Adversarial representation learning for synthetic replacement of private attributes

John Martinsson **Edvin Listo Zec** **Daniel Gillblad** **Olof Mogren**
RISE Research Institites of Sweden

Abstract

The collection of large datasets allows for advanced analytics that can lead to improved quality of life and progress in applications such as machine cognition and medical analysis. However, recently there has been an increased pressure to guarantee the privacy of users when collecting data. In this work, we study how adversarial representation learning can be used to ensure the privacy of users, and to obfuscate sensitive attributes in existing datasets. While previous methods using adversarial representation learning for privacy only aims at obfuscating the sensitive information, we find that adding new information in its place can improve the strength of the provided privacy. We propose a method building on generative adversarial networks that has two steps in the data privatization. In the first step, sensitive data is removed from the representation. In the second step, a sample which is independent of the input data is inserted in its place. The result is an approach that can provide stronger privatization on image data, and yet be preserving both the domain and the utility of the inputs.

1 Introduction

The list of success stories from big data analytics can be made long. Statistical studies for medical diagnosis [18] and business analytics [5] are merely a couple of examples. The amounts of data required for training machine learning models for the processing of natural language and images are growing with the scale and capabilities of the models [8]. However, collecting large datasets which potentially contain sensitive information about individuals can be difficult because getting the consent from people may be challenging. Furthermore, privacy laws are being incorporated in many countries to protect the rights of individuals, making it hard to use data with sensitive information. Being able to give privacy guarantees on a dataset may be a way to allow the distribution of the data, while protecting the rights of individuals, and thus unlocking the large benefits for individuals and for society that big datasets can provide.

In this work, we study techniques for selective anonymization of image datasets. The end goal is to provide the original data as detailed as possible and retain the most information from it, while at the same time making it hard for an adversary to detect specific sensitive attributes. The proposed solution is agnostic to the downstream task, with the objective to make the data as private as possible given a distortion constraint.

Previous research has addressed this issue using adversarial representation learning with some success: a filter model is trained to hide sensitive information while an adversary model is trained to recover the information. While previous work had the objective of training the filter model so that it is hard for the adversary to uncover the sensitive attributes [12], we instead explore this task under the assumption that *it is easier to hide sensitive information if you replace it with something else*: a sample which is independent from the input data.

In our setup, the adversary can make an arbitrary number of queries to the model, each time another sample will be produced from the distribution of the sensitive data, while keeping as much as possible of the non-sensitive information about the requested data point.

Besides the adversary module, our proposed solution includes two main components: one filter model that is trained to remove the sensitive information, and one generator model that inserts a synthetically generated new value for the sensitive attribute. The generated sensitive information is entirely independent from the sensitive information in the original input image. Following a body of work in privacy-related adversarial learning we evaluate the proposed model on faces from the CelebA dataset, and consider, for example, the smile of a person to be the sensitive attribute. The smile is an attribute that carries interesting aspects in the transformations of a human face. Even if the obvious change reside close to the mouth, subtle changes occur in many other parts of the face when a person smiles: eyelids tighten, dimples may occur and the skin may wrinkle. The current work also includes a thorough analysis of the dataset, including correlations of such features. These correlations make the task interesting and challenging, reflecting the real difficulty that may occur when anonymizing data. What is the right trade-off between preserving the utility as defined by allowing information about other attributes to remain, and removing the sensitive information? Our method is easily generalizable and can be applied to other visual attributes such as gender, race, background cues, etc.

2 Related work

Adversarial learning is the process of training a model with the objective of being able to fool an adversary (a second model). The adversary is trained simultaneously, and both become increasingly good at their respective task during the training process. This approach has been successfully used to learn image-to-image transformations [15, 6], and synthesis of properties such as facial expressions [29, 30]. Privacy-preserving adversarial representation learning studies how to utilize this learning framework to learn representations of data that hide some sensitive information, yet retain the utility. Adversarial learning has also been used to train generative models [9].

A body of prior work has been dedicated to employ adversarial representation learning to hide sensitive attributes while retaining the useful information [7, 36, 34, 3, 25]. [2] presented a privacy preserving mechanism that minimizes the mutual information between the utility variable and the input image data conditioned on the learned representation. [4] proposed that the generator in the adversarial learning setup should be optimized to maximize the entropy of the discriminator output rather than to minimize the log likelihood. The authors show that this is beneficial in the multi-class setting. [24] treated the problem as an information-bottleneck problem. The resulting images are optimized to be useful to predict a specific binary attribute, while hiding the identify of a person. [33], [32], [26] studied how to learn transformations of video that respect a privacy budget while maintaining performance on the downstream task. [31] proposed an approach for solving pose-invariant face recognition. Similar to our work, their approach used adversarial representation learning to disentangling specific attributes in the data. [22] trained an obfuscator network to add a minimal amount of perturbation noise to the input to make the output fool a person recognition adversary.

All these proposed solutions, with the exception of [7], depend on knowing the downstream task labels. Our work has no such dependency: the data produced by our method is designed to be usable regardless of downstream task.

The work most closely related to ours is that by [13] and [12]. They use adversarial learning to minimize the mutual information between the private attribute and the censored image under a distortion constraint. We extend on these ideas by proposing a modular design consisting of a filter that is adversarially trained to obfuscate the data points, and a generator that further enhances the privacy by adding new independently sampled synthetic information for the sensitive attributes. Our work is closely connected to the learning of fair representations [35], and the proposed generator module could be applied to counterfactual reasoning in a fashion similar to [16] by allowing to control the private attribute input to the generator deterministically.

Some prior work has assumed access to a privacy-preserving mechanism, such as bounding boxes for faces, and has studied to what extent the identity of a person can be hidden when blurring [20], removing [23] or generating the face of another person [14] in its place. Other work has assumed

access to the utility-preserving mechanism and proposed to obfuscate everything except what they want to retain [1]. But this raises the question: how do we find the pixels in an image that need to be modified to preserve privacy with respect to some attribute, or alternatively the pixels that need to be kept to preserve utility? Furthermore, [21] showed that blurring or removing the head of a person has a limited effect on privacy with respect to person recognition. The finding is crucial; we cannot rely on modifications of an image such as blurring or overpainting to achieve privacy.

An adversarial set-up instead captures the signals that the adversary uses, and can attain a stronger privacy.

3 Privacy-preserving adversarial representation learning

In the current work, we focus on utility-preserving transformations of data: we use privacy-preserving representation learning to obfuscate information in the input data, and seek to output results that retain the information and structure of the input.

3.1 Problem setting

Generative adversarial privacy (GAP) [12] was proposed as a method to provide privacy while maintaining the utility of an image dataset, which will be used as the baseline in the current work. In GAP, one assumes a joint distribution $P(X, S)$ of public data points X and sensitive private attributes S where S is typically correlated with X. The authors define a privacy mechanism $X' = f(X, Z_1)$ where Z_1 is the source of noise or randomness in f. Let $h_f(X')$ be an adversary's prediction of the sensitive attribute S from the privatized data X' according to a decision rule h_f. The performance of the adversary is thus measured by a loss function $\ell(h_f(f(x, z_1)), s)$ and the expected loss of the adversary with respect to X, S and Z_1 is

$$L_f(h_f, f) = \mathop{\mathbb{E}}_{\substack{x,s\sim p(x,s)\\ z_1\sim p(z_1)}}[\ell_f(h_f(f(x, z_1), s)], \tag{1}$$

where $p(z_1)$ is the source of noise.

The privacy mechanism f should be privacy-preserving and utility-preserving. That is, it should be hard for an analyst to infer S from X', but X' should be minimally distorted with respect to X. The authors [12] formulate this as a constrained minimax problem

$$\begin{aligned}\min_f \max_{h_f} & -L_f(f, h_f)\\ \text{s.t. } & \mathop{\mathbb{E}}_{\substack{x,s\sim p(x,s)\\ z_1\sim p(z_1)}}[d(f(x, z_1), x)] \le \epsilon_1,\end{aligned}$$

where the constant $\epsilon_1 \ge 0$ defines the allowed distortion for the privatizer and $d(\cdot, \cdot)$ is some distortion measure.

In the current work, we call the privacy mechanism f the *filter* since the purpose of $f(x, z_1)$ is to filter the sensitive information from x. A potential drawback with this formulation is that it only removes the sensitive information in x which may make it obvious to the adversary that x' is a censored version of x. Instead, in addition to removing the sensitive information we propose to replace it with a new value s' that is sampled from the uniform distribution[1], independent of s.

3.2 Our contribution

We extend the filter with an additional module which we call the *generator*. We define the generator mechanism as $X'' = g(f(X, Z_1), S', Z_2)$ where S' denotes the random variable of the new synthetic sensitive attribute. Z_1 and Z_2 denote the sources of randomness in f and g respectively. The discriminator h_g is trained to predict s when the input is a real image, and to predict the "$fake$" output when the input comes from g as in the semi-supervised learning setup in [28]. The objective of the generator $g(x', s', z_2)$ is to generate a new synthetic (independent) sensitive attribute s' in x', that will fool the discriminator h_g. We further define the loss of the discriminator h_g as

[1]The uniform distribution is a close approximation of the marginal distribution of the smiling attribute.

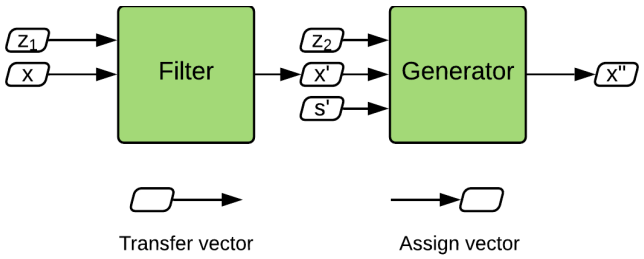

Figure 1: An overview of the training setup (figure does not show the two discriminators).

$$L_g(h_g, g) = \mathop{\mathbb{E}}_{\substack{x,s\sim p(x,s)\\ s'\sim p(s')\\ z_1,z_2\sim p(z_1,z_2)}}[\ell_g(h_g(g(f(x,z_1),s',z_2)),fake)] + \mathop{\mathbb{E}}_{x,s\sim p(x,s)}[\ell_g(h_g(x),s)],$$

where $p(z_1, z_2)$ is the source of noise, $p(s')$ is the assumed distribution of the synthetic sensitive attributes s', $fake$ is the fake class, and ℓ_g is the loss function. We formulate this as a constrained minimax problem

$$\min_g \max_{h_g} -L_g(g, h_g) \quad \text{s.t.} \mathop{\mathbb{E}}_{\substack{x,s\sim p(x,s)\\ s'\sim p(s')\\ z_1,z_2\sim p(z_1,z_2)}}[d(g(f(x,z_1),s',z_2),x)] \leq \epsilon_2,$$

where the constant $\epsilon_2 \geq 0$ defines the allowed distortion for the generator. An overview of the setup can be seen in Figure 1.

3.3 Data-driven implementation

Typically, we do not have access to the true data distribution $P(X, S)$, but we have access to a dataset of samples $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^n$ which are assumed to be identically and independently distributed according to the unknown joint distribution $P(X, S)$. We assume that the sensitive attribute is binary and takes values in $\{0, 1\}$. However, the proposed approach can easily be extended to categorical sensitive attributes. We model the filter mechanism $f(X, Z_1; \theta_f)$ with a convolutional neural network parameterized by θ_f and the generator mechanism $g(X', S', Z_2; \theta_g)$ with a convolutional neural network parameterized by θ_g. We use the UNet [27] architecture for both the filter and the generator, as illustrated in Figure 2. The orange blocks are convolution blocks each of which, except for the last block, consist of a convolution layer, a batch normalization layer and a rectified linear activation unit, repeated twice in that order. The number of output channels of the convolution layers in each block has been noted in Figure 2. The last convolution block with a 3 channel output (the RGB image) consists of only a single convolutional layer followed by a sigmoid activation. The green blocks denote either a max pooling layer with a kernel size of two and a stride of two if marked with "/2" or a nearest neighbor upsampling by a factor of two if marked with "2x". The blue block denotes an embedding layer, which takes as input the categorical value of the sensitive attribute and outputs a dense embedding of 128 dimensions. It is then followed by a linear projection and a reshaping to match the spatial dimensions of the output of the convolution block to which it is concatenated, but with a single channel. The same type of linear projection is applied on the 1024 dimensional noise vector input, but this projection and reshaping matches both the spatial and channel dimensions of the output of the convolutional block to which it is concatenated. Concatenation is in both cases done along the channel dimension.

Algorithm 1

input: $\mathcal{D}, lr, \lambda, \epsilon, \beta_1, \beta_2$
$\epsilon_1, \epsilon_2 \leftarrow \epsilon$
repeat
Draw m samples uniformly at random from the dataset
$(x_1, s_1), \ldots, (x_m, s_m) \sim \mathcal{D}$
Draw m samples from the noise distribution
$(z_1^{(1)}, z_1^{(2)}), \ldots, (z_m^{(1)}, z_m^{(2)}) \sim p(z^{(1)}, z^{(2)})$
Draw m samples from the synthetic distribution
$s'_1, \ldots, s'_m \sim p(s')$
Compute censored and synthetic data
$x'_1, \ldots, x'_m = f_{\theta_f}(x_1, z_1^{(1)}), \ldots, f_{\theta_f}(x_m, z_m^{(1)})$
$x''_1, \ldots, x''_m = g_{\theta_g}(x'_1, s'_1, z_1^{(2)}), \ldots, g_{\theta_g}(x'_m, s'_m, z_m^{(2)})$
Compute filter and generator losses

$$\Theta_f(\theta_f) = -\frac{1}{m}\sum_{i=1}^{m} \ell_f(h_f(x'_i; \phi_f), s_i) + \lambda \max(\frac{1}{m}\sum_{i=1}^{m} d(x'_i, x_i) - \epsilon, 0)^2$$

$$\Theta_g(\theta_g) = \frac{1}{m}\sum_{i=1}^{m} \ell_g(h_g(x''_i; \phi_g), s_i) + \lambda \max(\frac{1}{m}\sum_{i=1}^{m} d(x''_i, x_i) - \epsilon, 0)^2$$

Update filter and generator parameters
$\theta_f \leftarrow \theta_f - lr \nabla_{\theta_f} \Theta_f(\theta_f)$
$\theta_g \leftarrow \theta_g - lr \nabla_{\theta_g} \Theta_g(\theta_g)$
Compute discriminator losses
$\Phi_f(\phi_f) = \frac{1}{m}\sum_{i=1}^{m} \ell_{f_d}(h_f(x'_i; \phi_f), s_i)$

$$\Phi_g(\phi_g) = \frac{1}{m}\sum_{i=1}^{m} \ell_g(h_g(x''_i; \phi_g), fake) + \frac{1}{m}\sum_{i=1}^{m} \ell_g(h_g(x_i; \phi_g), s_i)$$

Update discriminator parameters
$\phi_f \leftarrow \text{Adam}(\phi_f; lr, \beta_1, \beta_2)$
$\phi_g \leftarrow \text{Adam}(\phi_g; lr, \beta_1, \beta_2)$
until stopping criterion
return $\theta_f, \theta_g, \phi_f, \phi_g$

The discriminators $h_f(X'; \phi_f)$ and $h_g(X''; \phi_g)$ are modeled using ResNet-18 [10] and a modified version which we refer to as ResNet-10[2], respectively. The last fully connected layer has been replaced with a two and three class output layer for each model, respectively.

The algorithm used to train the filter, the generator, and the discriminators is described in Algorithm 1 where the loss function ℓ_f for the filter is the entropy of its output, the loss functions ℓ_{f_d} and ℓ_g are categorical cross entropy, the distortion measure d is defined as the $L2$-norm, and $p(s')$ is assumed to be the uniform distribution $\mathcal{U}\{0, 1\}$. The hyperparameters consist of the learning rate lr, the quadratic penalty term coefficient λ, the distortion constraint ϵ, and the (β_1, β_2) parameters to Adam [17]. We also include results where ℓ_f is the categorical cross entropy as in the baseline.

[2]ResNet-10 has the same setup as ResNet-18, but each of the "conv2_x", "conv3_x", "conv4_x", and "conv5_x" layers consists of only one block instead of two.

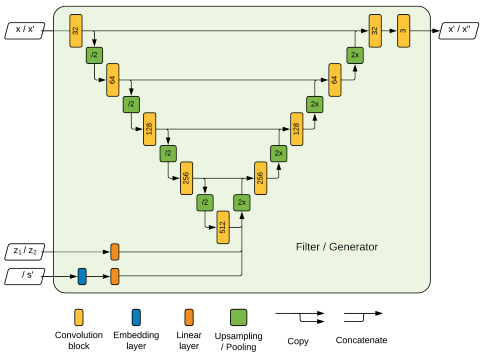

Figure 2: The architecture of the filter and generator networks. The notation x/x', z_1/z_2, and $/s'$ means that the network takes these inputs and gives as output x'/x'' for the filter / generator respectively. In the filter we do not use the embedding branch.

4 Experiments

In this section we describe our experiments, the dataset used to evaluate the method and the evaluation metrics. We have evaluated our method on the facial images from the CelebA dataset [19][3].

The preservation of utility for a downstream task is measured by training a classifier to detect non-sensitive attributes in the censored images, which are known from the original dataset.

4.1 CelebA

The CelebA dataset [19] consists of 202,599 face images of size 218x178 pixels and 40 binary attribute annotations per image, such as age (old or young), gender, if the image is blurry, if the person is bald, etc. We include experiments where we use the *smiling* attribute and the *gender* attribute from CelebA, respectively as the sensitive variable to censor and synthesize. The dataset has an official split into a training set of 162,770 images, a validation set of 19,867 images and a test set of 19,962 images. We resize all images to 64x64 pixels and normalize all pixel values to the region $[0, 1]$, unless otherwise is specified.

4.2 Filtering and replacement of sensitive data

Let $\mathcal{D}_{train} = \{(x_i, s_i\}_{i=1}^n$ be a set of training data where x_i denotes facial image i and $s_i \in \{0, 1\}$ denotes the sensitive attribute (either *smiling* or *gender*). Further let $\mathcal{D}_{test} = \{(x_i, s_i)\}_{i=1}^m$ be the held out test data. Since we evaluate on CelebA [19] we have $n = 162,770$ for the training data and $m = 19,867$ for the test data. We also, for the purpose of evaluation only, assume access to a number of utility attributes $u_i \in \{0, 1\}$ for each x_i. We consider the utility attributes *gender*, *wearing lipstick*, *young*, *high cheekbones*, *mouth slightly open*, *heavy makeup*, and exclude the one currently used as the sensitive attribute for each experiment. The utility attributes are used to evaluate how well non-sensitive attributes are preserved after the images have been censored.

Let θ_f and θ_g denote the parameters of the filter and generator, respectively, obtained by Algorithm 1 applied to $\mathcal{D}_{train}$ with $lr = 0.0005$, $\lambda = 10^5$, $\epsilon \in \{0.01, 0.05, 0.001, 0.005\}$, and $(\beta_1, \beta_2) = (0.9, 0.999)$. We can then define the training data censored by the filter as $\mathcal{D}'_{train} = \{(x'_i, s_i)\}_{i=1}^n$, where $x'_i = f(x_i, z_i^{(1)}; \theta_f)$ and the training data censored by the filter and the generator as $\mathcal{D}''_{train} = \{(x''_i, s_i)\}$ where $x''_i = g(x'_i, s'_i, z_i^{(2)}; \theta_g)$, $z_i^{(1)}, z_i^{(2)} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, and $s'_i \sim \mathcal{U}\{0, 1\}$. We do the same transformations to the test data and denote them $\mathcal{D}'_{test}$ and $\mathcal{D}''_{test}$ respectively.

Each experiment is run on a Tesla V100 SXM2 32 GB in a DGX-1 machine. The training is restricted to 100 epochs which takes about 13 hours.

[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

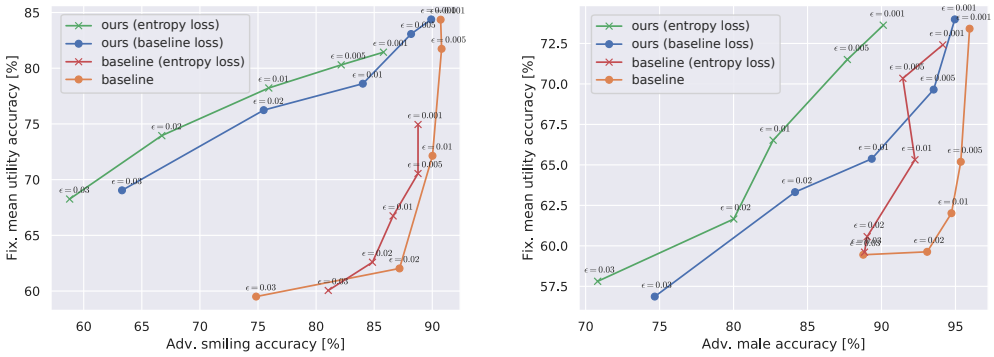

Figure 3: Privacy vs utility trade-off where the sensitive attribute is smiling (left), and gender (right). Privacy is measured using the adversarial classifiers $p_f(s|\cdot)$, $p_g(s|\cdot)$ and $p_{g\circ f}(s|\cdot)$, which had access to the non-censored labels for the censored images in the training set. Average utility is the accuracy of fixed classifiers of attributes other than the sensitive one (for *Gender*, *Wearing_Lipstick*, *Young*, *High_Cheekbones*, *Mouth_Slightly_Open*, *Heavy_Makeup*). Our approach with entropy loss consistently performs better than all other explored approaches.

To evaluate the two methods we use four different metrics. We train three adversarial classifiers $p_f(s|\cdot)$, $p_g(s|\cdot)$ and $p_{g\circ f}(s|\cdot)$ to predict the ground truth smile of a person given an image censored by the baseline, only the generator, and our method respectively. That is, we train the adversary $p_f(s|\cdot)$ on $\mathcal{D}'_{train}$ and evaluate the accuracy of the adversary on $\mathcal{D}'_{test}$ to measure how predictable the sensitive attribute is from an image censored by the baseline in this adversarial setting, and similarly we train the adversary $p_g(s|\cdot)$ and $p_{g\circ f}(s|\cdot)$ on their respective censored training sets and evaluate the accuracy on their respective testing sets.

In addition, we train a classifier $p_\psi(u|\cdot)$ to predict the gender of a person in an image on the original uncensored training data and use it as a measure to see how well the gender attribute is preserved after censoring the data. We also train a classifier $p_\eta(s|\cdot)$ to predict the smile of a person in the image on the original uncensored training data and use it as a measure to see how well the smile is censored.

To quantify the image quality of the censored images x' and x'' we use the Fréchet Inception Distance (FID) [11]. This is a metric frequently used in GAN literature to measure image quality.

5 Results

In this section we present quantitative and qualitative results on the facial image experiments.

5.1 Quantitative results

Figure 3 shows the trade-off between privacy and utility using the strong adversarially trained classifiers $p_f(s|\cdot)$, $p_g(s|\cdot)$ and $p_{g\circ f}(s|\cdot)$ when evaluated on images censored by the baseline and our method respectively. With these adversaries, which are much stronger than the fixed classifiers, our method consistently has a higher utility at any given level of privacy compared to the baseline. Remember: these adversaries require to run tagged training data through the privacy mechanism to be able to train.

To further show that the privatization mechanism is necessary we have conducted a very similar experiment using StarGAN [6] to randomly change the sensitive attribute in the image. We then train adversarial classifiers in the same way on images censored by StarGAN, which look very convincing to a human, and we then evaluate these adversaries on the held out test set censored by StarGAN. The adversaries can successfully detect the sensitive attributes when transformed using StarGAN with an accuracy of 90%. We explored $\lambda \in \{0, 5, 10, 50\}$.

In Table 1 we present the results of evaluating the accuracy of the fixed classifier $p_\eta(s|\cdot)$ on the dataset $\{x''_i, s'_i\}_{i=1}^m$ where $x''_i = g(x'_i, s'_i z_i^{(2)})$ is the image censored with our method and s'_i is the

Table 1: The mean accuracy and standard deviation over five different random seeds for $p_\phi(s|\cdot)$ evaluated on the data set $\{(x_i'', s_i')\}_{i=1}^m$ for varying ϵ. That is, the success rate of our method to fool the fixed classifier that the synthetic sensitive attribute s' is in the censored image x''. Higher is better.

Dist.		**Synthetic**		
ϵ	Smiling	Gender	Lipstick	Young
0.001	82.4 ± 2.1	72.9 ± 1.5	62.2 ± 3.1	59.2 ± 1.9
0.005	86.4 ± 3.7	79.4 ± 0.6	71.7 ± 2.4	62.1 ± 2.5
0.01	87.4 ± 2.1	85.6 ± 1.4	77.6 ± 2.4	59.6 ± 1.6
0.05	91.2 ± 2.9	90.3 ± 4.3	90.6 ± 4.2	67.7 ± 2.1

Table 2: The value of each cell denotes the Pearson's correlation coefficient between predictions from a fixed classifier trained to predict the row attribute and a fixed classifier trained to predict the column attribute, given that the column attribute has been censored.

	Smiling	Gender	Lipstick	Young
Smiling	1.00	-0.04	0.08	-0.06
Gender	-0.07	1.00	-0.44	-0.21
Lipstick	0.14	-0.30	1.00	0.26
Young	0.05	-0.11	0.23	1.00
High Cheekbones	0.14	-0.07	0.15	-0.01
Mouth Open	0.04	0.00	0.03	-0.02
Heavy Makeup	0.12	-0.24	0.47	0.22

new synthetic attribute uniformly sampled from $\{0, 1\}$. That is, we measure how often the classifier predict the new synthetic attribute s_i' when applied to x_i''. We can see that with $\epsilon = 0.001$ the method is on average able to fool the classifier 82.4% of the time, and this increases with larger ϵ to a success rate of 91.2% on average with $\epsilon = 0.05$. We also include these results when the images have been censored with respect to the attributes *gender*, *wearing lipstick* and *young*.

We also include results on correlations between classifier predictions on a pair of attributes when one attribute has been synthetically replaced, as seen in Table 2. This is further discussed in Section 6.

5.2 Qualitative results

In this section we present qualitative results of the baseline and our method. In Figure 4 we show, from the top row to the bottom row, the input image x, the censored image x', the censored image x'' with the synthetic attribute $s' = 0$ (non-smiling), and the censored image x'' with the synthetic attribute $s' = 1$ (smiling). A value of $\epsilon = 0.001$ is used in the first four columns, and $\epsilon = 0.01$ in the last four columns. The images censored by our method look sharper and it is less obvious that they are censored. We can see that the method convincingly generates non-smiling faces and smiling faces while most of the other parts of the image is intact. These images are sampled from models trained on images of 128x128 pixels resolution. Figure 6 shows corresponding samples on the same input images, but using *gender* as the sensitive attribute.

The gap in visual quality between our method and the baseline becomes clear when ϵ increase from 0.001 to 0.01. The images censored by the baseline look blurry while the images censored by our method still seem to maintain a lot of the structure from the original image. However, in some cases, for example in the eighth column, it is perhaps no longer obvious that the image is of the same person, but attributes such as gender, hair color and skin tone are still preserved.

In Figure 5 we present results where the attributes *smiling*, *gender*, *wearing lipstick*, and *young* have been censored. The method fails to disentangle the gender attribute and the lipstick attribute. This is discussed in more detail in Section 6.

Figure 4: Qualitative results for the sensitive attribute smile. In the first four columns: $\epsilon = 0.001$, and in the last four columns: $\epsilon = 0.01$. From top to bottom row: input image (x), censored image (x'), censored image with synthetic non-smile ($x'', s' = 0$), censored image with synthetic smile ($x'', s' = 1$). The model is able to generate a synthetic smiling attribute while maintaining much of the structure in the image. These images were generated from a model trained using 128x128 pixels.

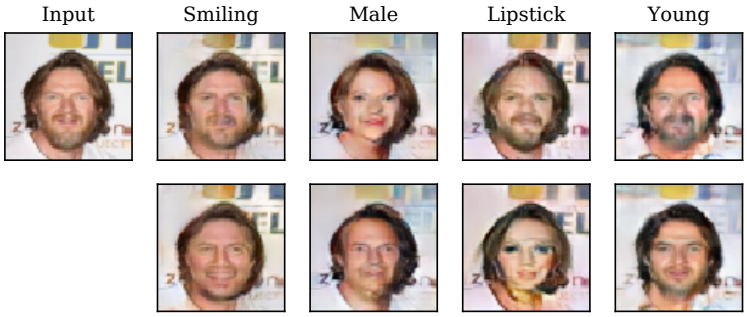

Figure 5: Examples of changing other attributes in an image. Each column (except the first which is the input) corresponds to an attribute that is being censored, and the top and bottom row show the censored image with the synthesized attribute set to false and true respectively.

6 Discussion

In Figure 3 our method consistently has a higher utility at any given level of privacy compared to the baseline. Furthermore, we can observe that using the entropy loss function for the filter benefits both the baseline and our method. Both for smiling and gender being the sensitive attribute, our method outperforms all other evaluated approaches.

In our experiments with StarGAN, we observe no noticable privacy when using this standard attribute manipulation method. This motivates our approach which is explicit about the privacy objective.

In Figure 3, we recognize that the baseline achieves a higher privacy for small ϵ, but point out that for a large value of $\epsilon = 0.05$ the baseline maintains a very low utility as the utility prediction is not better than random guessing.Further, for $\epsilon = 0.01$ our method achieves a higher utility but with a lower privacy.

Most importantly, our method demonstrates a higher privacy in the evaluation with the much stronger adversary as seen in Figure 3. This shows that our method makes it more difficult for the adversary to see through the privatization step. To show the effect of the filter we have included results using only

Table 3: Pearson correlation coefficient between the smiling attribute and 10 other attributes in the CelebA training dataset, ordered from high to low absolute correlation.

Attribute	Correlation
High cheekbones	0.68
Mouth slightly open	0.53
Rosy cheeks	0.22
Oval face	0.21
Wearing lipstick	0.18
Heavy makeup	0.18
Wearing earrings	0.17
Attractive	0.15
Gender	-0.14
Bags under eyes	0.11

the generator to privatize the images. Note that the combination of the filter and the generator works best.

A possible reason why using only the generator does not achieve high privacy is that the generator architecture is designed to easily learn the identity function to promote transformations that changes the input. Assuming enough capacity, the generator could learn the following rule: if the sensitive attribute in the image that is being censored is the same as the randomly sampled attribute, let the image through without changes. Otherwise, apply the ϵ-constrained change that transforms the image into a realistic image with the new attribute.

If the transformed image is indistinguishable from a real image this is not a problem, but if it is not we can easily reverse the privatization by detecting if the image is real or not. To mitigate this problem the filter *always* removes the sensitive data in the image which forces the generator to synthesize new data. Since the censored image is now guaranteed to be synthetic, we can no longer do the simple real/fake attack.

In Table 1, we see that the fixed smile classifier is fooled by our privatization mechanism in 82.4% to 91.2% of the data points in the test set (depending on the distortion ϵ). These results indicate that it may be harder for an adversarially trained classifier to predict the sensitive attribute when it has been replaced with something else, as compared to simply removed. We assume that this is due to the added variability in the data. Or intuitively: it is easier to "blend in" with other images that have similar demonstrations of smiles.

In Table 3 we see that some of the other attributes in the facial images of the CelebA dataset are highly correlated with whether or not the person is smiling. Two attributes that are highly correlated with smiling are high cheekbones and slightly open mouths. It is not obvious that a person with high cheekbones should be predisposed to smile more often, rather it may be that a person that is smiling is perceived as having high cheekbones due to the contraction of the facial muscles. We can see in Figure 4 that this is captured in all of the images with a synthetic smile (fourth row). The cheek muscles are visibly contracted and there are clear dimples in all images. On the other hand, the synthetic images with non-smile (third row) seem to have much less contracted facial muscles. We also note that the images with generated smiles tend to have an open mouth, while the images with a generated non-smile tend to have a closed mouth.

The fact that many important attributes in facial images correlate leads to the reflection that disentangling the underlying factors of variation is not entirely possible. For example, in this dataset lipstick is highly correlated with female gender. This means that if we want to hide all information about whether or not the person is wearing lipstick we also need to hide its gender (and other correlating attributes). This problem can be seen in Table 2 where changing whether or not a person is wearing lipstick correlates with changes of gender.

The question we ask is: if we censor an attribute in an image, how does that correlate with changes of other attributes in the image? In the lipstick column of Table 2 we have censored the attribute lipstick. We then make predictions on whether or not the person in the censored image is wearing lipstick, and compute the correlation between these predictions and predictions for the attributes for each

Figure 6: Qualitative results for the sensitive attribute gender. In the first four columns: $\epsilon = 0.001$, and in the last four columns: $\epsilon = 0.01$. From top to bottom row: input image (x), censored image (x'), censored image with synthetic female gender ($x'', s' = 0$), censored image with synthetic male gender ($x'', s' = 1$). The model is able to generate a synthetic gender while maintaining much of the structure in the image. These images were generated from a model trained using 128x128 pixels.

row. For example, we can see that changes in lipstick correlate negatively with changes in gender and positively with makeup. This highlights the problem of disentangling these underlying factors of variation. We also see this in Figure 5 where changing the lipstick attribute to true results in a transformation that changes the gender from male to female.

One core strength of our method is that it is *domain-preserving*, meaning that $x, x'' \in \mathcal{X}$ where $\mathcal{X}$ denotes the domain of images. This allows a utility provider to use the censored image x'' in existing algorithms without modifications. Together with our results on the utility preservation, we envision that we can stack several different such privatization mechanisms in a chain to add a selection of privatizations to an image. This could come in handy for example in settings in social media where people may want to share images, but would like a selection of attributes to be censored.

7 Conclusions

In this work we have addressed the problem of learning privacy-preserving adversarial representations for image data that can censor sensitive attributes by generating a new synthetic attribute in its place. While previous work has proposed this method for removing information from a representation, our approach extends on this and can also generate new information that looks realistic in its place. We evaluate our method using adversarially trained classifiers, and our results show that it is possible to preserve non-sensitive attributes of the image when performing the censoring. Further, the results show that the synthetically added attribute helps in fooling the adversary in the most challenging setting where the adversary is allowed to be trained using the output of the privacy mechanism.

References

[1] Rawan Alharbi, Mariam Tolba, Lucia C. Petito, Josiah Hester, and Nabil Alshurafa. To mask or not to mask? balancing privacy with visual confirmation utility in activity-oriented wearable cameras. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3), September 2019.

[2] Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially learned representations for information obfuscation and inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 614–623, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[3] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

[4] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4), 2012.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[9] 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.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.

[12] C. Huang, P. Kairouz, and L. Sankar. Generative adversarial privacy: A data-driven approach to information-theoretic privacy. In *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, pages 2162–2166, Oct 2018.

[13] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-aware generative adversarial privacy. *Entropy*, 19(12), 2017.

[14] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Daniela Ushizima, Sek Chai, Shinjiro Sueda, Xin Lin, Aidong Lu, Daniel Thalmann, Chaoli Wang, and Panpan Xu, editors, *Advances in Visual Computing*, pages 565–578, Cham, 2019. Springer International Publishing.

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, July 2017.

[16] Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 3020–3029. JMLR.org, 2016.

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[19] 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.

[20] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 19–35, Cham, 2016. Springer International Publishing.

[21] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

[22] 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.

[23] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] S. A. Osia, A. Taheri, A. S. Shamsabadi, K. Katevas, H. Haddadi, and H. R. Rabiee. Deep private-feature extraction. *IEEE Transactions on Knowledge and Data Engineering*, 32(1):54–66, Jan 2020.

[25] 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.

[26] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 639–655, Cham, 2018. Springer International Publishing.

[27] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

[29] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. *CoRR*, abs/1712.03474, 2017.

[30] Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2052–2060, New York, NY, USA, 2019. Association for Computing Machinery.

[31] Luan Quoc Tran, Xi Yin, and Xiaoming Liu. Representation learning by rotating your faces. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[32] Haotao Wang, Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Privacy-preserving deep visual recognition: An adversarial learning framework and A new dataset. *CoRR*, abs/1906.05675, 2019.

[33] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 627–645, Cham, 2018. Springer International Publishing.

[34] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, pages 585–596, 2017.

[35] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, Proceedings of the 30th International Conference on International Conference on Machine Learning, pages 325–333. JMLR.org, 2013.

[36] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM, 2018.