Privacy-Protective-GAN for Face De-identification

Yifan Wu, Fan Yang, Haibin Ling
Temple University
{yifan.wu, fyang, hbling} @temple.edu

Abstract

Face de-identification has become increasingly important as the image sources are explosively growing and easily accessible. The advance of new face recognition techniques also arises peoples concern regarding the privacy leakage. The mainstream pipelines of face de-identification are mostly based on the k-same framework, which bears critiques of low effectiveness and poor visual quality. In this paper, we propose a new framework called Privacy-Protective-GAN (PP-GAN) that adapts GAN with novel verificator and regulator modules specially designed for the face de-identification problem to ensure generating de-identified output with retained structure similarity according to a single input. We evaluate the proposed approach in terms of privacy protection, utility preservation, and structure similarity. Our approach not only outperforms existing face de-identification techniques but also provides a practical framework of adapting GAN with priors of domain knowledge.

1. Introduction

Beneficial from the blooming development of media and network techniques that makes huge amount of images more approachable, image analysis techniques bear its prosperity in the past decade and brings unprecedented convenience to our daily life. Among those image sources exposed to the public with or without our awareness, a considerable number of them contain our identity especially the biometric information. Not only will the unprotected exposing cause the leak of privacy, the common approaches of protection like blurring and pixelization may also not be satisfied in thwarting face recognition software [22, 24]. The other extreme side is that we simply mask identity area off, which is perfect for identity removal. But it causes serious loss of data utility in application of visual understanding as the scene information is changed with objects removal. Thus it is critically important to build a framework that can properly de-identify the privacy information from the image while keeping its utility at the same time.

Figure 1. **Illustration of face de-identification.** The top row shows the original face images, the bottom row represents corresponding de-identified images synthesized by our method.

Specially for the face de-identification problem, the dilemma is that on the one hand, we want the de-identified image to look as different as possible from the original image to ensure the removal of identity; on the other hand, we expect the de-identified image to retain as much structural information in the original image as possible so that the image utility remains. Previous approaches on this problem are mostly based on k-same algorithm [9, 11, 12, 22] for the de-identificatin procedure and some apply additional models like the Active Appearance Models (AAMs) [19] to explicitly construct faces preserving the utility attributes like gender, race, and age [6, 16]. Yet these models fail to make full use of existing data and deliver fairly poor visual quality due to the unnatural synthesis.

The Generative Adversarial Networks (GANs) provide an inspiring framework on generating sharp and realistic natural image samples via adversarial training [8, 15, 35]. GAN can be naturally used for the face de-identification as it can generate new samples from the gallery following original input data distribution. Since GAN implicitly models a dataset distribution, it cannot involve processing individual images. To achieve an image-to-image face de-identification, conditinal GAN (cGAN) is intuitively leveraged here, which is an extention of GAN that fits a conditional data distribution [21].

Unfortunately, cGAN is not directly applicable to our scenario, which requires privacy protection as well as data utility preservation. First, the generative loss in cGAN is

not specific for distinguishing identities, thus the generated samples could be either too similar with each other to pass external face verification, or visually ghosted as none-faces. Second, the structure information is unensured to be preserved due to the inconsistency in feature spaces between GAN-type and structural losses.

Thus to make an adaptive GAN framework for the de-identification problem, we introduce additional mechanisms to enlarge the sampling range away from the given samples in the identity-related embedding feature space properly and to constrain the variation of the part of image structure from the full image feature space simultaneously. When unified in the GAN loss, these two tasks are actually contradictory to each other because the GAN loss is unable to explicitly tell the embedding feature from the structure feature and they contribute to the performance of distinguishing two faces in the same direction. Thus here we explicitly express them as external losses to compensate their contradiction in GAN and consequently achieve a novel Privacy-Protective-GAN model (PP-GAN, see Figure. 2), which can plausibly balance between id-removal and structure retains.

Indeed, in our PP-GAN model, we introduce two additional types of external modules: the verificator with contrastive loss [13] to allow cGAN's generator for a larger range of sampling exploration, and the regulator with Structural Similarity Index (SSIM) [31] to account for the structure preservation. Intuitively, the verificator adds prior information of identity feature matching in the embedding space to help remove biometic information while the regulator tells the generator how to maintain similar image utility via luminance, contrast and structural similarities. The involvement of these two prior knowledge on what is identity information versus the structure information significantly improve the model performance.

In the experimental setting, we quantitatively demonstrate the effectiveness of proposed method in several aspects: (i) The generative de-identified image can thwart the face verificator; (ii) The de-identified image is not simply switched with others in the training data; (iii) The luminance, contrast and structure can be partially preserved and (iv) The attributes utility can be preserved when training an attribute specific generator.

In summary, our contributions are as following:

• To the best of our knowledge, we are the first to propose a GAN-based framework that is trainable in an end-to-end manner directly for the face de-identification task. The integrated framework can synthesize de-identified output for each single probe face.

• We manage to balance the trade-off between image quality and privacy protection by introducing novel modules, the verificator and the regulator. The model learns to retain the structure similarity with the original image on the pix level, while distinguishes the input and synthesized output in the identity-related feature space. It is the first time that a de-identification metric is properly aggregated into the objective function in a controllable and measurable manner, which largely ensured the effectiveness of de-identification and privacy protection.

• We are the first to propose a systematic plan of evaluating the results of de-identification, where we employ a state-of-the-art face verificator [27] to determine the de-identification rate and the a face detector [34] to quantify the detection rate. These two measurements combinely indicate how well the de-identification procedure works in its target of with utility retained. The proposed method achieves 100% de-identification rate and obtains best trade-off between identity removal and visual similarity.

The rest of the paper is orgnaized as following: Related works about de-identification task are summarized in Section 2; In the Section 3, we introduce our PP-GAN framework; A quantitative evaluation on a public dataset as well as visual illustrations are provided in Section 4; The paper concludes with a discussion in Section 5.

2. Related Work

Privacy protection in visual data has been drawing increasing amount of research attention recently, with focuses on different scenarios such as video analysis and security and sampled studies in [28, 3, 32, 2, 5]

Face de-identification is an important tool for visual privacy protection. The goal of face de-identification has two folds: privacy protection and data utility preservation. In such way, de-identified images conceal the identity privacy of the original image, while preserving non-identity-related aspects for further data analysis. Earlier works on face de-identification simply use masking, blurring or pixelation [24]. While these Ad Hoc methods are easily applicable, it is shown that they provide no privacy assurance since they may fail to thwart face recognition software [10]. Moreover, how to conduct a *sufficient blur* itself is non-trivial [7].

To address this problem, Newton *et al.* [22] propose the k-same algorithm based on the k-anonymity concept [30]. By applying the k-same algorithm, a given image is represented by average face of k-closet faces from the gallery. This procedure theoretically limits the performance of recognition to $1/k$, but usually suffers from ghosting artifacts in de-identified images. Some variants of k-same are proposed to improve the data utility and the naturalness of de-identified face images. The k-Same-Select algorithm [9] partitions the image set into mutually exclusive subsets and applies the k-same algorithm independently to each subsets, in this way attempts to preserve attributes of each subset. In order to overcome undesirable artifacts due to misalignments and produce de-identified images of better quality, the k-same-M algorithm [11] based on Active Appearance Models (AAMs) [19] is proposed. This algo-

rithm fits an AAM to input images then applies *k*-same to the AAM model parameters. To further keep the data practically useful, Du *et al*. explicitly preserve race, gender and age attributes in face de-identification [6]. Jourabloo *et al*. propose to jointly model face de-identification and attribute preservation in a unified optimization framework [16].

The majority of sate-of-the-art approaches are designed based on the *k*-same and implemented using AAM models. However, these methods have notable limitations: (i) The *k*-same assumes that each subject is only represented once in the dataset, but this may be violated in practice. The presence of multiple images from same subject or images share similar biometric characteristics can lead to lower levels of privacy protection; (ii) *k*-same operates on a closed set and produces a corresponding de-identified set, which is not applicable in situations that involve processing individual images or sequences of images; (iii) Generated images by AAM models do not yet look natural enough.

Recently, GANs present promising effectiveness on image generating problems. It is a natural tool to synthesize de-identified images. Karla*et al*. [1] build a GAN-based model to generate full body images for de-identification, but the quality in face areas is not guaranteed. Blaž *et al*. use GAN to synthesize de-identified faces, but still based on the *k*-same algorithm [20].

In this paper, we only focus on biometric information, i.e., we do not remove soft biometric features like age or race and non-biometric information like hairstyles [24]. We propose a novel *Privacy-Protective-GAN* framework to address the problems mentioned above.

3. Method

Face de-identification can be formulated as a transformation function δ which maps a given face image x to a de-identified image $\hat{x}$, i.e. $\delta(x) = \hat{x}$, aiming to misleading the face verification. In this work, we propose a PP-GAN model and adopt the final generator as the de-identification function. First, we pretrain a verificator to determine whether two faces are from the same subject. Next, we freeze this verificator and utilize it to compute the contrastive loss in the whole system training phase. In each iteration, the synthesized output and the original image will be forwarded in: 1) the identity-related feature space by the pre-trained verificator to enforce the removal of identity; 2) the pixel level structure similarity by SSIM to preserve visual correspondence. Then the contrastive loss and the SSIM loss will be backpropagated to update the generator. Fig. 2 gives the big picture of the whole framework.

In the following subsections, we will describe the structure of the PP-GAN and explain its learning procedure in detail. The overview of the whole system is illustrated in Section 3.1. The detailed explanations of the Generator *v.s.* Discriminator are given in Section 3.2. The verificator and structure similarity are introduced in Sections 3.3 and 3.4 respectively.

3.1. Objective

Face de-identification is a complex task involving multiple constraints at the same time. Specially, given an original face, we generate a new face image accordingly and hope that it meets three types of qualification: 1) The generated face should look natural and realistic, i.e. it looks like a face; 2) The generated face should not be discriminated as the same person with the original face, i.e. it should be far enough from the original one in the identity-related feature space; 3) The de-identified image should be consistent with the original image on the pixel level as much as possible, i.e., it shares similar luminance, contrast, and structure with the original image.

For the first objective, it can be achieved by utilizing the cGAN, which adversarially trains a generator (G) taking the original image as input and producing de-identified images versus the co-trained discriminator (D). The cGAN loss is denoted as $\mathcal{L}_{\text{cGAN}}(G, D)$ here. The details of cGAN settings will be clarified in Section 3.2.

For the second goal, we pre-train a verificator to produce a verification loss in the identity embedding space for each pair of original and de-identified images. The verification loss for each generator G is then formularized as the expected loss across subjects and denoted as $\mathcal{L}_{\text{verif}}(G)$. This is used to guide the generator to enforce that the output holds an enough distance with the input in the identity-related feature space. The details about verificator will be elaborated in Section 3.3.

For the third objective, we adopt the Structural Similarity Index (SSIM) [31] to quantify the similarity between the original and generated images. The SSIM loss combines the luminance, contrast, and structural differences in a product form with different power coefficients. Similarly as the verification loss, the SSIM loss for the generator G is then formularized as the expected loss across subjects and denoted as $\mathcal{L}_{\text{sim}}(G)$. The details of structure loss will be discussed in Section 3.4.

Combining these three types of losses, our final objective of face de-identification is defined as

$$
\begin{aligned}
&\mathcal{L}_{\text{face}}(G, D) \\
&= \mathcal{L}_{\text{cGAN}}(G, D) + \lambda_1 \mathcal{L}_{\text{verif}}(G) + \lambda_2 \mathcal{L}_{\text{sim}}(G),
\end{aligned}
\tag{1}
$$

and the optimal generator G^* is solved through the min-max procedure

$$
G^* = \arg \min_G \max_D \mathcal{L}_{\text{face}}(G, D),
\tag{2}
$$

where λ_1 and λ_2 are the hyperparameters for multiple losses.
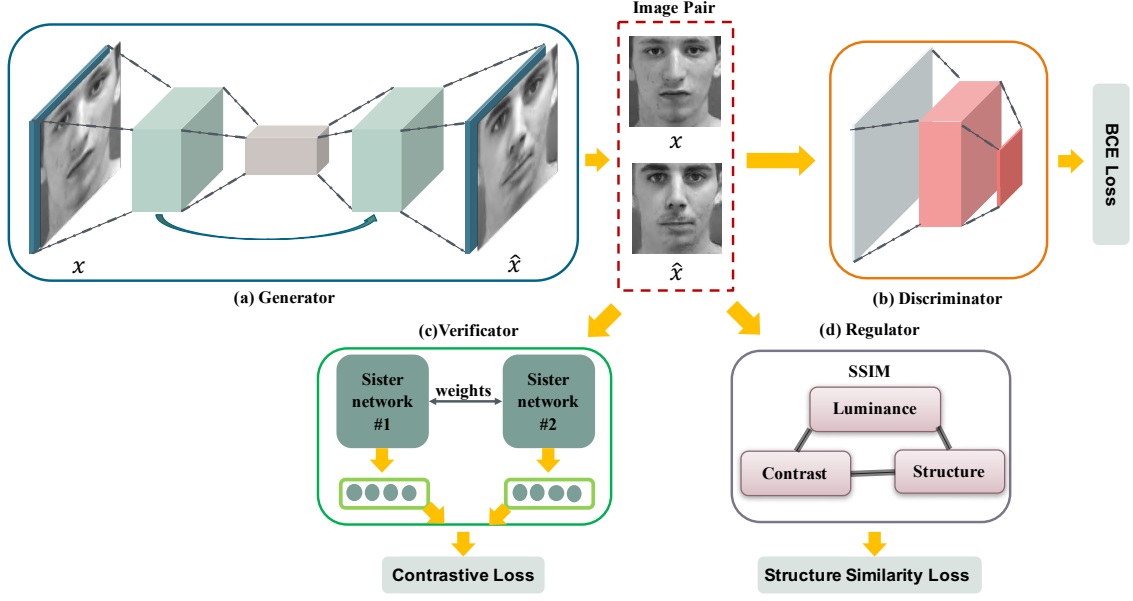
Figure 2. **The Structure of the Privacy-Protective-GAN (PP-GAN) Framework.** The PP-GAN framework consists of four components: generator, discriminator, verificator, and regulator. *(a)* The generator is a "U-Net" based auto-encoder, synthesizing a de-identified image $\hat{x}$ for given original image $\hat{x}$. *(b)* The discriminator is adversarially trained with the generator to encourage the output to be sharp and realistic. *(c)* The verifiactor is to determine whether two faces are from the same person. *(d)* The regulator is to determine how similar the two faces are in pixel level.

3.2. Architecture of GAN

The cGAN learns a mapping $G : \{x, z\} \to \hat{x}$ from an observed image x with additional random noise z to a synthesized image $\hat{x}$, where x is referred as a 'real' sample or the condition from the original dataset, $\hat{x}$ is referred as a 'fake' sample generated by the trained generator G, and z is the random noise to ensure the image variability. The adversarial procedure trains a generator to produce outputs that can hardly be distinguished as a 'fake' by the co-trained discriminator (D). Mathematically, the objective function of the cGAN is expressed as

$$\mathcal{L}_{\mathrm{cGAN}}(G, D) = \mathbb{E}_{x,\hat{x} \sim P_{\mathrm{data}}(x,\hat{x})}[\log D(x, \hat{x})] + \\ \mathbb{E}_{x \sim P_{\mathrm{data}}(x), z \sim P_z(z)}[\log(1 - D(x, G(x, z)))], \quad (3)$$

where G is the generator and D is the discriminator. The optimal G^* minimizes this objective against an adversarial D that maximizes it, and it can be solved via a min-max procedure $G^* = \arg\min_G \max_D \mathcal{L}_{\mathrm{cGAN}}(G, D)$.

In practice, we adapt our generator and discriminator architectures from the Image-to-Image translation [15]. As shown in Fig. 2, both of them use modules consisting of Convolution-BatchNorm-ReLu [14]. The generator is a "U-Net" [26] by adding skip connections between symmetric layers in the convolution and deconvolution steps. The stride keeps to be 2 so it resizes by 2 in the encoding part, expands by 2 in the decoding part. The size of the em-

bedding layer is 1×1. Two drop-out layers in the middle serve as random noise z. For the discriminator, instead of feeding it with 'real' or 'fake' pairs, we follow the standard approach from [8] and input either 'real' image x or 'fake' image $\hat{x}$ since there is no unique solution for the de-identified $G(x)$. Also, a patch-design [15] is adopted in the discriminator thus it outputs a $N \times N$ patch, rather than a single value to represent the probability of current input to be "real". Here we set N to 30 with receptive field of size 34. The network architectures are shown in Fig. 3.

3.3. Face Verification

Face verification is to determine whether two images express the same person. Metric learning is a widely used approach for this task, which learns semantic distance measurements and the associated embeddings [4, 29, 27, 23] from the original image space to feature space. Motivated by FaceNet [27], we strive for an embedding $f(x)$ so that the squared distance between all faces are independent of imaging conditions. In the embedding spaces, samples with the same identity are closer whereas those with different identities are further apart.

Mathematically, the embedding is denoted as $f : \mathrm{data} \to \mathbb{R}^d$, which embeds an image x into a d-dimensional Euclidean space. We adopt the following contrastive loss function originally proposed by Hadsell *et al.* [13]. The loss
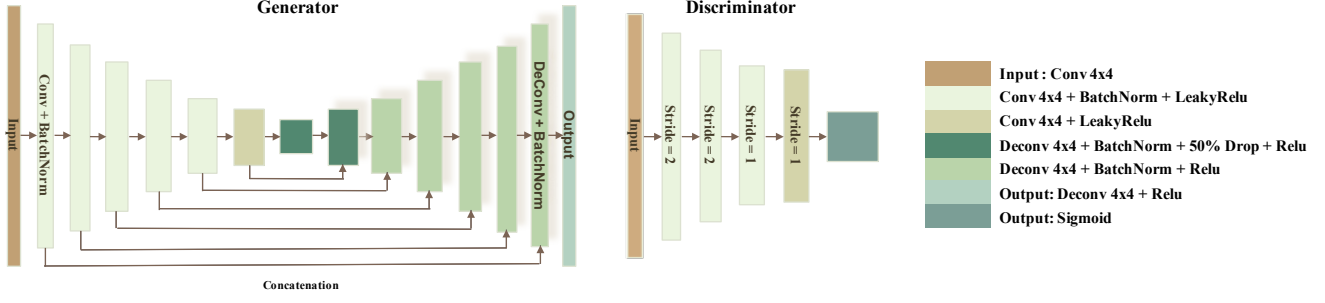
Figure 3. **The network architecture of the generator G and the discriminator D.** The generator has a "U-net" architecture with input size of $1 \times 128 \times 128$, embedding size of $256 \times 1 \times 1$, and output size that is consistent with the input size. The input size of discriminator is $1 \times 128 \times 128$, while the output size is $1 \times 30 \times 30$.

function is defined as

$$
\begin{aligned}
&\text{Verif}(x_i, x_j, \eta_{x_i, x_j}, \alpha) = \\
&\begin{cases}
\dfrac{1}{2} \|f(x_i) - f(x_j)\|_2^2, & \text{if } \eta_{x_i, x_j} = 0 \\
\dfrac{1}{2} \max\left(0, \ \alpha - \|f(x_i) - f(x_j)\|_2\right)^2, & \text{if } \eta_{x_i, x_j} = 1
\end{cases}
\end{aligned}
\tag{4}
$$

where α is the margin that is enforced between positive and negative pairs, f is the embedding function, and η_{x_i, x_j} is the indicator where $\eta_{x_i, x_j} = 0$ means positive pairs, and $\eta_{x_i, x_j} = 1$ means negative pairs.

For the network architecture, we choose the new Light CNN-9 model [33], which introduces the MaxFeature-Map (MFM) operation as an alternative of ReLU and claims that the MFM can play a role of feature selection. With this design, the Light CNN-9 model achieves state-of-the-art results with less parameters and time-consumption. The Light CNN-9 model contains 5 convolution layers, 4 Network in Network (NIN) layers [18], Max-Feature-Map function and 4 max-pooling layers. The final layer is a fully connected one that outputs a 256-dimensional representation. A detailed explanation of this architecture can be found in [33].

In the pre-training phase of the verificator, we randomly sample 50% positive and 50% negative pairs as train data to avoid data unbalance. These pairs are separately fed to a shared weight Siamese network, where the model is updated by the contrastive loss defined in Eq. (4). Once the verificator training is finished, parameters are frozen in the whole system training followed.

Next in the PP-GAN training phase, the pair of original image x and synthesized image $\hat{x}$ are fed into pre-trained verificator with $\eta_{x, \hat{x}} = 1$ to enforce their different identities, i.e., the de-identification.

The term of verification loss in Eq. (1) is then defined as

$$
\begin{aligned}
&\mathcal{L}_{\text{verif}}(G) \\
&= \mathbb{E}_{x \sim P_{\text{data}(x)}, z \sim P_z} \left[\text{Verif}(x, G(x, z), \eta, \alpha) \right],
\end{aligned}
\tag{5}
$$

where x is the original input I, z is the noise, $G(x, z)$ is the generated image, $\eta = 1$ is the always false indicator, and the margin is set to 2 on the normalized feature vector to enforce a positive loss.

3.4. Structure Similarity

For the de-identification problem, we want to hold the correspondence between original image x and synthesized image $\hat{x}$. The vanilla GAN simply generates new samples from the data distribution. For cGAN, although it samples under the condition of the input image x, it does not ensure that the synthesized image $\hat{x}$ is close enough to x in the image space. For example, if we have two images x_i and x_j with the same condition, the cGAN would generate $\hat{x}_i$ and $\hat{x}_j$ with the same distribution so that we cannot tell whether $\hat{x}_i$ is the de-identified image from x_i or x_j. Thus an additional structural loss is necessary here to regulate the generator to assure this matching. The two popular choices, the mean squared error (MSE) and the related quantity of peak signal-to-noise ratio (PSNR), are not well suited because they do not match very well to the perceived visual quality by humans, and the pixel-level matching with the reference image is not what we want to optimize. Instead we use the *Structural Similarity Index* (SSIM) [31] as an objective measurement for assessing perceptual image degradation after de-identification.

SSIM consists of three components including luminance similarity $l(x, \hat{x})$, contrast similarity $c(x, \hat{x})$ and structural similarity $s(x, \hat{x})$:

$$
\text{SSIM}(x, \hat{x}) = l(x, \hat{x})^\alpha \cdot c(x, \hat{x})^\beta \cdot s(x, \hat{x})^\gamma
\tag{6}
$$

For a detailed explanation of SSIM, one can refer to [31]. Here we define $\mathcal{L}_{\text{sim}}(x, \hat{x}) = \frac{1}{2}(1 - \text{SSIM}(x, \hat{x}))$. Then in the GAN training, we have

$$
\mathcal{L}_{\text{sim}}(G) = \mathbb{E}_{x, \sim P_{\text{data}(x)}, z \sim P_z} \left[\frac{1}{2} \left(1 - \text{SSIM}(x, G(x, z)) \right) \right],
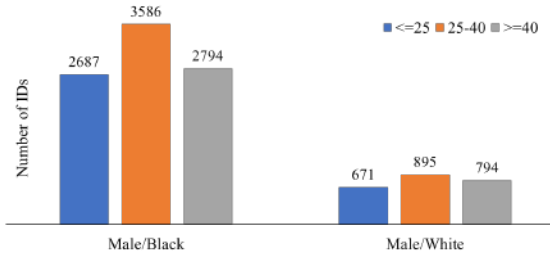$$

Figure 4. **The demographic distribution of the MORPH dataset.** The black group contains more samples than the white group, while both have the similar age distribution.

Table 1. De-identification rate (percent) of Male/Black.

	Black	Black-Youth	Black-Middle	Black-Senior
Original Test	1.5	1.6	1.6	2.7
De-identified Train	100.0	100.0	100.0	97.2
De-identified Test	100.0	100.0	97.0	93.7

Table 2. De-identification rate (percent) of Male/White.

	White	White-Youth	White-Middle	White-Senior
Original Test	1.8	2.8	3.3	5.9
De-identified Train	100.0	96.2	89.7	91.7
De-identified Test	100.0	94.4	90.8	84.7

where z is the noise, and $G(x, z)$ is the generated image.

4. Experiments

4.1. experimental Setup

We evaluate the proposed framework on a publicly available MORPH [25] dataset, which contains 55,000 unique images of more than 13,000 individuals, with diverse demographic information like age, gender, and race. Here we only use male data since the number of femalel subjects is limited. The details of the demographic distribution in the MORPH dataset are shown in Fig. 4.

To validate that our framework can preserve attributes utility, data is divided into black and white according to race, then we divide each category into three age groups: youth (age $\leqslant 25$), middle-aged ($25 <$ age < 40) and senior (age $\geqslant 40$). We refer these 8 different data groups as *Black, White, Black-Youth, Black-Middle, Black-Senior, White-Youth, White-Middle* and *White-Senior*. We randomly pick 90% from each group for training and use the remaining for testing. The training and testing sets are consistent in the verificator pretraining and the whole system training. For the data preprocessing, we use OpenFace[1] to detect, align, and crop face areas. During Verificator pretraining, we set initial learning rate as 1e-4 and margin as 2. During the whole system training, the learning rate is set to be 1e-5. For all the experiments, we use minibatch SGD and apply the Adam solver [17].

For each group, we separately conduct four experiments to analyze our objective function: only use cGAN, cGAN with SSIM loss, cGAN with pretrained verificator, amd cGAN with both pretrained verificator and SSIM constraint (i.e. ours). Here we refer these different set-ups as: *cGAN-only, cGAN+Sim, cGAN+Verif*, and *cGAN+Sim+Verif*.

4.2. Privacy Protection

To demonstrate the proposed framework successfully generates face images for privacy protection, we need to

prove that the produced face images neither remain the same as the corresponding raw images, which means successful de-identification, nor simply changing the identity to other identities in dataset, which means no switching identity occurs.

De-identification To evaluate if the proposed model successfully removes the identity of original images, we assess if a pair of images are from same identity in a verification way. Specifically, given a verificator, for each pair of images $(x, \hat{x})$, it gives a distance value indicating the similarity of the two images. If the value is smaller than a threshold, the two images are treated as the same identity, which means de-identification fails; otherwise, from different identities, which means de-identification successes. Here we use de-identification rate (ERR rate) to represent the effectiveness of privacy protection, i.e., $(x, \hat{x})$ is not determined as the same person.

To eliminate the overfitting caused by the pre-trained verificator since it guides the training of the whole system, we fine-tune the FaceNet [27][2] with triplet loss as our evaluation verificator. The FaceNet is trained on the 8 subgroups respectively. Specifically, the test data are grouped into numerous pairs with half positive and half negative. Then the trained model is evaluated on these pairs of images in a 10-fold way. The finally gained optimal threshold is utilized in de-identification evaluation.

Table 1 and Table 2 show our de-identification performance. The FaceNet achieves accuracy greater than 97% on all Black groups and 94% on White groups. The overall de-identification rates are both very high, and that of the black group is a bit higher due to the larger sample size.

The majority state-of-the-art de-identification algorithms based on k-anonymity theory, and k-value will directly affect the de-identification rate. The APFD algorithm [16] shows the performance of de-identification saturates when $k = 8$, and achieves the de-identification rate around 90%, note that the Bayesian classifier used for evaluation only achieves verification accuracy of 70%. Recently GAN-based face de-identification achieves best de-identification rate of 43.2% when $k = 2$ [20]. Thus it is obviously that our system outperforms previous methods regrading the ef-

---

[1] https://cmusatyalab.github.io/openface

[2] https://github.com/davidsandberg/facenet

Table 3. Detection rate of Male/Black.

Data Settings	Black	Black-Youth	Black-Middle	Black-Senior
Original w/o padding	0.733	0.817	0.767	0.598
De-identified w/o padding	0.719	0.847	0.715	0.570
Original w padding	0.988	0.993	0.994	0.977
De-identified w padding	0.956	0.995	0.990	0.969

Table 4. Detection rate of Male/White.

Data Settings	White	White-Youth	White-Middle	White-Senior
Original w/o padding	0.758	0.808	0.738	0.621
De-identified w/o padding	0.784	0.767	0.739	0.499
Original w padding	0.973	0.967	0.979	0.975
De-identified w padding	0.975	0.940	0.988	0.902

fectiveness of privacy protection by explicitly adding de-identification metric into the optimization objective.

No switching-identification In addition, in order to demonstrate the yielded image does not possess the identity of the other original images, for each generated image, we compute its similarity, i.e., distance, with every original image, and compare the distance of image pair $(x_i, \hat{x}_j)$ with a threshold, where i, j are the identity of image. If the distance is smaller than the threshold and $i \neq j$, identification switching occurs. If the distance is smaller than the threshold and $i = j$, identification switching does not occur. We use the number of identification switches (IDS) times as a measurement.

We conduct experiments on *Black-Youth* and *White-Youth*. As the data are unbalanced, to fairly compare the IDS on both groups, we randomly sample same number of people from *Black-Youth* test as that in *White-Youth* test. On both groups, we obtain 0 IDS, which means no identification switching occurs.

4.3. Utility Preservation

Face-like The most important point for data utility is that the generated images should look natural and realistic, i.e, they look like face. Here we utilize a face detector MTCNN [34] to determine whether the generated images are face or not. To alleviate the variance of the face detector, we conduct experiments on both original and generated face images. Table 3 and 4 show the detection rate on both the original and generated data. From the tables, we note that the detection rates on original and de-identified data are similar, proving that our system preserves the face-like utility very well. As cropped face images are not suitable for face detector trained on face image with context, to alleviate the effects from the inconsistency, we pad the cropped image with 50 pixels of 0 along each axis. For some groups, the detection rate on de-identified images is higher than that on original images, e.g. *Black-Youth*. This is because some original face images are distorted, but yielded images have less distorted images.

Attributes Preservation To examine the performance of the proposed model for attribute preservation, we train sep-
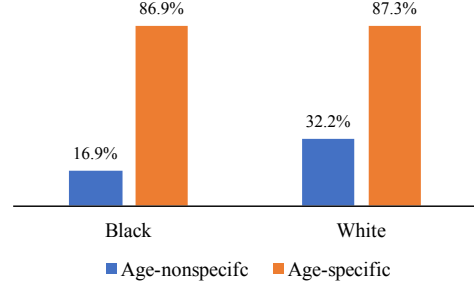

Figure 5. **Age classification accuracy.** The blue bar displays the classification accuracy of preserving attributes for age-nonspecific group and the orange bar is for age-specific group.
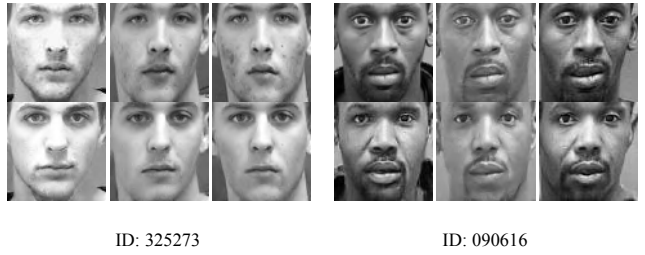

ID: 325273   ID: 090616

Figure 6. **The visualization of correspondence between original and de-identified images.** The top row shows the original face images, the bottom row shows the corresponding de-identified images synthesized by our PP-GAN model.

arate classifiers for each attribute and compute the classification accuracy for the generated images. The classification accuracy rate is used as a measurement to evaluate experimental results. From Figure 5, we can see that the age-specific models well preserve its group attributes compared to the age-nonspecific group. The experiments on original images display similar comparisons.

4.4. Visual Similarity

For the de-identification problem, in addition to the utility preservation, we want to hold the correspondence between original image x and synthesized image $\hat{x}$ rather than just replace the face area randomly. In other words, we only want to remove the privacy related characteristics, but keep the visual similarity, e.g. contours and luminous condition, as much as possible. As shown in Fig. 6, the original and generated images share consistent SSIM-consisting features like luminance, contrast, and structure information while distinguish in privacy related characteristics such as the sizes and shapes of eyes, nose, and mouth. This indicates the effectiveness of the regulator.

To demonstrate the proposed framework achieves both de-identification and visual similarity, we compute id-removal rate and SSIM of each yielded image with the cor-
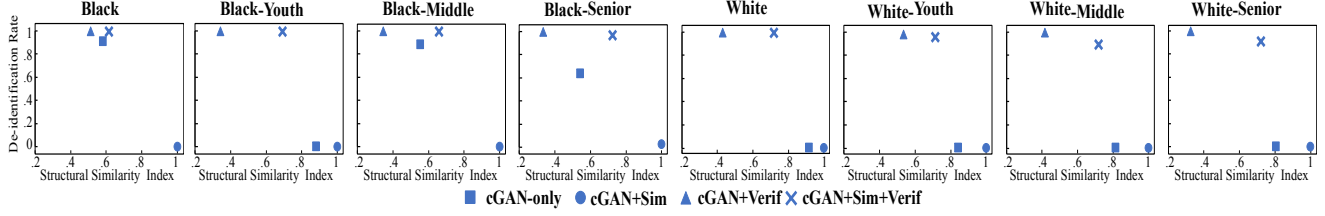
Figure 7. **The trade-off between identity removal and visual similarity.** The horizontal axis represents values of Structural Similarity Index (SSIM), the vertical axis represents the de-identification rate.
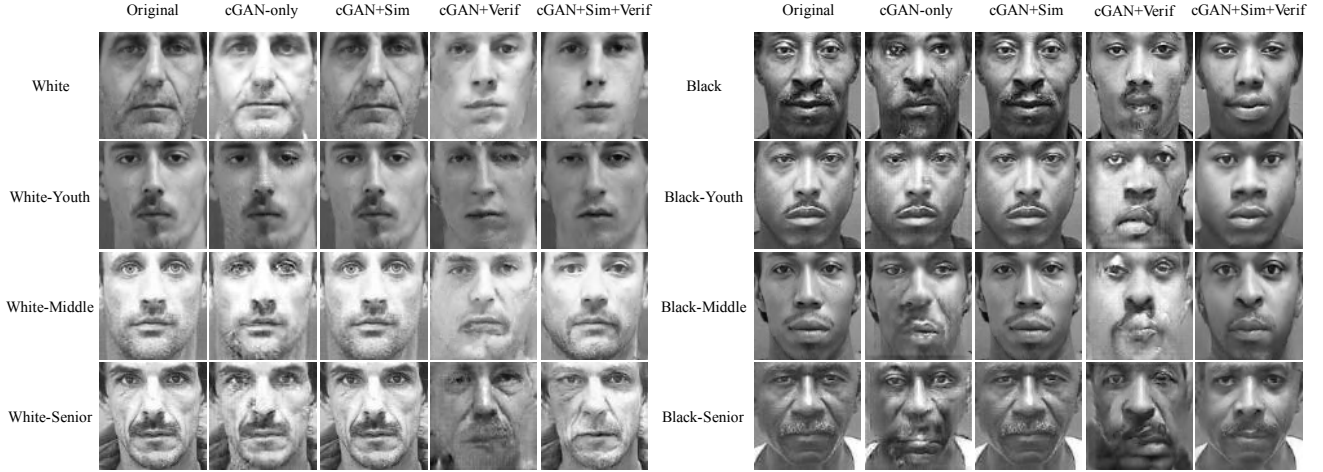

Figure 8. **The visualized results of different objective designs on different groups.** Our PP-GAN generates visually agreeable results beyond all the other settings on every group.

responding original image on *cGAN-only, cGAN + Verif, cGAN + Sim*, and *cGAN + Verif + Sim* (see Fig. 7), and display representative results (see Fig. 8). Overall, our framework maintains high id-removal rate and SSIM at the same time. One can also visually tell that the PP-GAN model maps the original face to a different face while somewhat keeps certain implicit relationship.

For the *cGAN-only* setting, it shows a high id-removal rate with uncertain SSIM for the black (Fig. 7 *Left-half*) and a low id-removal rate with high SSIM for the white (Fig. 7 *Right-half*), which may come from the different sizes of samples. When the sample size is small, the cGAN-only generates new images quite similar to existing one (Fig. 8 *Left*) and gains high SSIM. When the sample size is large, the cGAN-only generator is able to step away from the original figures but may generate improper visualizations (Fig. 8 *Right*) and provides low insurance for SSIM.

For the *cGAN+Sim* setting, the SSIM cost degenerates cGAN to the same manner of cGAN with limited samples, which functions almost as an identity mapping (Fig. 8). Thus it holds low id-removal rate and high SSIM.

For the *cGAN+Verif* setting, the verification loss here enlarges the cGAN's sampling range thus results in a high

id-removal rate (Fig. 7), while the inconsistency between the two feature spaces and lack of structure constrains may cause the visually strange consequences (Fig. 8 *Right*). An interesting phenomenon is that when the sample size is small (the white group), the *cGAN+Verif* indeed obtains a visually moderate result (Fig. 8 *left*) due to the enlarged sampling range from the verificator.

These results confirm the necessity and efficiency of simultaneously including the verificator and regulator to balance the id-removal and utility retaining.

5. Conclusion

In this paper, we present a new face de-identification framework, Privacy-Protective-GAN, which generates de-identified output according to a single input. We explicitly integrate the de-identification metric into the objective function to ensure the privacy protection. Meanwhile, we try to preserve visual similarity as much as possible to retain data utility by adding a regulator. In the experiments, we quantitatively demonstrate the effectiveness of proposed method in terms of privacy protection, utility preservation, and visual similarity.

References

[1] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *CVPRW*, 2017.

[2] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *CVPR*, 2008.

[3] D. Chen, Y. Chang, R. Yan, and J. Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007.

[4] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546, 2005.

[5] B. DeCann and A. Ross. 'has this person been encountered before?' modeling an anonymous identification system. In *IEEE Computer Society Workshop on Biometrics at the Computer Vision and Pattern Recognition (CVPR) Conference, (Providence, USA), June 2012.*, 2012.

[6] L. Du, M. Yi, E. Blasch, and H. Ling. Garp-face: Balancing privacy protection and utility preservation in face de-identification. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8, 2014.

[7] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent. Large-scale privacy protection in google street view. In *ICCV*, pages 2373–2380, 2009.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[9] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *International Workshop on Privacy Enhancing Technologies*, pages 227–242, 2005.

[10] R. Gross, L. Sweeney, J. Cohn, F. de la Torre, and S. Baker. Face de-identification. *Protecting Privacy in Video Surveillance*, pages 129–146, 2009.

[11] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *CVPRW*, pages 161–161, 2006.

[12] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Semi-supervised learning of multi-factor models for face de-identification. In *CVPR*, pages 1–8, 2008.

[13] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, pages 1735–1742, 2006.

[14] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[16] A. Jourabloo, X. Yin, and X. Liu. Attribute preserved face de-identification. In *International Conference on Biometrics (ICB)*, pages 278–285, 2015.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[19] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[20] B. Meden, R. Mallı, S. Fabijan, H. Ekenel, V. Struc, and P. Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 2017.

[21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] E. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[23] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

[24] S. Ribaric, A. Ariyaeeiniaand, and N. Pavesic. De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47:131–151, 2016.

[25] K. Ricanek and T. Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[26] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[28] A. Senior. *Protecting Privacy in Video Surveillance*, chapter Privacy Protection in a Video Surveillance System, pages 35– 47. 2009.

[29] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[30] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[32] M. J. Wilber and T. E. Boult. Secure remote matching with privacy: Scrambled support vector vaulted verification (s^2v^3). In *WACV*, pages 169–176, 2012.

[33] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv preprint arXiv:1511.02683*, 2015.

[34] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

(a)

(b)

(c)

(d)

Figure 9. **The visualized results of the black group.** In each pair of double rows, the top one is for the original faces and the bottom one is for the de-identified faces. *(a)* This panel shows the results for the whole black group. We can see that the age factor is not retained. The *(b)* shows the results of *Black-Youth* subgroup. The *(c)* shows the results of *Black-Middle* subgroup. The *(d)* shows the results of Black-Senior subgroup.

(a)

(b)

(c)

(d)

Figure 10. **The visualized results of the white group.** In each pair of double rows, the top one is for the original faces and the bottom one is for the de-identified faces. *(a)* This panel shows the results for the whole white group. We can see that the age factor is not retained. The *(b)* shows the results of *White-Youth* subgroup. The *(c)* shows the results of *White-Middle* subgroup. The *(d)* shows the results of White-Senior subgroup.