StyleGAN as a Utility-Preserving Face De-identification Method

Seyyed Mohammad Sadegh Moosavi Khorzooghi
The University of Texas at Arlington
seyyedmohammads.moosavikhorzoog@mavs.uta.edu

Shirin Nilizadeh
The University of Texas at Arlington
shirin.nilizadeh@uta.edu

Abstract

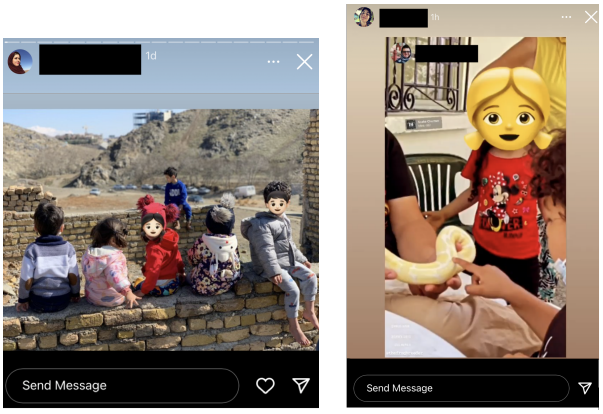
Several face de-identification methods have been proposed to preserve users' privacy by obscuring their faces. These methods, however, can degrade the quality of photos, and they usually do not preserve the utility of faces, e.g., their age, gender, pose, and facial expression. Recently, advanced generative adversarial network models, such as StyleGAN [20], have been proposed, which generate realistic, high-quality imaginary faces. In this paper, we investigate the use of StyleGAN in generating de-identified faces through style mixing, where the styles or features of the target face and an auxiliary face get mixed to generate a de-identified face that carries the utilities of the target face. We examined this de-identification method with respect to preserving utility and privacy, by implementing several face detection, verification, and identification attacks. Through extensive experiments and also comparing with two state-of-the-art face de-identification methods, we show that StyleGAN preserves the quality and utility of the faces much better than the other approaches and also by choosing the style mixing levels correctly, it can preserve the privacy of the faces much better than other methods.

Keywords: Face Obfuscation; Face De-identification; StyleGAN; Privacy; Utility; Social Media

1. Introduction

Posting photos and videos is an integral feature of the design and functioning of the most popular social network sites, such as Instagram and TikTok. usually, these photos include faces of others, e.g., family, friends, acquaintances, children, and even bystanders, who may not even be aware of being captured in a photo. However, social network sites do not provide effective functions for obfuscating or de-identifying such faces in posts, mainly because they make a profit from images of faces, as they have become valuable to everyone from marketers to police forces.

Recently, however, some users spend the time and energy to modify their photos and somehow obfuscate the faces of others before sharing them. For example, they might hide the faces by adding a black circle on them or blur the faces using some external blurring application. There is also a trend now among some users, who as is shown in Figure 1, replace the faces with some emojis, which match the subjects' gender, race, and emotion. These recent common practices show that there is a high incentive for developing a de-identification method that provides three properties: (1) *effectively hides the subjects' identity* such that humans and machines can not re-identify faces; (2) *preserves the realism of visual data*, i.e., makes de-identified subjects look realistic, and (3) *preserves the utility of the visual data*, i.e., preserves facial features, such as age, gender, pose, and expression. Providing these three properties may encourage users to employ such methods before sharing their photos online. The third property specifically may motivate social media platforms to enable such services for their users, because maintaining facial properties is essential for systems that rely on facial attributes such as age and gender to perform services such as face clustering or targeted advertisement [9], and this can be the reason that they do not provide de-identification functionality on their platforms.



(a) Example 1 (b) Example 2

Figure 1. Some social media users share their everyday experiences, obfuscating the faces of bystanders and children, using emojis matching the gender and emotion of the subjects.

Existing face de-identification methods have evolved

from *image filtering* to more advanced *face de-identification* methods. Image filtering modifies the information using common image filters, such as blurring [10, 29, 30, 45] or pixelation [4,22,30], which often give unpleasant occlusion. The more advanced face de-identification methods either make imperceptible changes to the photo to evade recognition by specific recognition algorithms [7, 8, 31, 37, 38], or substantially modify faces, thus making them unrecognizable for generic recognition algorithms [17, 40]. Some recent work has used Generative Adversarial Networks (GANs) [12] for face de-identification, where they generate synthesized objects for the k-same algorithm [5, 28, 32, 33, 33, 44]. However, these methods may not preserve the characteristics of the original face because they cannot control the image-generation process. The resulting faces may have artifacts resulting from inpainting faces of unfitting face poses, expressions, or implausible shapes.

StyleGAN [20] is a GAN, trained on a set of images, that can create high-resolution, realistic but imaginary images. Unlike traditional GANs, it can have control over the features of the generated image or face by style mixing, in which, the generated image will inherit the styles or the features of two images. This property makes StyleGAN a good candidate for generating high-quality and realistic-looking de-identified faces. In this paper, we investigate the effectiveness and robustness of StyleGAN for utility-preserving face de-identification. We explain how using StyleGAN a de-identified face can be generated, and evaluate the privacy of the generated faces, through different re-identification attacks, i.e., face detection, verification, and identification. Face++ is used to measure its utility preserving. Our findings show that StyleGAN can be used for face de-identification and performs better than some famous GAN-based de-identification methods in both privacy and utility preserving if some certain style-mixing level is used. Thus, this paper has the following contributions:

1. Utilized StyleGAN to generate de-identified faces based on the latent vectors of the target's face so that the de-identified faces look different but have the same utility features as the target.

2. Implemented extensive experiments to examine the utility- and privacy-preserving properties of StyleGAN for all possible style-mixing levels under different attack models against different re-identification attacks including verification and identification.

3. Compared StyleGAN with two recent famous GAN-based face de-identification methods.

4. Created high-quality datasets that can be used by the community to evaluate face de-identification methods.

2. Related Work

Traditional Methods and K-same Methods. Traditional methods such as pixelation, blurring, and masking obfuscate faces to hide their identity. The level of privacy-preserving of pixelation and blurring is highly dependant on the size of their blocks. To defeat face recognition systems, the block size should be very high which heavily damages utility preservation [23].

GAN-based face de-identification. GAN-based face de-identification can transfer the face to a whole new fake face, which also looks realistic. Early research on applying GANs for face de-identification began by applying parametric face models [41]. The GAN-based de-identification methods can be divided into two sets: (1) the methods which rely on conditional inpainting [19, 34, 39, 40] and (2) the manipulating facial representation methods [11, 24, 42, 43]. We used two GAN-based utility-preserving face de-identification methods to compare with StyleGAN. The reason is that they reached state-of-the-art results and are open source. DeepPrivacy [19] generator consisted of a UNet network. The method outperformed different filtering methods in face detection. They stated that their method does not follow the pose in some images correctly. CIA-GAN's [27] goal was to reach controllable anonymity by generating new realistic faces which follow temporal consistency in videos. They evaluated their systems on three datasets including CelebA, which we have also used in our evaluation. They employed several face detection and face identification systems including FaceNet to show that their system reached state-of-the-art results.

De-identification Method Evaluation. Prior research for evaluating face de-identification method can be divided into three categories: (1) privacy vs. utility analysis, (2) human or viewer experience, and (3) de-identification robustness against adversarial attacks or machine learning attacks [6]. Li et al. [25] created a user study to investigate the privacy and user preference of several identification attacks. They showed inpainting and avatar outperformed other methods in face de-identification. Chattopadhyay et al. [6] investigated the robustness of de-identification methods using re-identification attacks in different videos. In another work [13, 14], the researchers investigated data utility preservation using an SVM face expression classifier. A recent work [18] evaluated eight commonly-used de-identification methods including GAN-based ones, against three attack scenarios: face identification, verification, and reconstruction. Results showed that UP-GAN was the best or one of the bests in the reconstruction and verification attack, but for the identification attack, K-same Net performed slightly better than UPGAN and K-same. However, there was not much UPGAN performance testing of user experience or image quality except FID, which is still much worse than that StyleGAN.

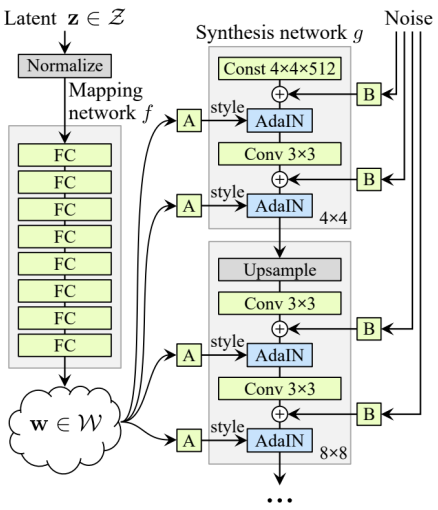


Figure 2. StyleGAN architecture [20].

3. StyleGAN for Face De-identification

StyleGAN2 is a GAN which generates high-resolution imaginary images that look more authentic compared to images produced by previous methods [21]. StyleGAN2 generator consists of different convolutional layers with an increased resolution output, starting from a learned constant (1*1) and continuing to a high resolution (1024*1024) (see Figure 2). StyleGAN2 maps the input, a random latent vector, to an intermediate latent space, which can control the generator at each convolution layer and transfer the styles related to the latent vector. It is possible to transfer the styles to selected layers, and since each layer is responsible for certain features, we can have a control on the generated image. For example, the initial convolutional layers control coarser features, such as face shape, and pose, while the last ones control finer features, such as face and hair color. Therefore, StyleGAN2 can generate a synthetic face based on two latent vectors, where initial or the coarser styles are inherited from the first face and the rest or finer styles from the second face. *Thus, we propose to employ StyleGAN2 for face de-identification, where the first face is that of a subject (target) aiming to be de-identified, and the second face is an auxiliary face, which is an imaginary StyleGAN-generated face used to transfer non-utility related styles to the de-identified face.*

Style layers. In StyleGAN2, it is possible to configure the level of style mixing by choosing $r \in \{0, 1, ..., 9\}$, representing blocks with sizes of 2^{r+1} in the generator. When setting style mixing to r, styles 0 to r are inherited from the first face, which is the target here, and styles $r + 1$ to 9 are inherited from the second face, which is the auxiliary face here. From now on, by referring to style mixing $0 - n$, we mean that styles 0 to n are inherited from the target and the rest from the auxiliary face.

Latent vector. The goal of StyleGAN is to generate an imaginary but realistic-looking face. To do style mixing, two random latent vectors are required. However, to use StyleGAN for face de-identification, instead of random latent vectors, the latent vectors for the target and auxiliary faces should be provided to the model. The latent vectors for the auxiliary faces could be any random latent vectors, which could be any imaginary face. The auxiliary faces in our work were chosen by observing their corresponding faces generated by the pre-trained StyleGAN2. Therefore, there is only need to obtain the corresponding latent vectors of the target faces. The enhanced version of StyleGAN encoder [21] was used to generate the latent vector of the target face. In particular, a ResNet encoder [1] was trained with a set of faces, as inputs, and their corresponding latent vectors, as outputs, to generate an estimate of the latent vector. The optimization of the estimate latent vector is performed using L2-optimization on the feature maps of StyleGAN2 output and the target face. The optimization was performed on the feature maps instead of the actual images to speed up the training process. The feature maps are generated using a pre-trained VGG network. The optimized latent vector is then considered a very close estimate of the target latent vector [1].

4. Methodology

To examine the effectiveness and robustness of StyleGAN for face de-identification, we first used StyleGAN to create a set of de-identified faces, then we tried identifying these faces by implementing several attacks considering various threat models, ranging from white to gray to black-box adversarial settings. We then investigated the quality and utility preserving of de-identified faces generated by StyleGAN by passing them through Face++ [3], a well-known cloud system for face analysis. Moreover, we compared the performance of the attacks and Face++ on faces generated by StyleGAN with those generated by two other state-of-the-art face obfuscation methods, CIA-GAN [27] and Deep Privacy [19].

4.1. Threat Model

We assume that the adversary can have different levels of capabilities and knowledge. We can divide this knowledge into three categories: black box, grey box, and white box. In a black-box setting, the attacker has no knowledge about the method parameters while they know all the parameters in the white-box attack. If they have partial knowledge, the attack is grey-box. The adversary's knowledge can be about: (1) the StyleGAN style levels that are used for generating the de-identified faces, (2) access to a set of target's face photos, and (3) access to a set of auxiliary faces that are

Table 1. Threat models based on adversaries' knowledge

	Attacker's Knowledge about:		
Model	Style levels	Target photos	Auxiliary photos
m_1	yes	yes	no
m_2	yes	no	yes
m_3	yes	no	no
m_4	no	yes	yes
m_5	no	yes	no
m_6	no	no	yes
m_7	no	no	no

used during the generation of the de-identified faces. Table 1 shows different grey-box attack assumptions based on the adversary's knowledge, listing them from the highest knowledge to the lowest. For example, for the threat model, m_1 in Table 1, the attacker is assumed to know about the style level and target photos, however, they do not have access to the set of auxiliary photos used for de-identification. Therefore, during the attack, a different set of auxiliary pictures are used to create datasets to train the neural network. For another example, in the threat model, m_4, the attacker does not know the styles used for de-identification, so they would use another mixing style for training. We then considered the attacker to be highly capable and has access to a lot of training data. We used the above-mentioned scenarios in our implemented identification attacks. If the adversary does not have access to any of this knowledge then the attack is called *black-box*. The face verification attack that we implemented is a black-box attack. The table does not include an adversary who has knowledge about all of the three categories, because having all that information, the attacker easily can recognize the identity. Note that if the target photos are known then the attack is *targeted*, but when the attacker has no access to the photos of the target then the attack is *untargeted*.

4.2. Data Generation

Dataset of Target Faces: CelebA dataset was used for our experiment [26]. This dataset consists of 10,177 identities with a total 202,599 number of face images, i.e., there are about 20 images for each identity. For each face image, there are 40 binary attributes along with 5 landmark locations. Because of process time constraints, we randomly selected a subset of 200 identities with at least 20 images, using an equal number of identities for males and females. Therefore, the total number of images in our dataset is 5,007, approximately 25 images per identity.

Dataset of Auxiliary Faces: We used StyleGAN2 to generate auxiliary faces. Using the trained StyleGAN2, new faces can be generated based on seed numbers given as input. Each seed represents a latent vector. Using this capability of StyleGAN2, 400 faces were generated using the first 400 seeds of the pre-trained StyleGAN2. Then, 20 of these faces, 2 categories of 10, one for training and the other for validation, were chosen. In each category, we tried to have images from different genders, ages, ethnicity, hair colors, etc. Each of these four categories is used for different experiments. For efficiency and to save computation time, we reduced the resolution of images from 1024*1024 to 256*256.This didn't make any problem because this size is larger than than the sizes of all images in the dataset.

Generating De-identified Faces: The de-identified faces were created by using StyleGAN2 and mixing 5,007 target faces with the 20 auxiliary faces and using all the 9 possible style mixing levels, 0-0 to 0-8. We generated the de-identified faces using all the style mixing levels because they determine the level of information (utility) that is passed from the targets' faces into the de-identified faces. For example, in 0-0, the first style, 0, is inherited from the target, and styles 1 to 9 from the auxiliary, so the output image will be almost identical to the auxiliary face, while for 0-8, styles 0 to 8 are inherited from the target and style 9 from the auxiliary face. In this case, the generated face will be almost identical to the target face. Therefore, there is a trade-off between preserving the utility of a target's face and protecting their identity. Each face was mixed with 20 auxiliary faces generating $20 * 5,007 = 100,140$ StyleGAN-generated faces for each style. Therefore, there will be $9 * 100,140 = 901,260$ StyleGAN-generated faces in total. In identification attacks, we use some of these de-identified faces for training and some for validation.

We also employed two state-of-the-art GAN-based de-identification methods: Deep Privacy [19] and CIAGAN [27] to generate de-identified faces for the 5,007 face images. Deep Privacy used conditional neural networks to generate de-identified realistic faces while retaining pose and background. CIAGAN also uses conditional neural networks to reach anonymity by removing identifiable information from the face or body in photos and videos. While Deep Privacy could generate a de-identified image for each face in the dataset, CIAGAN failed to generate de-identified faces for 831(17%) images due to failing to detect faces in its processing phase, in which the dataset is prepared for de-identification. Therefore, one advantage of StyleGAN and Deep Privacy compared to CIAGAN is that they can successfully generate faces for all faces in the dataset.

Figure 3 shows the generated images of the nine mixing levels (0-0 to 0-8) mixed with an auxiliary photo along with de-identified photos obtained from Deep Privacy and CIAGAN for 4 different target samples chosen from our datasets. See how the StyleGAN faces change from the most similar to the auxiliary photo (0-0) to the most similar to the target photo (0-8). StyleGAN-generated photos also seem to be more natural and have a better quality than Deep Privacy and CIAGAN. StyleGAN modifies the background of faces while other methods do not change it. However, the background can easily be extracted and replaced with that

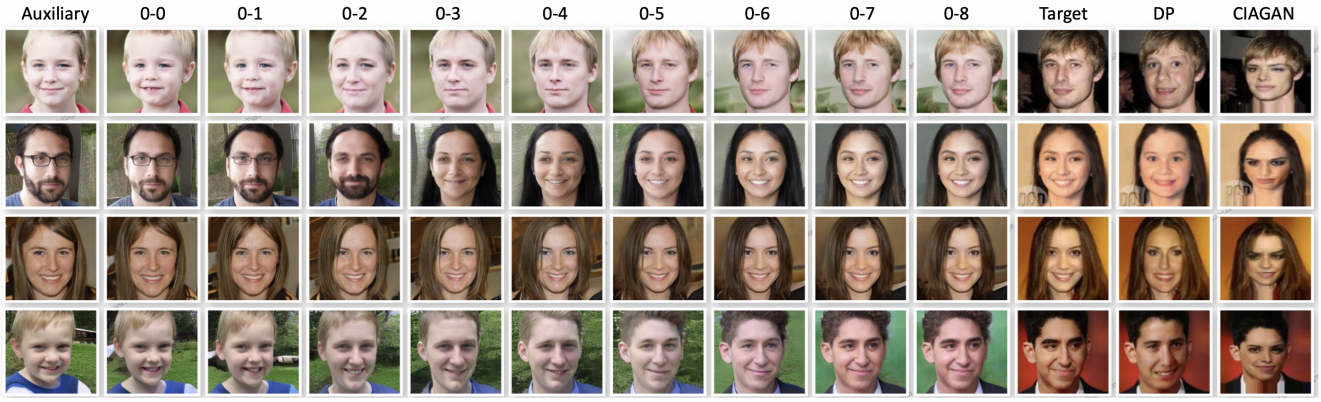


Figure 3. StyleGAN-generated images for 4 samples along with the corresponding de-identified versions made by Deep Privacy and CIAGAN.

of the generated photo [1].

5. Utility and Privacy Analysis

After creating datasets of de-identified faces for different style-mixing levels, we examine how well StyleGAN preserves the privacy of the targets and the utility of their faces, compared to CIAGAN and Deep Privacy.

5.1. Face Detection

Some face de-identification methods are not even able to generate a face [15]. Face quality has also been defined as the usability of an image for the purpose of recognition [15]. Therefore, to measure the effectiveness of StyleGAN in generating recognizable faces, we passed all the de-identified images through the Face++ detection module [2]. We found that the Face++ detection module could successfully identify a face in all StyleGAN-generated photos, while it failed for 29 (%0.6) and 44 (%1.1) images generated by Deep Privacy and CIAGAN, respectively. Note that we had 20 (auxiliary photos) * 9 (styles) = 180 times more images for StyleGAN compared to the other two methods and there was not even one photo without any face detected. Therefore, we can conclude that the quality of StyleGAN-generated faces is higher than that of CIAGAN and Deep Privacy. We manually checked all the photos that Face++ failed to detect a face in them and found that they actually did not have any face in them.

5.2. Utility Preserving Evaluation

Many works have tried to develop utility-preserving face de-identification methods [24, 34, 39, 44]. Transferring non-identity-related features or the utility of the face to the de-identified face not only increases the quality of the photo and its visual appeal but also the de-identified photo can still be analyzed by applications, such as recommendation systems, that provide services based on these attributes. With recent advances in image processing and recognition, many companies, including Face++ [3], offer face analysis services to extract face attributes, such as age, gender, emotion, smile, eye status, head pose, mouth status, blurriness, ethnicity, etc. These attributes are also called utility. Therefore, we employed some experiments using the Face++ API to check the utility preserving of StyleGAN2. We did not evaluate the effectiveness of StyleGAN in preserving race, because this feature does not exist in Face++.

Experiments: As mentioned before, 9 datasets of 100,140 faces were used. The attributes are age, gender, emotion, smile, eye status, head pose, mouth status, blurriness, and ethnicity. Gender could be male or female. Smile is presented with a real number between 0 and 100. Emotion is presented as percentages for 6 different emotion states: anger, disgust, fear, happiness, neutral, sadness, and surprise. Mouth status has 4 states: surgical mask or respirator, other occlusions, close, and open, and eye status has 6 states for each eye: occlusion, no glass with an open eye, no glass with a closed eye, normal glass with an open eye, normal glass with a closed eye, dark glasses, and occlusion. The head pose consists of three angles between -180 and 180 degrees: pitch angle, roll angle, and yaw angle.

Metrics: To compare the results of the attributes for the original face and the de-identified face, the mean and standard deviation of differences are calculated for age, blurriness, and smile. For the gender, the ratio of times when the genders are the same is presented. For the emotion, and the eye and mouth statuses, the state with maximum value is chosen, and the ratio of times when the original and de-identified face have the same chosen state. For the head pose, the mean of differences for each angle is calculated, and then, the mean of the three numbers is presented as the

final value for the head pose difference.

Results: Table 2 shows the results of attribute preserving for different style mixing levels, CIAGAN, and Deep Privacy. The values of the first 4 attributes, gender, emotion, eye status, and mouth status, are matching rates in percent (0-100). As expected, our results show that the higher the mixing style is the utility is more preserved in the generated faces. For example, while gender and emotion are preserved for 47% and 51% of the faces when the mixing style is 0-0, these values are 98% and 84% preserved when the mixing style is 0-8. This is because when more styles are inherited from the target, thereby have more similar attributes to the target. Interestingly, our results show that the attribute values for style 0-3 on-wards are as good as or better than those of CIAGAN and Deep Privacy. For example, gender and emotion are preserved for 81% and 52% of de-identified faces generated by CIAGAN, and 89% and 60% of de-identified faces generated by Deep Privacy, while they are preserved for 88% and 73% of de-identified faces generated by StyleGAN when style mixing is as low as 0-3.

The values for the other 4 attributes, age, smiling, head pose, and blurring, decrease as style mixing level n increases since they indicate the difference between attribute values. In other words, the lower the metric values are the more the attributes of the original and de-identified faces match. The maximum difference when two attributes are completely different is 100. Our results show very small values as small as 5, 6, 2.1, and 9 for age, smiling, head pose, and blurring when StyleGAN 0-8 is used to generate the de-identified faces. Interestingly, the blurring for StyleGAN is much less than that for CIAGAN and Deep privacy, no matter what style mixing is used. We observe that StylGAN with style mixing levels 0-3 to 0-8 has better performance preserving the smiling and head pose attributes compared to CIAGAN and Deep Privacy. The only attribute that is not as well-preserved as others when using StyleGAN is age, where CIAGAN and Deep Privacy show better results, i.e., 10, compared to 16 and 13 for StyleGAN 0-3 and 0-4, respectively.

5.3. Verification Attack

Face de-identification methods are mostly evaluated against two types of attacks [16]: *verification* and *identification* attacks. In a verification attack, which is a targeted attack, the attacker has a suspect and tries to verify if the de-identified face belongs to that suspect. In practice, in this attack two face images are given to a face recognition system, and the system should report to what extent they belong to the same person. One of the images contains the unknown or the de-identified face, and the other is a face image of the target or suspect. We used Face++ to implement our verification attacks.

Here we can define two attack scenarios: (1) when the attacker has the exact image that the target might have used for generating their de-identified face, and (2) when the attacker does not have access to the exact photo, but has another image of the target. We hypothesize that it would be harder for the attacker to recognize the identity in the second scenario.

Experiments: All the de-identified images with their corresponding non-de-identified images are given to the Face++ *compare* module, which returns a confidence score for each pair indicating to what extent (from 0 to 100) the comparing module finds the identity of the faces is the same. The attack is successful if the system has high confidence in two images being belonged to the same person. In our experiments, we specify a threshold variable, ranging from 0 to 100, and for every image pair if their obtained confidence score is more than a given threshold, then we label that verification attack, a success for the attacker and a failure for the de-identification method, i.e., the de-identification is not successful in obscuring the face, and vice versa.

Figures 4 and 5 show the average de-identification success rate (or 1 - attack success rate) for different threshold values, from 0 to 100, for StyleGAN with different style-mixing levels, CIAGAN, and Deep Privacy. Figure 4 shows the results for the first scenario when the attacker has access to the same image used for de-identification, and Figure 5 shows the results for the second scenario when the attacker has access to another image of the target. The curves have shifted upwards in scenario 2 compared to scenario 1 confirming our hypothesis that it is harder to re-identify a face when the attacker does not have access to the exact image that the target has de-identified. The results for both scenarios are pretty consistent. The higher the threshold, the de-identification success rate is higher and the attack success rate is lower. As expected, de-identified faces generated by StyleGAN 0-0 and 0-1 cannot be re-identified, even when using a very low confidence threshold of 55. The de-identified faces generated by higher style-mixing levels are easier to re-identified, especially those generated by 0-7 and 0-8. Interestingly, the performance of Deep Privacy is better than CIAGAN and the performance of StyleGAN with style mixing levels 0-3 is far better than both. Results for StyleGAN with style mixing levels 0-4 are comparable with those of Deep Privacy and still far better than those of CIAGAN. Although de-identification success rates for StyleGAN 0-5 are worse than those of Deep Privacy, they are near those of CIAGAN.

Discussion: Remembering the attribute analysis results, StyleGAN with style-mixing levels 0-3 onwards were as good as or better than the state-of-the-art de-identification methods in terms of high utility. Here we observe that the de-identification success rates are far better for 0-3. Therefore, based on these results, we argue that StyleGAN, even though was not designed for face obfuscation, performs on

Table 2. Utility preserving of StyleGAN, CIAGAN, and Deep Privacy for 8 attributes using Face++. Gender, emotion, eye status, and mouth status values are the rate (percentage) that the attributes match for the target and de-identified image, while for age, smiling, head pose, and blurring, the mean and standard deviation of absolute differences are shown.

Method	gender	emotion	eye	mouth	age		smiling		head pose		blurring	
					M	Std	M	Std	M	Std	M	Std
StyleGAN 0-0	47%	51%	76%	46%	19	24	42	58	5.9	5	10	22
StyleGAN 0-1	48%	51%	80%	47%	19	24	43	59	5.2	4.4	10	22
StyleGAN 0-2	60%	62%	84%	50%	17	21	30	47	4.3	3.8	10	22
StyleGAN 0-3	88%	73%	87%	65%	16	20	16	32	3.3	3.1	10	22
StyleGAN 0-4	94%	77%	89%	75%	13	17	12	26	3.1	2.9	10	22
StyleGAN 0-5	96%	80%	91%	81%	10	13	10	23	2.8	2.7	10	22
StyleGAN 0-6	96%	82%	92%	83%	7	10	8	20	2.6	2.5	9	22
StyleGAN 0-7	97%	83%	93%	85%	6	8	7	17	2.3	2.3	9	21
StyleGAN 0-8	98%	84%	93%	86%	5	7	6	15	2.1	2.1	9	21
CIAGAN	81%	52%	73%	72%	10	12	25	39	3.8	4.1	26	39
Deep Privacy	89%	60%	87%	57%	10	13	36	46	3.6	3.7	16	29

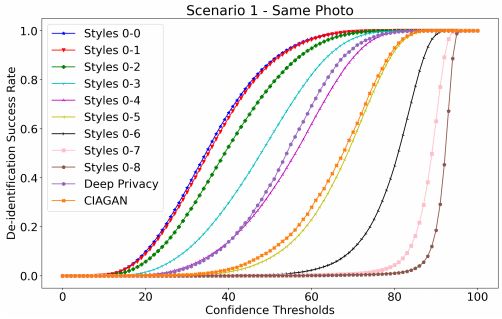


Figure 4. Verification Attack results Scenario 1: De-identification success rates when the de-identified photo has been obtained by the de-identification of the given target photo. Based on the attribute analysis results, Styles 0-3 and 0-4 are better than CIAGAN and Deep Privacy if we consider both utility and privacy.

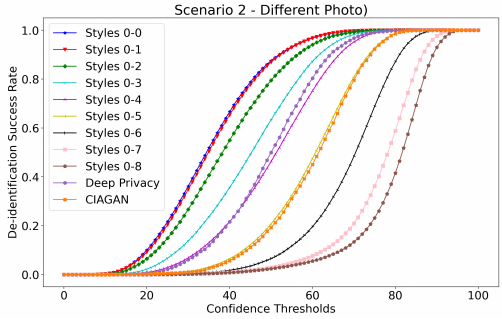


Figure 5. Verification Attack results Scenario 2: De-identification success rates when the de-identified photo has been obtained from an unknown photo of the target identity. The results are consistent with scenario 1.

par or even better than the state-of-the-art de-identification methods. In addition, because of the ability to choose the mixing levels, the users have the ability to tune the trade-off between privacy and utility. For example, if the user is more concerned about privacy, they can choose StyleGAN with style-mixing levels 0-3, and still, the output image preserves adequate utility, or if the user prefers better quality and more utility-preserving de-identified faces, they can choose StyleGAN with style-mixing levels 0-4 and still obtain de-identified images that preserve privacy the same as other state-of-the-art de-identification methods.

5.4. Identification Attack

In the identification attack, which is an untargeted attack, the attacker aims to identify any unknown face by matching it to a person in a set of known faces. We used FaceNet [36], a common tool for evaluating face de-identification methods [27, 35]. FaceNet is a face recognition model, developed by Google's researchers which reached a state-of-the-art accuracy of 99.63 when trained on the Google imageset. FaceNet outputs a distinct feature vector, called face embedding, with a size of 128 for each image. Obtaining the embedding for all the images, then, they can be fed to a classifier, e.g., a Support Vector Machines (SVMs) to perform the identification (i.e., classification).

Experiments: For this attack, we implemented all the threat models listed in Table 1. The attacker is assumed to be powerful as they know StyleGAN2 and how to use it to create mixed images. We used all images for training and validation. However, the number of images for training and validation changed based on the threat model. For example, if the attacker did not know the auxiliary photos used for de-identification, we considered half of the dataset for training and the rest for validation, while when the attacker did not know the target photo, we split 70% of images for training and 30% for validation. When the attacker does not know any of the auxiliary photos or the target photo, both conditions were applied, i.e., 35% of images were used for training and 15% for validation. For threat models, m4 to

Table 3. Identification attack results (T: training accuracy, V: validation accuracy) using FaceNet and SVMs. The validation accuracy values for styles 0-3 and 0-4 are better than those of CIAGAN and Deep Privacy which is compatible with the verification results.

Method	m1		m2		m3		m4		m5		m6		m7	
	T	V	T	V	T	V	T	V	T	V	T	V	T	V
StyleGAN 0-0	9.8	.7	6.3	1.8	5.5	1.9	5	12.8	9.8	16.2	4.3	5	9.2	11.1
StyleGAN 0-1	10.2	.9	6.7	2.3	6	2.4	5.5	13.4	10.2	17.6	5.4	6	9.6	11.5
StyleGAN 0-2	15.3	2.7	98.7	2.3	10.9	6	10.5	14.2	31.7	24.7	13.1	7.4	15.3	9.9
StyleGAN 0-3	31.7	11.7	98.7	3.8	27.5	18.6	26.8	25	31.7	24.7	40.8	13.7	39.1	14.8
StyleGAN 0-4	50.1	27	98.7	6.7	46.4	34.2	45.7	33.7	50.1	33.3	67.6	20	65.3	20
StyleGAN 0-5	76.5	58.2	98.7	18.2	75.7	63.4	75.4	41.2	76.5	40.2	93.9	25.8	92	26.2
StyleGAN 0-6	93.6	87.8	96.4	67.3	94.3	88.6	94.4	39.6	93.6	38.9	99.4	27.5	99.1	27.3
StyleGAN 0-7	99	98.6	99.7	82.4	99.3	98.3	99.3	27.7	99	26.7	100	19.5	99.9	19
StyleGAN 0-8	99.6	99.5	98.7	84.4	99.8	99.3	99.8	16.2	99.6	34.2	100	25.2	100	24.8

m7, the attacker does not know the style-mixing level. In that case, we randomly sampled 20% of whole images of other styles for validation with other assumptions applied. For CIAGAN and Deep Privacy, we simply considered one threat model in which the attacker has access to 70% of the identities and can create de-identified photos for training, and use the trained model to recognize the identity of the de-identified target photos (30% of the photos). Note that the number of classes or identities here is all 200 identities we used for our experiments.

Results: Table 3 shows the results of the identification attacks for all style-mixing levels and the seven threat models. As expected, the de-identified faces generated by higher style-mixing levels are easier to re-identified, especially those generated by StyleGAN 0-6 upwards. For example, in the threat model m2 that the attacker has knowledge about both the style levels and auxiliary photos, the accuracy of the identification attack (validation) for style-mixing levels 0-2, 0-3, 0-4, 0-5, 0-6, 0-7 and 0-8 are 2.3, 3.8, 6.7, 18.2, 67.3, 82.4, and 84.4. However, there are a few exceptions. For example, in threat models m4 and m5, the accuracy increases until 0-5 and achieves 41.2 and 40.2, respectively, but then it starts decreasing from style levels 0-6, and it gets to 16.2 and 34.2 for style levels 0-8. We see a similar but less severe trend for threat models m6 and m7 as well. The common setting among these threat models is that the attacker does not know the style levels, and based on our results, this knowledge is learned in the training phase very well, especially when de-identified images are generated using higher style levels and therefore, during the validation, the classifier does not perform well as the images are generated using other style levels. This can explain the huge difference between the accuracy of the training and validation phases for these four threat models.

Comparing the validation results for different threat models, we observe that having knowledge about the style levels has the biggest impact on the accuracy of the identification attack. We observe that if the attacker has this knowledge, i.e., in threat models m1, m2, and m3, the accuracy is much higher compared to the threat models that do not have this assumption, i.e., m4, m5, m6, and m7. For example, for style levels, 0-6, the validation accuracy for m1, m2, m3, m4, m5, m6, and m7 are 87.6, 67.3, 88.6, 39.6, 38.9, 27.5, and 27.3. Note that this assumption that the attacker has knowledge about the style levels is a strong assumption and in practice might not be realistic.

Running the identification attack on images generated by CIAGAN and Deep Privacy, we obtained 31.2 and 30.1. Note that this threat model is black-box, which is equivalent to m7 in our table of threat models. Comparing their accuracy with those of m7, we see that no matter the style level, it is harder to re-identify faces generated by StyleGAN, as the best verification accuracy for m7, is 24.8. Moreover, even no matter the threat model, we observe that the performance of the attack, when StyleGAN 0-3 and 0-4 are used, is better or on par with that of CIAGAN and Deep Privacy.

Overall, through our extensive experiments, we could show that StyleGAN is a better de-identification method in terms of both privacy and utility preserving, especially if style-mixing levels 0-3 and 0-4 are used; 0-4 offers preserve utility better while 0-3 preserves privacy better.

6. Conclusion

In this paper, we investigated if StyleGAN can be used as a utility-preserving face de-identification method. We used Face++ to assess the quality of generated de-identified faces. All StyleGAN-generated faces were detected, but this was not the case for the images generated by the other two state-of-the-art methods, CIAGAN and Deep Privacy. For utility preserving analysis, we employed Face++ to quantify 8 face attributes of all the generated faces. We found that style-mixing levels from 0-3 to 0-8 transfer the utility to the generated face as good as or better than CIAGAN and Deep Privacy. We also implemented verification and re-identification attacks. The results indicated that style-mixing levels 0-0 to 0-4 preserve privacy more than CIAGAN and Deep Privacy. Considering the findings of the utility and privacy evaluation, we conclude that StyleGAN with style levels 0-3 and 0-4 provides better performance compared to CIAGAN and Deep Privacy.

References

[1] Encoding images into stylegan's latent space. https://colab.research.google.com/drive/1jrSki9OXahtnS2Okcf7_ubvLkPnboJey. 3, 5

[2] Face++ - Face++ Cognitive Services. https://www.faceplusplus.com/. 5

[3] Face plus plus. https://www.faceplusplus.com/. 3, 5

[4] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 1–10, New York, NY, USA, 2000. ACM. 2

[5] Karla Brkic, Ivan Sikiric, Tomislav Hrkac, and Zoran Kalafatic. I know that person: Generative full body and face de-identification of people in images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1319–1328. IEEE, 2017. 2

[6] Ankur Chattopadhyay, Robert Ruska, and Levi Pfantz. Determining the robustness of privacy enhancing deid against the reid adversary: An experimental study. In *The 16th International Conference on Availability, Reliability and Security*, pages 1–11, 2021. 2

[7] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021. 2

[8] Ivan Evtimov, Pascal Sturmfels, and Tadayoshi Kohno. Foggysight: a scheme for facial lookup privacy. *arXiv preprint arXiv:2012.08588*, 2020. 2

[9] GWIR Fonseka and HMM Naleer. Face-recognition billboard display to target advertisement with gender and age recognizing. 2019. 1

[10] Andrea Frome, German Cheung, Ahmad Abdulkader, Marco Zennaro, Bo Wu, Alessandro Bissacco, Hartwig Adam, Hartmut Neven, and Luc Vincent. Large-scale privacy protection in google street view. In *2009 IEEE 12th international conference on computer vision*, pages 2373–2380. IEEE, 2009. 2

[11] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[13] Ralph Gross, Edoardo Airoldi, Bradley Malin, and Latanya Sweeney. Integrating utility into face de-identification. In George Danezis and David Martin, editors, *Privacy Enhancing Technologies*, pages 227–242, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 2

[14] R. Gross, L. Sweeney, F. de la Torre, and S. Baker. Model-based face de-identification. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 161–161, 2006. 2

[15] Patrick J Grother, Mei L Ngan, Kayee K Hanaoka, et al. Face recognition vendor test part 3: demographic effects. 2019. 5

[16] Hanxiang Hao, David Güera, János Horváth, Amy R Reibman, and Edward J Delp. Robustness analysis of face obscuration. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 176–183. IEEE, 2020. 6

[17] Hanxiang Hao, David Güera, Amy R Reibman, and Edward J Delp. A utility-preserving gan for face obscuration. *arXiv preprint arXiv:1906.11979*, 2019. 2

[18] Hanxiang Hao, David Güera, János Horváth, Amy R. Reibman, and Edward J. Delp. Robustness analysis of face obscuration. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 176–183, 2020. 2

[19] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019. 2, 3, 4

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018. 1, 2, 3

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 3

[22] Itaru Kitahara, Kiyoshi Kogure, and Norihiro Hagita. Stealth vision for protecting privacy. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 404–407. IEEE, 2004. 2

[23] Karen Lander, Vicki Bruce, and Harry Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 15(1):101–116, 2001. 2

[24] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 5

[25] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–24, 2017. 2

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 4

[27] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 4, 7

[28] Blaž Meden, Refik Can Mallı, Sebastjan Fabijan, Hazım Kemal Ekenel, Vitomir Štruc, and Peter Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 11(9):1046–1054, 2017. 2

[29] Carman Neustaedter and Saul Greenberg. The design of a context-aware home media space for balancing privacy and awareness. In *International Conference on Ubiquitous Computing*, pages 297–314. Springer, 2003. 2

[30] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, Mar. 2006. 2

[31] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017. 2

[32] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 791–799. IEEE, 2019. 2

[33] Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. Protecting visual secrets using adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1329–1332. IEEE, 2017. 2

[34] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. 2, 5

[35] Harrison Rosenberg, Brian Tang, Kassem Fawaz, and Somesh Jha. Fairness properties of face recognition and obfuscation systems. *arXiv preprint arXiv:2108.02707*, 2021. 7

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 7

[37] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020. 2

[38] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2

[39] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. 2, 5

[40] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018. 2

[41] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. A hybrid model for identity obfuscation by face replacement. *ECCV*, LNCS 2016, 2018. 2

[42] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3281–3289, 2021. 2

[43] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3281–3289, June 2021. 2

[44] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34(1):47–60, 2019. 2, 5

[45] Cha Zhang, Yong Rui, and Li-wei He. Light weight background blurring for video conferencing applications. In *2006 International Conference on Image Processing*, pages 481–484. IEEE, 2006. 2