Multimedia Tools and Applications (2023) 82:15535–15551
https://doi.org/10.1007/s11042-022-13917-6

Semantic-aware deidentification generative adversarial networks for identity anonymization

Hyeongbok Kim[1] · Zhiqi Pang[1] · Lingling Zhao[1] · Xiaohong Su[1] · Jin Suk Lee[2]

Received: 7 February 2022 / Revised: 25 April 2022 / Accepted: 12 September 2022/
Published online: 7 October 2022

Abstract

Privacy protection in the computer vision field has attracted increasing attention. Generative adversarial network-based methods have been explored for identity anonymization, but they do not take into consideration semantic information of images, which may result in unrealistic or flawed facial results. In this paper, we propose a Semantic-aware De-identification Generative Adversarial Network (SDGAN) model for identity anonymization. To retain the facial expression effectively, we extract the facial semantic image using the edge-aware graph representation network to constraint the position, shape and relationship of generated facial key features. Then the semantic image is injected into the generator together with the randomly selected identity information for de-Identification. To ensure the generation quality and realistic-looking results, we adopt the SPADE architecture to improve the generation ability of conditional GAN. Meanwhile, we design a hybrid identity discriminator composed of an image quality analysis module, a VGG-based perceptual loss function, and a contrastive identity loss to enhance both the generation quality and ID anonymization. A comparison with the state-of-the-art baselines demonstrates that our model achieves significantly improved de-identification (De-ID) performance and provides more reliable and realistic-looking generated faces. Our code and data are available on https://github.com/kimhyeongbok/SDGAN

Keywords Deep learning · Generative adversarial networks · Image generation · Identity anonymization

✉ Xiaohong Su
sxh@hit.edu.cn

Jin Suk Lee
jslee@testworks.co.kr

[1] Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China

[2] Testworks, Inc., Seoul, South Korea

1 Introduction

The success of computer vision methods based on deep learning [10, 11, 40] requires a large amount of training data. A large number of private images have been shared on various application platforms [21, 48], which has aroused people's concerns about personal privacy and security. For example, the General Data Protection Regulation (GDPR) is in force in Europe, requiring organizations to define privacy policies based on user preferences [5]. In this case, it is useful to use tools to support users in understanding how their sensitive data are exchanged [2]. However, existing computer vision tasks, such as person reidentification [31, 32] and action recognition [4, 49], do not require clear face information. Therefore, we can use anonymization technology to process faces when publishing these data.

Facial de-identification (De-ID) techniques [14, 35, 39] thus came into being; these approaches aim to remove one person's identity by replacing the real face with a generated or simulated face while keeping the original head pose, facial expression and background unchanged. As facial De-ID plays a vital role in privacy protection, it has attracted extensive attention.

In this context, the earliest De-ID techniques obfuscated privacy-sensitive identity information via image distortion operations, such as mosaics and image blur. However, these methods also destroy any privacy-insensitive information and decrease the visual quality and authenticity of the images or videos, thus compromising their utility [8]. Other related approaches replace the faces in the image to be processed by finding new face images in a predefined reference image set. Their disadvantage is that the quality of the resulting image depends heavily on the selected reference image, and they cannot protect the face identity privacy in the reference image. Segmentation-based methods [36] are also used for anonymizing faces, but these methods often make faces undetectable. [14, 35] have the problem of insufficient erasure. On the other hand, even though the images generated by some methods [6] can fool recognition systems, they can be easily recognized by humans.

Recently, more sophisticated De-ID techniques were proposed based on generative adversarial networks (GANs) [7]. The conditional identity anonymization GAN (CIAGAN) [24] is a state-of-the-art method for private identification protection. However, the CIAGAN ignores the attribute information contained in the face identity features and only retains the attribute information through the key points of the face, which leads to many visual defects in the results.

To overcome these problems, a new face De-ID and generation framework based on a GAN and a semantic image [44] is proposed to realize high-quality and controllable face De-ID. Specifically, to preserve the basic attribute information of the original image, we use a semantic image instead of a landmark to guide the generation process. To prevent the normalization layers from washing out the information contained in the image, the semantic image is input into the generator with spatially adaptive normalization from [34]. By concatenating the latent space with the representation layer of the face classifier, we achieve a rich latent space, embedding both identity and expression information. In addition, a structural similarity index (SSIM) loss [45] and a perceptual loss [17] are added to the objective function to improve the quality of the generated image.

Our contributions in this work are fourfold.

- Facial semantic image extraction is exploited to maintain the key pose and expression of the original face.
- We integrate the features of the new identity into the original features by leveraging the output of the feature representation layer of a pretrained face classification model.

- We design a hybrid identity discriminator composed of an image quality analysis module, a Visual Geometry Group (VGG)-based perceptual loss function, and a contrastive identity loss to guide the identity anonymization process.
- A large number of experiments have shown that our approach surpasses most existing techniques in the De-ID field.

2 Related work

2.1 Conventional face De-ID methods

Face anonymity technology aims to protect the private information of faces. Traditional methods change directly the data distribution of original images for face deidentification [1, 29]. The problems with these methods lie in that all the objects in an image are blacked out, blurred or pixelated independently, and the De-ID efficiency cannot be guaranteed since the fixed image operations can be easily reconstructed [8].

Except for simple image processing operations such as image blurring, pixelation or adding random noise, the K-same method [30] improved by the k-anonymity algorithm [41] generates face images by calculating the average value of k in the dataset. This ensures the visual privacy protection but often contains "ghosting" artifacts [14]. Therefore, as the variants of the K-same scheme continue to emerge in the literature, they focused mainly on preserving important attribution information in the original images or improving the naturalness of resultant faces [18, 25].

2.2 Deep learning-based face De-ID methods

With the development of deep learning based computer vision, generative adversarial networks and their variants have been extended for the privacy De-IDentification. GAN has natural advantage due to its strong generation ability according to the guidance of discriminator. It inspires designing frameworks that generate realistic image samples via adversarial training. GAN has become the current main trend for the research on face De-IDentification.

GANs realize face deidentification by generating image pixels instead of deleting or modifying information. According to whether additional reference faces are used, the existing GAN based methods can be divided into two kinds: (1) one-to-one methods and (2) many-to-one methods. In the framework of one-to-one generation methods, no reference faces are used; thus, privacy leakage of reference faces does not exist. Among the one-to-one generation methods, some researchers proposed the privacy-preserving GAN (PPGAN) [47], which forces the removal of the identity from the identity-related feature space performed by the pretrained discriminator. Meanwhile, visual correspondence is maintained by the similarity of the pixel horizontal structure. However, the PPGAN tends to generate images with unique facial features, which leads to low image generation quality. Another one-to-one generation method, called DeepPrivacy, [14] has achieved good performance. This method first masks the face, and then generates it under the guidance of landmarks. Fawkes [38] aims to anonymize identity without changing the visual perception of the face. Different from previous methods, Gu et al. [9] proposed a method that can anonymize and deanonymize at the same time. In addition, this method no longer relies on the mask to eliminate the original identity but directly modifies the original face.

In the many-to-one method, at least one reference face is exploited to fuse the attribute or ID information in the generated faces. Conditional GANs (cGANs) [16, 27] have become popular tools for controlling the appearance of synthesized data. Meden et al. [26] proposed a generation model that can generate anonymous faces based on K faces with different identities closest to the input face. CIAGAN [24] takes the existing face image, landmark, masked face and specified identity as input, trying to generate a face image with a new identity. The controllable face anonymization network (CFA-Net) [22] aims to control the anonymous process by operating the identity vector in the feature space. This method can generate all kinds of new faces highly similar to the original image content.

3 The semantic-aware deidentification GAN

3.1 Overview

As illustrated in Fig. 1, our model takes an image, the corresponding semantic image, the masked face and an identity feature as inputs. We aim to erase the identifiable features in the facial image and preserve the other attributes of the original image, including its pose, expression, and background.

Our model consists of the following three blocks: Block I for semantic image extraction, Block II for identity transformation, and Block III for face generation with SPADE.

Block I aims to perform facial segmentation via facial semantic image extraction. We propose the use of a semantic image to maintain the key pose and expression of the original face. Block II provides randomly selected identity information extracted from the given face dataset. Block III aims to generate an anonymized face image according to the original face and new identity information. In this component, the generator is an encoder-decoder model where the encoder embeds the original image information into a low-dimensional space. Then, the decoder decodes the combined information of the source image, identity features and semantic image into a generated image. In addition, we use spatially adaptive (DE) normalization (SPADE) [34] to enhance the conditional GAN so that it can produce realistic-looking results. Furthermore, the identity discriminator ensures that the generated images are anonymized to the greatest extent possible.

In addition, the identity information represented by the feature layer extracted from a pretrained face classifier is fused in the generator, while adjustment by the identity discriminator ensures that the generated images are anonymized to the greatest extent possible.

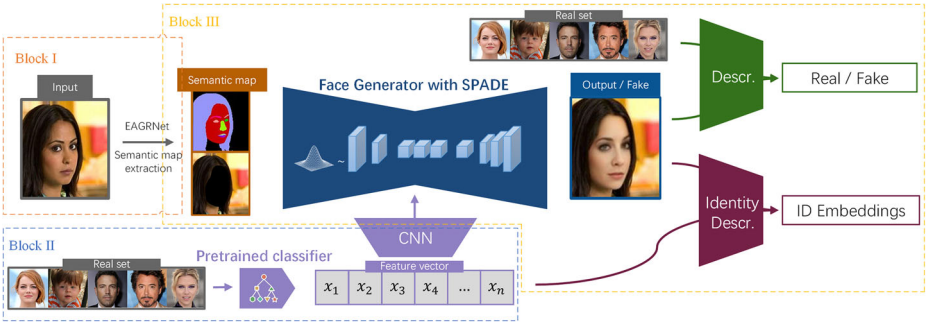

Fig. 1 The framework of the proposed model

3.2 Facial semantic image extraction

It has been proven that compared with generation methods based on random noise, due to the guidance of semantic information, generation methods based on semantic maps [16] can obtain higher-quality images. In addition, previous studies [46] have proven that correlations are present between facial between components, and it is difficult to mine these correlations in a generation model. This leads to randomly generated facial images that can be easily identified. To effectively preserve the basic attributes and the correlations between the facial components in the original face, the edge-aware graph representation network (EAGRNet) [44] is introduced into our model. The EAGRNet models the relationships between regions by learning graphical representations of facial images. In addition, it can capture long-distance correlations in facial images.

As illustrated in Fig. 2, the semantic image extraction process of the EAGRNet includes the following three stages: feature and edge extraction, edge-aware graph reasoning, and semantic decoding. In the feature and edge extraction phase, the EAGRNet takes the residual network (ResNet) [43] as the backbone to extract features at low levels and high levels for multiscale representation. Additionally, a spatial pyramid pooling operation is exploited to learn multiscale contextual information. Pyramid pooling outputs 16×16 size feature map. Furthermore, an edge perception module is constructed to acquire an edge map for the subsequent module. Edge perceiving module outputs a 32×32 size feature map.

Then, to build the long-range relations among facial components, the feature map and edge map are fed into the edge-aware graph reasoning module (EAGR module in Fig. 2). In the EAGR module, to learn intrinsic graph representations, the graph is projected into a collection of pixels that tend to reside in the same facial component to K ($K \geq 1$) vertices in the graph. Accordingly, the original features are projected onto vertices in an edge-aware fashion, the relations between the vertices (regions) are reasoned over the graph, and the learned graph representation is projected back to pixel grids, leading to a refined feature map with the same size as the original.

Finally, in the semantic decoding stage, EAGRNet designs a two-way decoder, both of which are based on 32×32 size feature map as input. The decoder combines the feature maps of the two paths to generate the final result of face parsing.

3.3 Identity transformation

The traditional encoder-decoder structure easily learns the reconstruction ability, which leads to anonymization failure. To realize the anonymization of the original identity, we integrate the features of the new identity into the original features. Different from the CIA-GAN [24], as shown in Block II in Fig. 1, we exploit a pretrained face classification model and leverage the output of the feature representation layer as the identity attribute. In this way, we achieve a rich latent space, embedding both identity and expression information. Based on the new identity features, the generator can learn the features of the reference image, not just the identity, to anonymize the original image.

3.4 Face generator with SPADE

The data distribution of the semantic image may result in the failure of traditional convolutional networks because their normalization layers tend to remove information contained in the input semantic masks. Inspired by SPADE [34], we employ spatially adaptive normalization to replace the traditional normalization layer. As shown in Fig. 2, the face generator

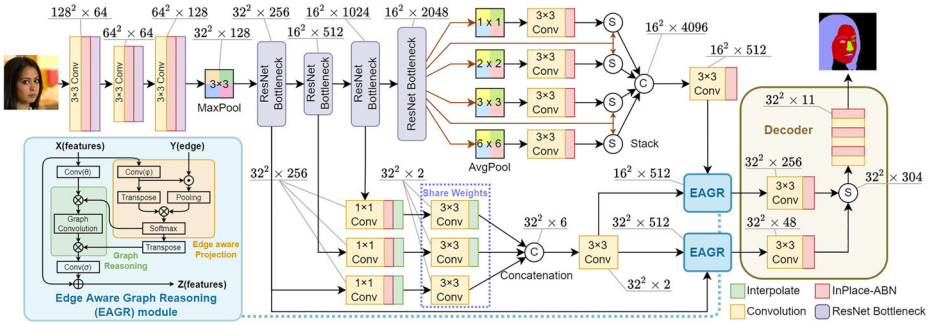

Fig. 2 The overall framework of EAGRNet

is built with the semantic image, the masked image and the identity feature from the pre-trained identity classifier as inputs. The masked image is reshaped by a CNN and then fed into a SPADE ResBlk along with the semantic image, where SPADE ResBlk is a residual block with the SPADE. To obtain a reasonable feature dimension, we downsample the output matrix by stacking 4 SPADE ResBlks and then concatenate the output with the reshaped identity feature. This concatenated feature vector is input into 4 ResNet blocks, 4 SPADE ResBlks with upsampling layers, and a convolutional layer again to generate a facial image matching the spatial resolution.

The spatially adaptive denormalization structure is shown in Fig. 3. The relationship between the output l^i_{out} and the input l_{in} of the module is defined as:

$$l^i_{out} = \alpha^i_{c,h,w} l_{in} + \beta^i_{c,h,w}, \tag{1}$$

where $\alpha^i_{c,h,w}$ and $\beta^i_{c,h,w}$ are modulation parameters, and the channel, height and width are (c, h, w), respectively. This conditional normalization layer modulates the activation process by using input semantic layouts through a spatially adaptive learned transformation and can effectively propagate the semantic information throughout the network (Fig. 4).

A discriminator network is used to differentiate between the generated face images and the real images. In this paper, we employ the least-squares GAN (LSGAN) [23] to train our face identity anonymization and generation network in an adversarial manner. The LSGAN

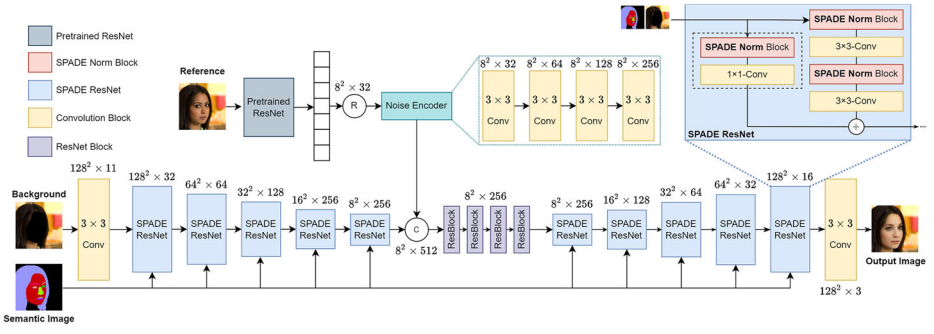

Fig. 3 Generator based on SPADE

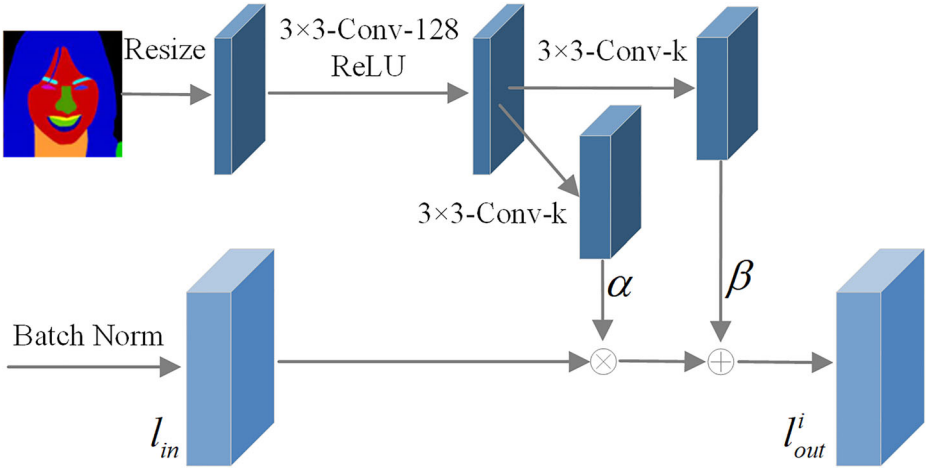

Fig. 4 Spatially adaptive denormalization

can boost training stability and produce more realistic images than the regular GAN [7]. The LSGAN loss functions for the discriminator and generator are:

$$L(D) = \frac{1}{2}E_{x\sim p_I(x)}[(D(x)-1)^2] + \frac{1}{2}E_{z\sim p_z(z)}[D(G(z))^2], \tag{2}$$

$$L(G) = \frac{1}{2}E_{z\sim p_z(z)}[(D(G(z))-1)^2], \tag{3}$$

where p_I is the distribution of the real face images and p_z is the distribution of the latent variable z. The adversarial loss L_{adv} is computed as follows:

$$L_{adv} = L(G) + L(D) \tag{4}$$

The SSIM was originally proposed for image quality analysis to overcome the limitations of the mean squared error (MSE). The SSIM is utilized here to measure the structural similarity between two images. SSIM is defined as:

$$SSIM(G(x), y) = \frac{2\mu_{G(x)}\mu_y + C_1}{\mu^2_{G(x)} + \mu^2_y + C_1} \cdot \frac{2\sigma_{G(x)y} + C_2}{\sigma^2_{G(x)} + \sigma^2_y + C_2}, \tag{5}$$

where μ_x and σ^2_x are the average value and the variance of x, respectively. σ_{xy} is the covariance of x and y. C_1 and C_2 are constants used to maintain stability. The SSIM ranges from 0 to 1. The SSIM loss is defined as follows:

$$L_S = 1 - SSIM(G(x), y) \tag{6}$$

To generate visually pleasing images, we also used the VGG-based perceptual loss [17] function. The perceptual loss function is used to determine the high-level feature differences between the target and generated output, such as content and style differences. In our proposed approach, we extract the high-level features (rectified linear unit 3 (ReLU3)-3 layer) of VGG-16 for both the real target image and the output of the generator. The L_1 distances between these features of the target and generated images are used to guide the generators G. The perceptual loss is defined as:

$$L_P = \frac{1}{C_p W_p H_p} \sum_{c=1}^{C_p} \sum_{w=1}^{W_p} \sum_{h=1}^{H_p} V(G(z|x))^{c,w,h} - V(y)^{c,w,h}, \tag{7}$$

$V(\cdot)$ denotes a particular layer of VGG-16, where the layer dimensions are C_p, W_p and H_p.

To guide the identity anonymization process, we design an identity discriminator. The identity discriminator uses the architecture of the Siamese network. We train the identity discriminator by using a contrastive loss as:

$$L_C(m, (Y, X_1, X_2)^i) = \begin{cases} ||X_1^i - X_2^i||_2 \ \mathrm{Y} = 1 \\ \max(0, m - ||X_1^i - X_2^i||_2) \ \mathrm{Y} = 0 \end{cases}, \tag{8}$$

where $||\cdot||_2$ denotes the l_2 norm of a vector and m is the margin. Finally, under the guidance of the identity discriminator, the generator learns to generate a face with some of the features of the desired identity while retaining the basic attributes of the real image.

The overall objective function for learning the network parameters in the proposed method is given as the sum of all the loss functions defined above:

$$L_{tot} = L_{adv} + \lambda_1 L_S + \lambda_2 L_P + \lambda_3 L_C, \tag{9}$$

where L_{adv} is the adversarial loss, L_P is the perceptual loss, and L_S is the SSIM loss. The variables λ_1, λ_2, and λ_3 are hyperparameters used to weight the different loss terms.

4 Experiments

4.1 Experimental settings

4.1.1 Datasets and baseline methods

The CelebA [20] dataset consists of 202,599 face images (218×178 pixels each) and 40 binary attribute annotations per image, such as age (old or young), gender, whether the image is blurry, whether the person is bald. The dataset has an official split into a training set containing 162,770 images, a validation set containing 19,867 images and a test set containing 19,962 images.

FG-NET Aging Dataset (FG-NET-AD) [33] contains 1002 images from 82 persons aged from newborn to 69 years old, but most of them are between 0 and 40 years old. Meanwhile, there are significant diversity in resolution, quality, illumination and viewpoint in the face images in the FG-NET-AD. To evaluate the model's generalization on diverse data, we also select some images from the CALFW [3] and LFW [13] dataset to set up new datasets with specific data distribution.

We evaluate our method compared to some currently advanced methods, including Fawkes [38], DeepPrivacy [14] and the CIAGAN [24]. Fawkes aims to reduce the probability of face identification without changing the visual feeling of the face. Unlike Fawkes, CIAGAN and DeepPrivacy are committed to generating new faces. DeepPrivacy includes a generator and a discriminator and uses landmarks to guide the generation process. Compared with DeepPrivacy, CIAGAN adds a discriminator for guiding identity. In addition, this method uses face silhouette to guide the generation process.

4.1.2 Implementation details

We resize all images to 64×64 for the quantitative experiments and to 128×128 pixels for the qualitative experiments, and we normalize all pixel values to the region [0,1]. We use a higher resolution for the qualitative results to make subtle visual changes more apparent.

We train our model on 35579 images from 1200 persons as done for the CIAGAN. We train our network on 128×128 resolution images. To evaluate our model's performance

accurately, we test the SDGAN and the baseline approaches on distinct datasets, including the 363 persons (each person has more than 30 images) from the same CelebA dataset and the 82 persons from FG-NET Aging Dataset. In addition, to test the model's generalization ability, we set up three mixed datasets by selecting images of distinct ages, genders, and skin tones from the CALFW, CelebA and LFW dataset, called the Gender dataset, containing 50 females and 50 males; the Age dataset, containing 17 children, 35 adults and 35 old people; the Skin dataset, containing 100 persons with white, yellow, brown and black skins separately, each 25 persons of the same skin tone type.

4.2 Detection and identification

We first evaluate two important attributes that an anonymization method should have: a high detection rate and a low identification rate. In other words, we do not want the generated face to be identified as the original ID by the identification system, but at the same time, we still want the face to be detected by the detection system. Additional, we hope the generated faces for the same person still can be identified as one person, which can enable more visual applications not to be influenced by the DeID, such as ReID and action recognition. Therefore, we exploit evaluation metrics in terms of the face detection, identification and re-identification metrics. It is known that a high detection rate, a low identification rate and a high re-identification rate indicate better anonymization.

We perform detection using the machine learning library (Dlib) [19] and SSH single-stage headless (SSH) detector [28]. For identification, We use a pretrained FaceNet model [37] based on the Inception-ResNet backbone [42] and use the standard Recall@1 evaluation metric to judge whether a generated face and its corresponding original face belong to the same ID, that is, to measure the effect of De-ID. With regard to Re-Identification, we detect Recall@1 of all generated face images to measure the ratio regarding the number of samples whose nearest neighbor is from the same class, which can implicitly evaluate the impact when applying De-ID in the Re-ID scenarios relying on the facial information.

In Table 1, we show the detection and identification results of the proposed SDGAN and baseline models. Among the existing methods, CIAGAN and DeepPrivacy achieve advanced performance; that is, they have higher detection rates and lower identification rates than the other methods. Although Fawkes can preserve the visual feeling of the original face, it is difficult to anonymize the face image. The detection rates of the classical Dlib [19] and deep learning-based SSH [28] detectors for our anonymized images are 98.12% and 99.76%, respectively, which are higher than those of the CIAGAN and DeepPrivacy. The detection rate of the SSH detector for our anonymized images is almost 100%. The testing results obtained with FaceNet show that the identification rate of our model nearly reaches 0.0%, which suggests that our model almost removes all the identity information, making it better than the CIAGAN and DeepPrivacy. The above experimental results demonstrate that our method can not only generate reliable faces but also has advanced De-ID performance.

In addition, in terms of the Re-ID score, Recall@1 scores of all the De-ID models are lower than the ones of the original face images, suggesting that the resultant faces from one person of each model can not maintain the same ID more or less. Our model achieves the better Recall@1 scores than CIAGAN and DeepPrivacy, which indicates that our model has less impact in the ReID scenarios compared to the baseline approaches. Fawkes provides best Recall@1 score due to its anonymization mechanism without changing the visual perception of face, which guarantees the identification consistency but loses the visual ID protection.

Table 1 Results of the tested detection and identification methods on the CelebA dataset. Lower (↓) results imply better anonymization. Higher (↑) results imply better detection

Models	Detection (↑)		Identification (↓)	Re-Identification (↑)
	Dlib	SSH	FaceNet	FaceNet
Original	99.61	99.85	–	95.46
Pixelization (16 by 16)	0.00	0.00	0.30	–
Pixelization (8 by 8)	0.00	0.00	0.30	–
Blur (9 by 9)	90.60	38.60	57.20	–
Blur (17 by 17)	68.40	0.30	0.50	–
Fawkes	99.67	99.80	23.70	86.05
DeepPrivacy	98.98	99.75	0.18	26.05
CIAGAN	97.19	97.96	0.14	67.12
Ours	98.12	99.76	0.05	71.72

Table 2 reports the comparison results on the FG-NET-AD dataset. Our method provides best detection rate in SSH, the identification rate, and re-identification rate and second best detection rate in Dlib (slightly lower than Fawkes), indicating the proposed SDGAN anonymizes successfully. It can also be observed that the detection rate measured by SSH of CIAGAN, DeepPrivacy and Fawkes drops significantly on the FG-NET-AD dataset compared to the CelebA dataset, while our model maintains steady, suggesting that the SDGAN is robust enough to overcome the impact of low image quality of the FG-NET-AD. Be limited by the space, the quantitative evaluation results on "Gender", "Age" and "Skin" dataset are available at https://github.com/kimhyeongbok/SDGAN, and similar results are reported on these datasets.

4.3 Generation quality

4.3.1 Quantitative results

In this section, we evaluate the visual quality of the generated images from a quantitative point of view by using the Fréchet inception distance (FID) [12], SSIM [45] and the peak signal-to-noise ratio (PSNR) [15]. The FID and SSIM are metrics that compare the statistics of generated samples to those of real samples. The lower the FID and the higher the SSIM are, the better the results are, corresponding to more similar real and generated samples.

Table 2 Results of the tested detection and identification methods on the FG-NET-AD

Models	Detection (↑)		Identification (↓)	Re-Identification (↑)
	Dlib	SSH	FaceNet	FaceNet
Original	99.89	85.47	–	75.28
Fawkes	97.90	66.77	0.011	46.36
DeepPrivacy	92.91	70.46	0.014	7.79
CIAGAN	85.15	84.74	0.001	40.84
Ours	97.85	1	0	52.81

The PSNR evaluates the image quality based on the errors between corresponding pixels. The higher the PSNR is, the higher the image quality.

As shown in Table 3, compared with Fawkes [38] and the CIAGAN [24], our method achieved significantly improved generation quality. For example, compared with the CIAGAN on the CelebA dataset, our method reduces the FID by 51.94 and improves the SSIM and PSNR by 0.12 and 6.58, respectively. DeepPrivacy provides slightly better PSNR score, but much worse FID score than our model, which indicates that the faces generated by DeepPrivacy have better peak signal-to-noise ratio but are more difficult to be detected, and have lower visual quality compared to our model. Overall, the quantitative results of FID, SSIM and PSNR show that the quality of the image generated by our method is better than that of the existing advanced methods. In Table 4, similar results can be observed, the proposed SDGAN provides much better FID and SSIM than CIAGAN and DeepPrivacy, and comparable PSNR with DeepPrivacy.

4.3.2 Qualitative results

In this section, we qualitatively evaluate the quality of the generated images. We compared the generated images under diverse views: normal, side, occlusion, and other challenging scenes. Figure 5 shows the generated images obtained under normal conditions. Compared with those of DeepPrivacy, the CIAGAN and our method, the image generated by Fawkes is highly similar to the original image, and it is difficult to anonymize the image from a visual point of view. Both DeepPrivacy and our model provide more realistic images than the CIAGAN; meanwhile, the generated faces from DeepPrivacy look natural compared to our model.

Figure 6 shows the generated images obtained from side scenes. Although Fawkes can generate high-quality images, it cannot effectively anonymize the images. Compared with those of DeepPrivacy and the CIAGAN, the images generated by our method are more realistic. For example, in the second and fourth columns, both DeepPrivacy and the CIAGAN generate unrealistic images.

Figure 7 shows the generated images obtained in occluded scenes. Fawkes still has difficulty anonymizing identities. Compared with the CIAGAN, DeepPrivacy and our method can generate images with higher quality. Furthermore, our method can generate more realistic images than DeepPrivacy. As shown in columns 3 and 9, DeepPrivacy destroys the integrity of the occlusion. In addition, as shown in column 7, our method can better maintain the expression of the original image than DeepPrivacy.

Figure 8 shows the generated images obtained in other scenes, including uneven illumination, different ages, skin colors and low image quality. These images are generated from the dataset "FG-NET-AD", "Age", and "Skin". Since the quantitative evaluation and the algorithmic principle have proven that Fawkes has difficulty anonymizing identities, its

Table 3 FID, SSIM and PSNR results on the CelebA dataset. The lower (↓) the FID and the higher (↑) the SSIM, the better the results are, corresponding to more similar real and generated samples. The higher (↑) the PSNR is, the higher the image quality

Models	FID (↓)	SSIM (↑)	PSNR (↑)
DeepPrivacy	18.50	0.87	24.55
Fawkes	28.13	1.00	Inf
CIAGAN	56.16	0.77	17.79
Ours	4.22	0.89	24.37

Table 4 FID, SSIM and PSNR results on the FG-NET-AD

Models	FID (↓)	SSIM (↑)	PSNR (↑)
DeepPrivacy	21.63	0.80	25.34
Fawkes	6.50	0.99	Inf
CIAGAN	33.22	0.57	18.85
Ours	17.53	0.90	25.16

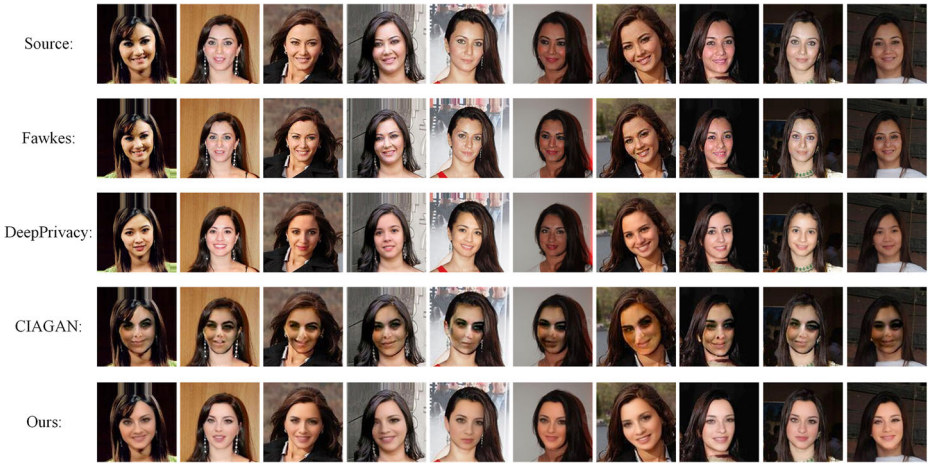

Fig. 5 Images generated by each method in normal scenes. Images in the same column correspond to the same original image

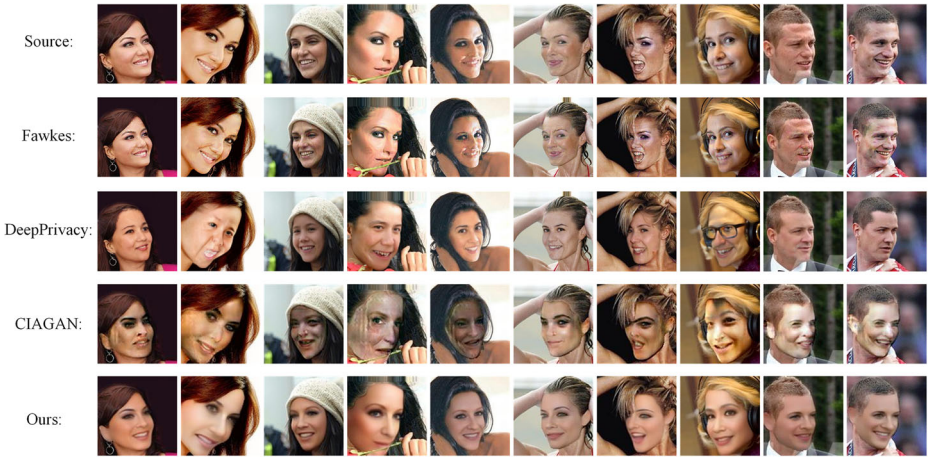

Fig. 6 Images generated by each method in side scenes. Images in the same column correspond to the same original image

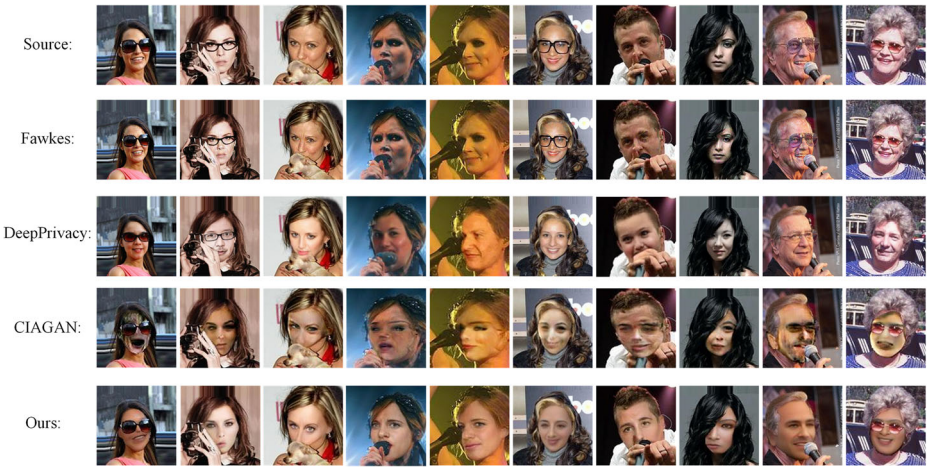

Fig. 7 Images generated by each method in occluded scenes. Images in the same column correspond to the same original image

resultant images are not provided here but are available online. Figure 9 provides the generated images with multiple faces only from our model and DeepPrivacy because CIAGAN cannot process multiface cases, and therefore, there are no results. In our model, the face recognition model Dlib is exploited to find the person faces in an image. Then image pieces are obtained by segmenting the original image and making sure each one piece only contains one face. Then these image pieces are de-identification. Finally, the generated faces replace their corresponding original faces to anonymize all the persons in the image.

In all these challenging cases, most of the generated faces from CIAGAN are flawed or unreal, while DeepPrivacy and our method can generate images with higher quality. In the testing of diverse age groups, our model and CIAGAN cannot guarantee age consistency

Fig. 8 Images generated by each method in diverse age, gender, skin color, and image quality scenes. Images in the same column correspond to the same original image

Fig. 9 Multiface images generated by each method. Images in the same column correspond to the same original image

because the age factor has not been considered in the design of their generators and discriminators. DeepPrivacy obtains better generation consistency with age and gender, especially in the children's images. However, in terms of maintaining expressions, it can be observed that compared with the original images, the facial images generated by DeepPrivacy essentially change in expression. Some of these images, such as column 10 in Fig. 6, column 8 in Fig. 7, column 2,3 in Fig. 8, and column 4 in Fig. 9, have changed from a smile to an appearance of displeasure, or vice versa, which violates the key requirements of De-ID. In contrast, the CIAGAN and our model can better maintain the expression of the original image. Overall, it is verified that the performance of our method is better than that of DeepPrivacy and CIAGAN.

4.4 Ablation studies

In this section, we perform an ablation study with our method to demonstrate the value of our design choices. In Table 5, we show several variants of our model.

Effectiveness of SPADE. Compared with that of the baseline, the quality of the image generated by V1 is significantly improved. Specifically, the FID decreases by 41.45, and the SSIM and PSNR increase by 0.05 and 3.12, respectively. Thus, the effectiveness of SPADE in our method is verified.

Table 5 Ablation study of our model

Model	SPADE	Feature	loss	Detection		Identification		Generation quality	
				Dlib	SSH	FaceNet	FID	SSIM	PSNR
Baseline	×	×	$L_{adv} + L_C$	97.19	97.96	0.14	56.16	0.77	17.79
V1	✓	×	$L_{adv} + L_C$	98.65	99.02	0.04	14.71	0.82	20.91
V2	✓	×	L_{tot}	98.55	99.71	0.22	6.17	0.91	25.14
V3	✓	F0	L_{tot}	98.31	99.78	0.09	5.67	0.90	24.70
V4	✓	F1	$L_{adv} + L_C + L_P$	98.05	99.44	0.22	10.19	0.90	24.85
V5	✓	F1	$L_{adv} + L_C + L_S$	97.43	97.62	0.05	16.20	0.85	22.48
V6	✓	F1	L_{tot}	98.12	99.76	0.05	4.22	0.89	24.37

Effectiveness of L_P and L_S. V2 and V1 have similar Dlib, and the SSH of V2 is 0.69 higher than that of V1, which verifies that L_P and L_S can improve the authenticity of the generated image. Compared with those of V1, the SSIM and PSNR of V2 are improved by 0.09 and 4.23, respectively, which verifies that L_P and L_S can improve the quality of the generated image. In addition, we verify the effectiveness of L_P and L_S. Table 3 shows that V4 removes L_S and obtains a worse FID than V6. Compared with V6, V5 removes L_P and obtains worse Dlib, SSH, FID and PSNR values. Thus, the effectiveness of L_P and L_S is verified.

Effectiveness of the identity features. We try two different identity features, F0 and F1, where F0 represents all features of the image and F1 is a simplified feature. Specifically, F1 eliminates the identity-independent features in F0. We find that V3 and V6 achieve higher performance than V2. This verifies the validity of the selected identity features. In addition, we find that V3 can obtain a lower FID than V6 because we eliminate the interference of irrelevant features so that the generator can focus more on identity-related features.

5 Conclusion

In this paper, we propose the SDGAN for high-fidelity face deidentification. Specifically, we add identity features and a semantic image to the generator. The introduction of identity features enables the generator to learn image features, not just identity features. The combination of SPADE and a semantic image can preserve the basic attributes in the original face. In addition, we introduce a perceptual loss and an SSIM loss to ensure the quality of the generated image. The results of ablation studies verify the effectiveness of the above components. In addition, extensive experimental results demonstrate the effectiveness and progressiveness of the proposed method in terms of identity anonymization. In contrast, as shown in column 7, our method can better maintain the expression of the original image than DeepPrivacy.

Declarations

Conflict of Interests We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

1. Boyle M, Edwards C, Greenberg S (2000) The effects of filtered video on awareness and privacy. In: Proceedings of the 2000 ACM conference on computer supported cooperative work, pp 1–10
2. Breve B, Caruccio L, Cirillo S, Desiato D, Deufemia V, Polese G (2020) Enhancing user awareness during internet browsing. In: ITASEC, pp 71–81

3. Chen BC, Chen CS, Hsu WH (2014) Cross-age reference coding for age-invariant face recognition and retrieval. In: Proceedings of the European conference on computer vision, pp 768–783
4. Cheng K, Zhang Y, He X, Chen W, Cheng J, Lu H (2020) Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 183–192
5. Desiato D (2018) A methodology for GDPR compliant data processing. In: SEBD
6. Gafni O, Wolf L, Taigman Y (2019) Live face de-identification in video. In: Proceedings of the IEEE international conference on computer vision, pp 9378–9387
7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014) Generative adversarial nets. In: Proceedings of Advances in neural information processing systems, pp 2672–2680
8. Gross R, Sweeney L, De la Torre F, Baker S (2006) Model-based face de-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshop, pp 161–161
9. Gu X, Luo W, Ryoo MS, Lee YJ (2020) Password-conditioned anonymization and deanonymization with face identity transformers. In: European conference on computer vision, pp 727–743
10. Guo K, Hu X, Li X (2021) MMFGAN: A novel multimodal brain medical image fusion based on the improvement of generative adversarial network. Multimedia Tools and Applications, pp 1–39
11. Guo J, Pang Z, Bai M, Xie P, Chen Y (2021) Dual generative adversarial active learning. Appl Intell, pp 1–12
12. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, pp 6626–6637
13. Huan GB, Mattar M, Berg T, Learned-Miller E (2008) Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition
14. Hukkelås H, Mester R, Lindseth F (2019) Deepprivacy: A generative adversarial network for face anonymization. In: International symposium on visual computing, pp 565–578
15. Huynh-Thu Q, Ghanbari M (2008) Scope of validity of PSNR in image/video quality assessment. Electronics Lett 44(13):800–801
16. Isola P, Zhu J, Zhou T, Efros A (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
17. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of the European conference on computer vision, pp 694–711
18. Jourabloo A, Yin X, Liu X (2015) Attribute preserved face de-identification. In: 2015 international conference on biometrics, pp 278–285
19. King D (2009) Dlib-ml: A machine learning toolkit. J Mach Learn Res 10:1755–1758
20. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3730–3738
21. Liu Z, Qi X, Torr P (2020) Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8060–8069
22. Ma T, Li D, Wang W, Dong J (2021) CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation. arXiv:2105.11137
23. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802
24. Maximov M, Elezi I, Leal-Taixé L (2020) Ciagan: Conditional identity anonymization generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5447–5456
25. Meden B, Emersic Z, Struc V, Peer P (2017) K-same-net: Neural-network-based face deidentification. In: 2017 international conference and workshop on bioinspired intelligence, pp 1–7
26. Meden B, Mallı RC, Fabijan S, Ekenel HK, Štruc V, Peer P (2017) Face deidentification with generative deep neural networks. IET Signal Process 11(9):1046–1054
27. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv:1411.1784
28. Najibi M, Samangouei P, Chellappa R, Davis LS (2017) Ssh: Single stage headless face detector. In: Proceedings of the IEEE international conference on computer vision, pp 4875–4884
29. Neustaedter C, Greenberg S, Boyle M (2006) Blur filtration fails to preserve privacy for home-based video conferencing. ACM Transactions on Computer-Human Interaction 13(1):1–36
30. Newton EM, Sweeney L, Malin B (2005) Preserving privacy by de-identifying face images. IEEE Trans Knowl Data Eng 17(2):232–243
31. Pang Z, Guo J, Ma Z, Sun W, Xiao Y (2021) Median stable clustering and global distance classification for cross-domain person re-identification. IEEE Trans Circuits Syst Video Technol, pp 1–15
32. Pang Z, Guo J, Sun W, Xiao Y, Yu M (2021) Cross-domain person re-identification by hybrid supervised and unsupervised learning. Appl Intell, pp 1–15

33. Panis G, Lanitis A (2014) An overview of research activities in facial age estimation using the FG-NET aging database. In: Proceedings of the European conference on computer vision, pp 737–750
34. Park T, Liu M, Wang T, Zhu JY (2019) Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2337–2346
35. Ren Z, Lee Y, Ryoo M (2018) Learning to anonymize faces for privacy preserving action detection. In: Proceedings of the European conference on computer vision, pp 620–636
36. Ryoo MS, Kim K, Yang HJ (2018) Extreme low resolution activity recognition with multi-siamese embedding learning. In: Proceedings of AAAI conference on artificial intelligence
37. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition
38. Shan S, Wenger E, Zhang J, Li H, Zheng H, Zhao BY (2020) Fawkes: Protecting privacy against unauthorized deep learning models. In: 29th Security Symposium, pp 1589–1604
39. Sun Q, Ma L, Oh SJ, Van Gool L, Schiele B, Fritz M (2018) Natural and effective obfuscation by head inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5050–5059
40. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision, pp 843–852
41. Sweeney L (2002) k-anonymity: A model for protecting privacy. Internat J Uncertain Fuzziness Knowledge-Based Systems 10(05):557–570
42. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of AAAI conference on artificial intelligence, pp 4278–4284
43. Targ S, Almeida D, Lyman K (2016) Resnet in resnet: Generalizing residual architectures. arXiv:1603.08029
44. Te G, Liu Y, Hu W, Shi H, Mei T (2020) Edge-aware graph representation learning and reasoning for face parsing. In: European conference on computer vision, pp 258–274
45. Wang Z, Simoncelli E, Bovik A (2003) Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, vol 2, pp 1398–1402
46. Wu Y, Ji Q (2019) Facial landmark detection: A literature survey. Int J Comput Vis 127(2):115–142
47. Wu Y, Yang F, Xu Y, Ling H (2019) Privacy-protective-GAN for privacy prerving face de-identification. J Comput Sci Technol 34(1):47–60
48. Yang S, Luo P, Loy C, Tang X (2016) Wider face: a face detection benchmark. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5525–5533
49. Yang C, Xu Y, Shi J, Dai B, Zhou B (2020) Temporal pyramid network for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 591–600