Landmarks distance				
	SDR (↑)	Chin (↓)	Nose (↓)	Eyes (↓)	Mouth (↓)	MSE (↓)
Original	1	0	0	0	0	0
CIAGAN[10]	0.9939	2.635	2.130	2.422	2.622	0.131
DeepPrivacy[11]	0.9989	2.070	1.631	1.384	2.712	0.344
Fawkes[21]	0.9990	0.720	**0.3921**	**0.389**	0.492	0.422
Ours	**1**	**0.704**	0.6664	0.484	**0.375**	**0.112**

Note: The face detection network used in this table is *dlib*. The landmarks distances are generated under the face recognition library.

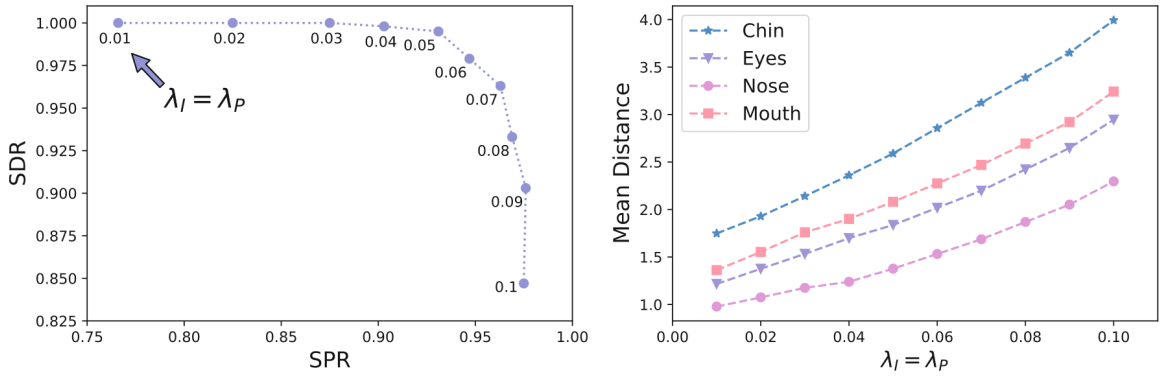
face ID for that image, the images protected by the ID-loss only sometimes make the image invisible to the face detection network, which lowers the protection rate.

4.2.2 | Results compared to other de-identification methods

This section compares our method with the state-of-the-art face de-identification methods.

Table 2 shows the privacy protection evaluation results with the widely used face recognition networks. We use both the Face Recognition Library and the FaceNet[55] network trained on VGGFace2 to evaluate the SPR. The results prove that our method can better de-identify the face image under the most widely used face recognition methods. Our method is better than CIAGAN, DeepPrivacy, and Fawkes by 4.9%, 2.8%, and 26.3%, respectively, in terms of the SPR under Face Recognition Library. Under FaceNet (VGGFace2) network, Ours (thick) improves the SPR by 1.7%, 14.4%, and 39.6% compared to CIAGAN, DeepPrivacy, and Fawkes, respectively.

Table 3 summarizes the utility performance of our method compared with the state-of-the-art methods. We evaluate our method with the SDR, MSE and the average distance of the face feature landmarks on the pixel level. The result shows that our approaches achieve the highest SDR. In other words, our de-identified faces lead to better performance for face detection tasks. Moreover, compared with GAN-Based methods, that is, CIAGAN,[10] DeepPrivacy,[11] our methods strike a compelling score on minimizing the distance of each facial feature. Our method lowers the average face feature distance of that on CIAGAN[10] and DeepPrivacy[11] by 77% and 71%, respectively. Both our method and Fawkes[21] have the lower average face feature distance, but Fawkes achieves the score at the expense of privacy-preserving effectiveness (lowest privacy score). In addition, our method achieves the lowest pixel-level distortions. Compared with CIAGAN, DeepPrivacy and Fawkes, our distortion is decreased by 14.5%, 67.4% and 73.5% respectively.



(A) The trade-off between privacy and utility under different parameters. **(B)** The analysis between the parameters and face features.

FIGURE 4 Parameters analysis. The values of λ_I (or λ_P) range from 0.01 to 0.1 with a step size of 0.01. The maximum iteration number $N = 30$. The threshold $\tau = 0.8$.

Taking into account the above all, our method has better performance in protecting the face ID while minimizing the impact on image utility. The images being protected by our method can be used in multi-image tasks.

4.2.3 | Analysis on the Hyperparameters

To evaluate the influence of the hyperparameters, we evaluate the perturbation performance with different controllable noise coefficients, that is, λ_I and λ_P. Figure 4A shows a trade-off between the SPR and the SDR under different values of λ_I and λ_P. We see that the SDR first decreases slowly when $\lambda_I = \lambda_P < 0.05$, and then decreases rapidly when $\lambda_I = \lambda_P \geq 0.05$. In contrast, the SPR increases rapidly when $\lambda_I = \lambda_P < 0.05$, and then flattens out at $\lambda_I = \lambda_P = 0.08$. In other words, with the growth of λ_I and λ_P, the success rate of protection (privacy) increases while the success rate of detection (utility) decreases.

Figure 4B gives the mean feature distance under different hyperparameter settings. The result shows that the distance of the feature will increase when the hyperparameter increase. From the first point 0.01 to the last point 0.1, the average distance increases 135%.

4.3 | Discussions

4.3.1 | Privacy protection against the commercial network

In this section, we test our protected images on the two most widely used commercial networks, Microsoft Face API[56] and Face Plus Plus API.[57] These two APIs are built on large face recognition networks, which use the advanced deep neural network and are trained on a large dataset. They provide several applications, including face detection and analysis, identity verification and finding similar faces. In this experiment, we use the identity verification service to evaluate our method. In identity verification, the APIs will take the original images and the protected images as the input and output a score of confidence that indicates the probability that two faces belong to the same person. The higher the confidence, the higher chance they belong to the same person. If the input is the images without protection and the original image, the confidence score equals 1.0.

The experiment results are shown in Figure 5. Microsoft Face API and Face Plus Plus use different confidence thresholds. While Microsoft Face API takes 0.5, Face Plus Plus takes the value of around 0.69. The sample with a score below the threshold is recognized as a different person in the API. The results show that our method lowers the score of 82.9% samples under the threshold of 0.5 against Microsoft Face API, which makes the network almost ineffective. And the score of 55.8% samples on Face Plus Plus API is under the threshold of 0.69, which makes the network works in a random guess. The experiment result proves that our method is transferable in different networks. Because these commercial APIs networks do not open their source code and we did this experiment in a black box mode.

4.3.2 | Limitations

Although our proposed framework for face identity protection can achieve compelling results in both privacy and utility, it has some room for further improvement. First, the face latent code generated by the auto-encoder architecture comes with a cost - the de-identity image quality is limited

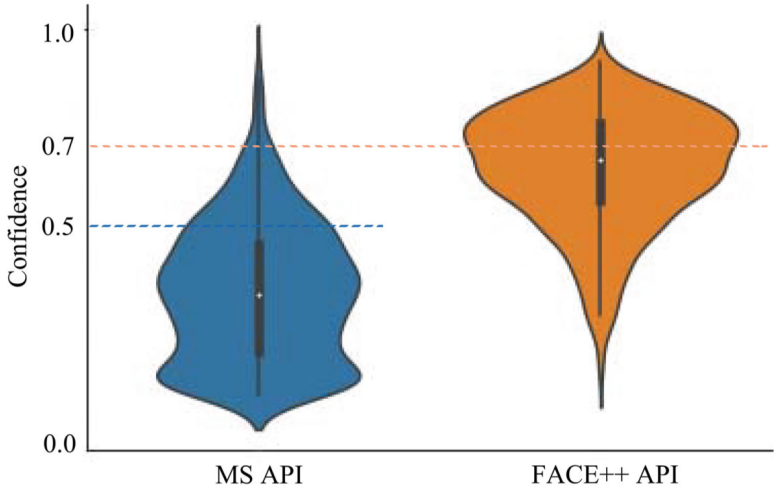


FIGURE 5 The confidence of commercial API.

FIGURE 6 Failure examples. Output 1 refers to the generated images without protection. Output 2 shows the de-identity image with our method.

to the image-to-image translation ability of the chosen auto-encoder. Thus, de-identifying the face image with complex attributes, for example, background, might be challenging if such examples were not synthesized well in the auto-encoder. Figure 6 presents a few examples of such images.

As can be seen in Figure 6, the complex attributes, for example, background and hands, which are non-identity-related attributes, can not display correctly due to the reconstruction failure.

5 | CONCLUSION

In this paper, we propose a novel face image de-identification framework. This framework de-identifies the face by adding feature space adversarial perturbation (FSAP). Moreover, we conduct intensive experiments to prove the effectiveness of the framework. With the latent vector W trained on the elaborate loss, the perturbed faces are equipped to reduce the risks of identity leakage under CNN face recognition technics while balancing the utility for computer vision tasks. The merits of the proposed framework are two-folded. First, compared with the GAN-based face de-identification network, instead of adding perturbations in a generalised direction, FSAP adds noise on the gradient, which ensures accuracy. Second, compared with the AP-based face de-identification network, feature-space adversarial perturbations have better transferability among different neural network models.

ACKNOWLEDGMENT

This work is supported by an ARC Linkage Project (LP180101150) from the Australian Research Council, Australia. Open access publishing facilitated by University of Technology Sydney, as part of the Wiley - University of Technology Sydney agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

No potential conflict of interest was reported by the authors.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request. The data that support the findings of this study are available in FFHQ at https://github.com/NVlabs/ffhq-dataset. These data were derived from the following resources available in the public domain: - images 1024 × 1024, https://drive.google.com/drive/folders/1tZUcXDBeOibC6jcMCtgRRz67pzrAHeHL.

ORCID

Hanyu Xue https://orcid.org/0000-0002-5150-1513
Bo Liu https://orcid.org/0000-0002-3603-6617
Tianqing Zhu https://orcid.org/0000-0003-3411-7947

REFERENCES

1. Xue H, Liu B, Din M, Song L, Zhu T. Hiding private information in images from AI. Paper presented at: ICC 2020 - 2020 IEEE International Conference on Communications (ICC); 2020; 1-6. doi: 10.1109/ICC40277.2020.9148656
2. Gu X, Luo W, Ryoo MS, Lee YJ. Password-conditioned anonymization and deanonymization with face identity transformers. Paper presented at: European Conference on Computer Vision; 2020; 727–743. doi: 10.1007/978-3-030-58592-1_43
3. Chen JW, Chen LJ, Yu CM, Lu CS. Perceptual indistinguishability-net (PI-Net): facial image obfuscation with manipulable semantics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021; 6478–6487.
4. Yeh RA, Chen C, Yian Lim T, Schwing AG, Hasegawa-Johnson M, Do MN. Semantic image inpainting with deep generative models. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017; 5485–5493.
5. Mirza M, Osindero S. Conditional generative adversarial nets. ArXiv preprint arXiv:1411.1784; 2014.
6. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of stylegan. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; 8110–8119.
7. Cao J, Liu B, Wen Y, et al. Hiding among your neighbors: face image privacy protection with differential private k-anonymity. 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB); 2022; 1-6. doi: 10.1109/BMSB55706.2022.9828699
8. Zhang K, Tian J, Xiao H, Zhao Y, Zhao W, Chen J. A numerical splitting and adaptive privacy budget allocation based LDP mechanism for privacy preservation in blockchain-powered IoT. *IEEE Internet Things J*. 2022;1. doi:10.1109/JIOT.2022.3145845
9. Hassan MU, Rehmani MH, Chen J. Anomaly detection in blockchain networks: a comprehensive survey. *IEEE Commun Surv Tutor*. 2022;1. doi:10.1109/COMST.2022.3205643
10. Maximov M, Elezi I, Leal-Taixe L. CIAGAN: conditional identity anonymization generative adversarial networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. doi: 10.1109/cvpr42600.2020.00549
11. Hukkelås H, Mester R, Lindseth F. Deepprivacy: a generative adversarial network for face anonymization. Paper presented at: International Symposium on Visual Computing; 2019; 565–578.
12. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. ArXiv preprint arXiv:1312.6199; 2013.
13. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. ArXiv preprint arXiv:1412.6572; 2014.
14. Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples. ArXiv preprint arXiv:1707.07397; 2017.
15. Huang Q, Katsman I, He H, Gu Z, Belongie S, Lim SN. Enhancing adversarial example transferability with an intermediate level attack. Proceedings of the IEEE International Conference on Computer Vision; 2019; 4733–4742.
16. Inkawhich N, Wen W, Li HH, Chen Y. Feature space perturbations yield more transferable adversarial examples. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019; 7066–7074.
17. Pidhorskyi S, Adjeroh DA, Doretto G. Adversarial latent autoencoders. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; 14104–14113.
18. Liu B, Ding M, Zhu T, Xiang Y, Zhou W. Adversaries or allies? privacy and deep learning in big data era. *Concurr Comput Pract Exp*. 2019;31(19):e5102.
19. Liu B, Xiong J, Wu Y, Ding M, Wu CM. Protecting multimedia privacy from both humans and AI. Paper presented at: Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting; 2019.
20. Oh SJ, Fritz M, Schiele B. Adversarial image perturbation for privacy protection a game theory perspective. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV); 2017; 1491–1500.
21. Shan S, Wenger E, Zhang J, Li H, Zheng H, Zhao BY. Fawkes: protecting privacy against unauthorized deep learning models. Paper presented at: 29th USENIX Security Symposium (USENIX Security 20); 2020; 1589–1604.
22. Li T, Lin L. AnonymousNet: natural face de-identification with measurable privacy. Paper presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2019; 56-65.
23. Rajabi A, Bobba RB, Rosulek M, Wright CV. Feng Wc. on the (Im) practicality of adversarial perturbation for image privacy. *Proc Privacy Enhanc Technol*. 2021;2021(1):85-106.

24. Zhang G, Liu B, Zhu T, Zhou A, Zhou W. Visual privacy attacks and defenses in deep learning: a survey. *Artif Intell Rev*. 2022;55:4347–4401. doi:10.1007/s10462-021-10123-y
25. Zhao Y, Chen J. A survey on differential privacy for unstructured data content. *ACM Comput. Surv.* 2022;54(10):1-28. doi:10.1145/3490237
26. Zhao Y, Yuan D, Du JT, Chen J. Geo-ellipse-indistinguishability: community-aware location privacy protection for directional distribution. *IEEE Trans Knowl Data Eng*. 2022;1-11. doi:10.1109/TKDE.2022.3192360
27. McPherson R, Shokri R, Shmatikov V. Defeating image obfuscation with deep learning. ArXiv preprint arXiv:1609.00408; 2016.
28. Oh SJ, Benenson R, Fritz M, Schiele B. Faceless person recognition: privacy implications in social media. *Eur Conf Comput Vis*. 2016;9907:19-35.
29. Sun Q, Ma L, Joon Oh S, Van Gool L, Schiele B, Fritz M. Natural and effective obfuscation by head inpainting. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018; 5050–5059.
30. Wu Y, Yang F, Xu Y, Ling H. Privacy-protective-GAN for privacy preserving face de-identification. *J Comput Sci Technol*. 2019;34(1):47-60.
31. Sun Q, Tewari A, Xu W, Fritz M, Theobalt C, Schiele B. A hybrid model for identity obfuscation by face replacement. Proceedings of the European Conference on Computer Vision (ECCV); 2018; 553–569.
32. Li L, Bao J, Yang H, Chen D, Wen F. Faceshifter: Towards high fidelity and occlusion aware face swapping. ArXiv preprint arXiv:1912.13457; 2019.
33. Wang HP, Orekondy T, Fritz M. InfoScrub: towards attribute privacy by targeted obfuscation. ArXiv preprint arXiv:2005.10329; 2020.
34. Zhao Y, Liu B, Zhu T, Ding M, Zhou W. Private-encoder: enforcing privacy in latent space for human face images. *Concurr Comput Pract Exp*. 2022;34(3):e6548. doi:10.1002/cpe.6548
35. Deb D, Zhang J, Jain AK. AdvFaces: adversarial face synthesis. Paper presented at: 2020 IEEE International Joint Conference on Biometrics (IJCB); 2020; 1-10. doi: 10.1109/IJCB48548.2020.9304898
36. Cherepanova V, Goldblum M, Foley H, et al. LowKey: leveraging adversarial attacks to protect social media users from facial recognition. Paper presented at: International Conference on Learning Representations; 2021.
37. Zhang J, Sang J, Zhao X, Huang X, Sun Y, Hu Y. Adversarial privacy-preserving filter. Proceedings of the 28th ACM International Conference on Multimedia; 2020; 1423–1431. doi: 10.1145/3394171.3413906
38. Wen Y, Liu B, Ding M, Xie R, Song L. IdentityDP: differential private identification protection for face images. *Neurocomput*. 2022;501:197-211. doi:10.1016/j.neucom.2022.06.039
39. Fan L. Image pixelization with differential privacy. Paper presented at: IFIP Annual Conference on Data and Applications Security and Privacy; 2018; 148–162.
40. Fan L. Practical image obfuscation with provable privacy. Paper presented at: 2019 IEEE International Conference on Multimedia and Expo (ICME); 2019; 784–789.
41. Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S. Certified robustness to adversarial examples with differential privacy. Paper presented at: 2019 IEEE Symposium on Security and Privacy (SP); 2019: 656–672.
42. Pan YL, Chen JC, Wu JL. A multi-factor combinations enhanced reversible privacy protection system for facial images. Paper presented at: 2021 IEEE International Conference on Multimedia and Expo (ICME); 2021; 1–6.
43. Cao J, Liu B, Wen Y, Xie R, Song L. Personalized and invertible face de-identification by disentangled identity information manipulation. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021; 3334–3342.
44. You Z, Li S, Qian Z, Zhang X. Reversible Privacy-Preserving Recognition. Paper presented at: 2021 IEEE International Conference on Multimedia and Expo (ICME); 2021; 1-6. doi: 10.1109/ICME51207.2021.9428115
45. Ren Z, Lee YJ, Ryoo MS. Learning to anonymize faces for privacy preserving action detection. Proceedings of the European Conference on Computer Vision (ECCV); 2018; 620–636.
46. Gafni O, Wolf L, Taigman Y. Live face de-identification in video. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019; 9378–9387.
47. Wen Y, Liu B, Cao J, Xie R, Song L, Li Z. IdentityMask: deep motion flow guided reversible face video de-identification. *IEEE Trans Circ Syst Video Technol*. 2022;1. doi:10.1109/TCSVT.2022.3191982
48. Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: a StyleGAN encoder for image-to-image translation. Paper presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021.
49. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. Paper presented at: CVPR; 2018.
50. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. ArXiv preprint; 2018. doi: 10.48550/ARXIV.1812.04948
51. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. Proceedings of International Conference on Computer Vision (ICCV); 2015.
52. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. ArXiv preprint arXiv:1512.03385; 2015.
53. Deng J, Guo J, Xue N, Zafeiriou S. Arcface: additive angular margin loss for deep face recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019; 4690–4699.
54. King DE. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*. 2009;10:1755-1758.
55. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. Paper presented at: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. doi: 10.1109/cvpr.2015.7298682
56. Microsoft. Facial Recognition | Microsoft Azure. Azure.microsoft.com 2022.
57. Faceplusplus. Face++ - Face++ Cognitive Services. Faceplusplus.com 2022.

How to cite this article: Xue H, Liu B, Yuan X, Ding M, Zhu T. Face image de-identification by feature space adversarial perturbation. *Concurrency Computat Pract Exper*. 2022;e7554. doi: 10.1002/cpe.7554