Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Differentially private facial obfuscation via generative adversarial networks[☆]

William L. Croft [a,*], Jörg-Rüdiger Sack [a], Wei Shi [b]

[a] School of Computer Science, Carleton University, Ottawa, Canada
[b] School of Information Technology, Carleton University, Ottawa, Canada

ARTICLE INFO

Article history:
Received 31 January 2021
Received in revised form 24 November 2021
Accepted 26 November 2021
Available online 13 December 2021

Keywords:
Privacy protection
Facial obfuscation
Differential privacy
Generative adversarial networks

ABSTRACT

From smartphones owned by the majority of teenage and adult populations to omnipresent closed-circuit television systems, the ubiquity of image-capturing devices in our everyday lives ensures that digital images of individuals are taken in the hundreds of millions on a daily basis. Many of these images capture individuals' faces which, through facial recognition techniques, identify the individuals and thus represent a major privacy concern. Many countries and companies require facial obfuscation to conform to privacy laws or policies. Since images should be useful and look realistic, a trade-off arises between privacy and utility. The task is therefore to find a method of obfuscation that offers a formal privacy guarantee while preserving visual quality and maintaining facial attributes deemed acceptable for release (e.g., the pose of the head, gender, etc.). We address this task by proposing facial identity obfuscation through the application of differential privacy to image encodings in a generative adversarial network. We provide details on the design of the model architecture and training process that allow for the generation of photo-realistic obfuscated images. Through the use of principal component analysis, we control the application of noise to the model encodings in order to achieve a favourable trade-off between privacy and utility. We demonstrate the effectiveness of our approach through an experimental comparison against other methods of obfuscation which also offer a formal guarantee of privacy.

1. Introduction

Digital images are captured from a wide variety of sources in massive quantities at an astounding rate. The number of images uploaded daily by users of social media is in the hundreds of millions [1]. Many of these images capture individuals' faces, thus providing a strong identifier of who the individual is [2,3]. It is often required, or at least respectful, to consider the privacy of the parties involved before sharing images online. Images of individuals are also routinely captured in large quantities from many other sources such as public surveillance [4], visual sensor networks [5] and image-based services such as Google Street View [6]. In all of these cases, the privacy of individuals should be carefully treated, whether it is a matter of law, company policies, or simply social responsibility [2,3].

The majority of teenagers and adults in countries with advanced economies own smartphones [7] and many of them actively share images on social media platforms [8]. Facebook now has over 2 billion users [9] with more than 350 million image uploads per day [1] and Instagram has over 1 billion users [10] with more than 100 million images and videos uploaded per day [11]. There are various social motivations, such as the desire to connect with others and share experiences, which encourage users to consistently share images on these platforms [12]. Yet due to the personal nature of many images, users must constantly weigh the benefits against the risks to privacy [13]. It is clear that users want to be able to control how, and with whom, their images are shared [14,15]. Furthermore, individuals appearing in images taken by others often want some say in what happens to the images. In cases involving multiple parties, conflicts are often resolved either through unaltered release of the image or complete redaction [16]. Yet this type of all or nothing approach is not ideal for handling nuanced cases and is likely unsatisfactory for some of the parties. It is also not uncommon for strangers to be captured in the background of images, making the issue of privacy more complex. Requesting consent from strangers to share images is likely to be impractical or even impossible in some cases.

[☆] Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [grant numbers RGPIN-2020-06482, RGPIN-2016-06253 and CGSD2-503941-2017].

[*] Corresponding author.

E-mail addresses: leecroft@cmail.carleton.ca (W.L. Croft), sack@scs.carleton.ca (J.-R. Sack), wei.shi@carleton.ca (W. Shi).

https://doi.org/10.1016/j.future.2021.11.032
0167-739X/© 2021 Elsevier B.V. All rights reserved.

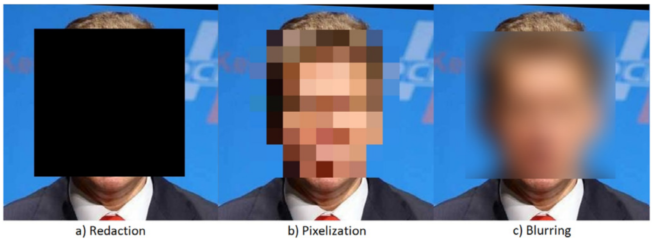


Fig. 1. Examples of basic methods of facial obfuscation. From left to right: redaction, pixelization, blurring.

Beyond privacy concerns regarding images being seen by others, rapid advances in machine learning approaches to computer vision also pose threats to privacy. Detection and identification of faces in images are now tasks that can be accomplished with great accuracy. Recent neural networks have achieved over 99% accuracy in tasks of facial verification and classification [17,18] and are increasingly able to handle difficult conditions such as low resolution, non-frontal poses and poor illumination [19,20]. These advances enable the development of software that can readily track and profile mass populations. This may, for example, raise concerns over the ability of governments and authorities to track populations on a highly personal level [21]. In some areas, closed-circuit cameras are present in the millions, capturing vast quantities of visual data [4]. While these technologies no doubt have benefits such as improved quality of services [2,22] and prevention of crime [2,23] or assistance in solving crimes, these benefits must be carefully weighed against the infringement on personal privacy. Furthermore, as with other types of sensitive information about individuals, the collection of visual data must be conducted in a manner that is compliant with privacy-centric laws such as GDPR [24,25].

One approach to ensure the privacy of individuals while still allowing for the release of images is to selectively redact portions of the image by placing a solid rectangle over top of individuals (Fig. 1-a). While this addresses the issue of privacy, it produces images that are found to be visually unappealing [26,27]. Less severe methods such as pixelization (Fig. 1-b) or blurring (Fig. 1-c) may be employed, but it is difficult to find a good trade-off between the aspects of privacy protection, visual appeal and information sufficiency [27,28]. In some cases, it may also be desirable to allow for the automatic extraction of information from images for machine learning tasks such as marketing and retail analytics [22,29]. In such cases, information sufficiency is especially important since machine learning algorithms typically rely on the ability to detect certain features in the images.

The balance between privacy protection and image quality and utility is further exacerbated by the fact that many ad hoc methods of privacy protection have been shown to ultimately be ineffective. The human visual system has remarkable robustness against blurring in images, allowing for recognition of faces against all but the most extreme applications of blurring [30,31]. Furthermore, super-resolution techniques can offer realistic reversals of pixelization [32] and machine learning algorithms have been shown to effectively defeat obfuscation via blurring as well as other forms effective against human recognition such as strong pixelization [33,34]. Although more recent approaches to privacy protection in facial images have improved greatly in this respect, many of them fail to offer a formal privacy guarantee (i.e., a mathematically provable guarantee such as k-same obfuscation [34] or differential privacy — described in Section 4.1) or suffer from susceptibilities to certain types of attacks [35]. In this work, we improve upon these issues by providing a method for the obfuscation of facial identity (i.e., the visual depiction of identity in facial images) which offers photo-realistic images of faces and is backed by the formal guarantee of differential privacy.

1.1. Contributions and paper outline

Our intent is to design a Generative Adversarial Network (GAN) that can obfuscate the facial identity in given input images while preserving certain types of information deemed acceptable for release such as the pose of the head (henceforth referred to simply as pose) and gender. To this end, we set out the following three goals:

1. The obfuscated images must be protected by a formal privacy guarantee.
2. The model must produce photo-realistic images and must have the ability to preserve targeted facial attributes (e.g., pose, gender, etc.).
3. The input images must be largely unconstrained with respect potential variability over aspects such as pose and image background.

To the best of our knowledge, differential privacy has not previously been applied via a GAN-based model for the obfuscation of facial identity in novel images fed to the trained model. A straight-forward application of differential privacy to GAN image encodings results in distorted, low-quality obfuscated images, as we demonstrate in Section 5. To achieve realistic obfuscated output that preserves desirable aspects in images, it is necessary to carefully consider the model design and training process as well as the method by which noise is applied to the encodings. To this end, we make the following contributions:

- We propose an approach that achieves the three goals we have specified for the task of facial obfuscation. For this, we provide architecture and training details that allow for differential privacy to be enforced on GAN image encodings in a manner resulting in realistic obfuscated images. We provide insight into how to carefully control the addition of noise to the image encodings such that visual quality is preserved.
- We propose a method to improve the preservation of utility in obfuscated images through an application of Principal Component Analysis (PCA) to distance-generalized differential privacy. To the best of our knowledge, ours is the first work to combine PCA with distance-generalized differential privacy. We explain how to apply noise to PCA-basis encodings to achieve the generalized privacy guarantee and how to adapt the GAN training process to produce high quality obfuscated images from noisy PCA-basis encodings.
- We provide details on both theoretical and practical interpretations of the formal privacy guarantee that we achieve for obfuscated images. We relate our instantiation of the privacy guarantee to that of standard differential privacy (for databases) and offer intuition for the interpretation of the privacy parameter ϵ as used in our setting.
- We provide an experimental evaluation of our work in which we empirically study the level of privacy achieved and compare our approach to others with respect to privacy-utility trade-offs.
- We demonstrate that our proposed approach achieves a strong level of privacy protection while preserving favourable levels of utility in the obfuscated images.

We emphasize that the work in this paper differs from the task of privacy-preserving machine learning, in which machine learning models are trained in a manner that offers privacy protection for sensitive information in the training data. In contrast to this, we use training data that is publicly available, resulting in a trained model that is devoid of sensitive information. All interactions with sensitive data occur after training, when the model is applied to accept private images as input and produce obfuscated versions as output.

The remainder of the paper is structured as follows. We review literature from the fields of study relevant to our work in Section 2 and provide a discussion on the works that achieve a formal privacy guarantee in Section 3. We then provide greater detail and notation for the works we build upon in Section 4. In Section 5, we cover the architecture and training details for our proposed GAN model for differentially private obfuscation of facial images. In Section 6, we show how to incorporate the use of PCA into the obfuscation pipeline in order to improve utility. In Section 7, we provide information on the interpretation of privacy in obfuscated images, both in the context of differential privacy and in more general terms. Finally, we provide an experimental evaluation of our work in Section 8 and conclude the paper in Section 9.

2. Literature review

Early approaches to facial obfuscation consisted of ad hoc methods such as pixelization [36] and blurring [37]. It has been shown that such ad hoc methods are often easily defeated by parrot attacks in which a machine learning algorithm is trained on instances of images that have been subjected to the targeted method of obfuscation [33,34]. As a result of this, the provision of a formalized guarantee of privacy has become an important aspect in methods of obfuscation.

To protect against parrot attacks, k-same obfuscation [34] offers a formal privacy guarantee which states that for each obfuscated image, at least k potential identities are equally probable to be the true identity. This is achieved by partitioning a gallery of images into clusters, each with cardinality at least k of visually similar images, and then averaging the images in each cluster to create obfuscated instances. This concept was first implemented using calculations in pixel-space [34] and was later applied to the parameter-space of Active Appearance Models (AAMs) [38,39] to improve the visual quality of the obfuscated images.

More recent advances in the field of facial obfuscation have largely turned to the usage of generative neural networks for the creation of obfuscated images. While k-same obfuscation and variants thereof have remained a popular choice for the privacy guarantee, other formalized approaches (e.g., redaction with in-painting [40,41] and differential privacy [35,42]) have recently been applied as well. We note that another direction of study, adversarial perturbation [43,44], has aimed to protect against machine learning recognition while still allowing humans to easily identify the images. However, in our work, we focus on obfuscation that protects against identification both by humans and machines.

2.1. Generative neural networks for images

As neural networks are becoming an increasingly popular tool for image processing, methods of facial obfuscation are now making use of them. These networks are able to offer impressive visual quality, often at a photo-realistic level, and provide convenient ways to manipulate parts of images while preserving desired aspects. We provide a brief review of relevant networks.

While not a particularly new area of study, convolutional neural networks have recently enjoyed a number of advances that have greatly improved their effectiveness in computer visions tasks such as classification of images [45]. These networks, primarily consisting of a layered combination of convolutional filters and non-linear activation functions, accept an image as input and produce a compact encoding that is well-suited for the intended task. Recent variations on this concept have essentially inverted the architecture by using transpose convolutions to take encodings as input and produce images as output. This has been applied to generate novel images having specific properties, potentially in combinations not seen in the training data [46]. This type of architecture has also been extended by pairing the generator network with a discriminator network that aims to distinguish between real or generated images [47]. The model, referred to as a Generative Adversarial Network (GAN), uses an adversarial training scheme in which the two networks compete in a minimax game, forcing the generator to learn how to produce realistic instances from a targeted distribution in order to fool the discriminator.

The concept of GANs has spawned a rich body of work on the generation and manipulation of highly realistic images. The DCGAN [48] architecture improved training stability, allowing for deeper network architectures and better image quality. DCGAN was also applied to demonstrate the ability to generate novel images by performing interpolation and arithmetic in the latent encoding space. Conditional GANs [49] extended the training process by teaching the network to generate images from a conditional distribution, allowing for the specification of desired properties in the generated images.

The concept of an encoder network was later introduced, allowing GANs to take an image as input and produce an encoding that could then be decoded (by the generator) back to an approximation of the input image [50,51]. This concept enabled the use of GANs for image editing tasks. Networks such as StarGAN [52] and AttGAN [53] have shown impressive results in the modification of specific features in facial images while preserving all other information in the images. It is these types of networks that are particularly useful for approaches to facial obfuscation.

2.2. Facial obfuscation using generative networks

The advances in generative neural networks, particularly those applied in the context of facial images, have provided a highly convenient tool for the task of facial obfuscation which has been explored in a number of recent works. By operating on the representation of an image within the encoding space of a neural network model, it is possible to generate realistic output of novel faces.

Using a model that learns facial features that are invariant to pose and illumination, k-same obfuscation has been applied to generate obfuscated faces from averaged clusters of facial features [54]. The authors propose an autoencoder-style architecture (i.e., a network that encodes and decodes images) such that the k-same algorithm can be applied to the intermediate image encodings. The concept of k-same obfuscation using a neural network has also been proposed via a transpose convolutional architecture [55]. Under this architecture, the network directly learns encodings for a set of training images. The authors therefore propose to use a mapping process from clusters of input images to clusters of training set images such that the encodings of the training image clusters can be averaged in a k-same fashion and used as the obfuscated replacements for the input images. The network architecture provides high visual quality and maintains separate features for facial expressions in order to allow for control and preservation of expressions in the obfuscated images. While k-same implementations have improved greatly over the years in visual quality, all such methods remain vulnerable to composition attacks and background knowledge, as this is a deficiency of the underlying privacy guarantee [35].

To provide a stronger level of privacy than what k-same offers, generative networks have been applied to achieve other types of formal privacy guarantees. The syntactic privacy guarantees of l-diversity and t-closeness have been applied over a set of binary facial attributes controlled via a GAN [56]. These privacy guarantees strengthen the k-same guarantee by additionally placing requirements over the distribution of attributes in order to render inferences about the original data more difficult. However, the proposed algorithm modifies only a predefined set of attributes while leaving all other information encoded from the image untouched, potentially leaking identifiable information.

Alternatively, privacy can be guaranteed by completely removing faces from images and then applying a generative network to in-paint the redacted area such that it produces a novel face. One such approach employs a pair of autoencoders, one to generate facial landmarks for a redacted facial image and the other to transform the generated landmarks and redacted image into an in-painted image displaying a new face matching the facial landmarks [57]. Although the redaction of a bounding box around the face guarantees perfect protection of the targeted area, the bounding box excludes most of the ears and hair, and in some cases, parts of the chin and forehead. The in-painting relies on this non-redacted contextual information to produce visual output of high quality. However, these excluded portions of the head may leak identifiable information. This information leakage can be empirically observed through the reported experiments in which facial re-identification was achieved at 5.6% accuracy over a set of 257 identities (much higher than the expected accuracy of roughly 0.4% for random guessing if all relevant information had been replaced). The concept of redaction and in-painting has also been applied using a conditional GAN [40]. Once again, facial landmarks are generated and these are provided along with the redacted image as the conditional information inputs to a GAN which generates an in-painted facial image. This work also uses a tight redaction bounding box that excludes much of the facial contour. Although re-identification accuracy is not reported in the original paper, we demonstrate the leakage of identifiable information in our own experiments in the present paper.

A variation on the in-painting approaches substitutes the use of facial landmarks for a parametric representation of a 3D facial model [41]. In this approach, a tightly cropped area around the face is first redacted, then 3D facial model parameters are generated from the redacted image using an autoencoder. Following this, a larger bounding box covering the full head is redacted and a second autoencoder is used to perform in-painting using the new redacted image and 3D model parameters as input. The redaction of the larger area addresses concerns of leakage of identifiable information through non-redacted facial contours. Further to this, many of the 3D model parameters passed to the second autoencoder are replaced with new values to prevent leakage of identifiable information though this channel. However, a subset of these parameters are allowed to keep their original values in order to ensure good visual quality during in-painting. These preserved parameters values have the potential to leak identifiable information, as seen by the reported facial re-identification accuracy of 7.1% over a set of 257 identities.

A more extreme method of redaction involves complete replacement of the original image with a newly generated facial image from a GAN [58]. In this approach, the authors first generate a facial image at random using the DCGAN architecture and then apply a feature editing GAN to modify basic facial features such as gender and hair colour to match the original image. This ensures that identifiable information is only revealed through the selected attributes. However, the approach suffers in its ability to preserve useful information in the obfuscated images. Aspects such as image background and pose are lost, and without a controllable privacy parameter, there is no means to adjust the degree of obfuscation.

The ability of neural networks to realistically manipulate facial features has also been exploited in contexts outside of formal privacy guarantees. Examples include the use of a training objective function that promotes changes in facial identity [59] and the application of a transformation to image encodings that induces large changes to identity-specific features while preserving certain desirable attributes [60]. Another informal approach to privacy protection in facial images is the concept of face swapping [61] which has recently been applied using neural networks to perform the transfer and blending process [62–64]. While these approaches may empirically demonstrate low re-identification risk in specific attack settings, their lack of a formal guarantee of privacy leaves uncertainty in their resilience and reliability against other types of attacks.

2.3. Differential privacy and facial obfuscation

Originally developed for usage on statistical databases, differential privacy [65] has become a widely accepted framework to achieve provably robust privacy guarantees on sensitive information about individuals. By adding controlled noise to responses for queries posed on sensitive data, a privacy guarantee is enforced which limits the distinguishability between potential query responses that could have led to the noisy response. The differential privacy guarantee has since been generalized to allow for its application to other domains beyond statistical databases [66].

Recently, differential privacy has been applied in the context of obfuscation of images. Through the addition of sufficient noise to pixel intensities, uncertainty is induced regarding the original content of the image [67]. The author proposes a method to determine the degree of noise required based on the maximum possible difference in the sum of pixel intensities over a window of pixels considered to cover a sensitive piece of information such as a face. Through a coarsening of image detail via pixelization, it is shown that the required noise can be reduced while achieving the same level of privacy. However, direct modification of pixels in this way leads to severe visual distortions in the images, rendering them highly unnatural. The visual quality has been improved slightly through the use of an invertible transformation applied to images via Singular Value Decomposition (SVD) [42]. Using this method, an image is decomposed into a product of three matrices, one of which contains the singular values that are said to encode the magnitude of geometric features in the decomposed image. Noise is then added to the singular values and the decomposition is inverted to produce an obfuscated image as output. This process, however, leaves the other two matrices unaltered, potentially leaking identifiable information. Even under a parameterization for a very strong level of privacy, experimental results show that re-identification accuracy remains at 17.5% over a set of 40 identities. Furthermore, although visual quality is improved over direct modification of pixels, the obfuscated images still do not resemble realistic human faces.

Photo-realistic visual quality has recently been achieved through the application of differential privacy to the image encodings learned by a transpose convolutional neural network [35]. The network, trained to produce realistic faces from noisy encodings, generates obfuscated faces through the alteration of high-level facial features. A separation of specific facial attributes from the image encodings also allows for the preservation of gender and facial expression, if desired. While the original approach is limited to obfuscation of identities contained in the network training data, a later extension allows for obfuscation of arbitrary identities [68]. This is achieved through the use of linear programming to approximate the encodings of novel identities. Noise is then applied to the approximated encodings and they are passed through the neural network to generate obfuscated facial images. However, the approach is still limited in its practical applicability due to a requirement for images taken in a heavily controlled setting (e.g., only frontal facing images against a solid white background).

In independent and concurrent work to our own, the concept of applying differential privacy to facial image encodings has also been treated in the context of using a GAN as the generative model.[1] One approach involves a pipeline of an image encoder network, followed by the application of differential privacy to the encoding, followed by the use of a GAN to generate the obfuscated output image from the noisy encoding [69]. However, the authors provide only an empirical estimate of the sensitivity that is required to configure a differentially private mechanism. This implies that privacy is not actually guaranteed. Furthermore, the values of the privacy parameter ϵ used in the reported experiments are in the range of tens of thousands, making them far too high to offer an acceptable level of privacy. A similar approach using an encoder and a GAN has also been taken with an added step in the pipeline to optimize the image encoding prior to adding noise in order to better capture the features in the original image [70]. In this work, the authors additionally enforce a bound on the maximum difference between encoding features in order to calculate an exact sensitivity and achieve the differential privacy guarantee in obfuscated output. However, the values of the privacy parameter ϵ used in the reported experiments also range from thousands to tens of thousands, again making them too high for acceptable levels of privacy.

Our work advances the concept of differentially private facial obfuscation via generative models by proposing an effective approach to design and train an encoder-based GAN architecture, allowing for high quality facial obfuscation for largely unconstrained image settings. Through the integration of PCA into the network training and image obfuscation pipelines, we are able to achieve differential privacy with vastly lower (i.e., stronger) privacy parameters than other approaches.

We note that differential privacy has also been applied in the context of training GANs in a privacy-preserving fashion. However, this goal is fundamentally different from that of facial obfuscation. Privacy-preserving training of GANs is used to protect sensitive training data, whereas facial obfuscation via GANs is intended to protect new images presented to the GAN after training (typically on public data) is complete. One method to protect sensitive training data involves adding noise to the gradients calculated for the discriminator network during training [71,72]. This allows for control, in the form of a differential privacy guarantee, over the influence of the sensitive information on the weights and distribution learned by the generator network. The generator can then produce new samples from the learned distribution while protecting the privacy of the training samples. A different approach replaces the discriminator network with an ensemble of teacher discriminators and a student discriminator [73]. The outputs of the teachers are aggregated and made differentially private before being passed as input to the student. Since training data is only ever seen by the discriminator, use of the student network as the new discriminator provides an alternative network that has a differentiable output and operates only on privacy-protected data. While both of these approaches (and others related to differentially private training) protect the training data, they do not offer a method to obfuscate new data presented to the network after training, and thus are not applicable for the task of facial image obfuscation.

3. Deficiencies in existing methods of facial obfuscation

We begin by providing a description of deficiencies in existing methods of facial obfuscation in order to clearly motivate the need for a stronger approach to privacy-protection that is simultaneously able to preserve realism and utility in obfuscated images. For ease of reference, we provide in Table 1 a summary of all reviewed works that offer a formal privacy guarantee for the task of facial image obfuscation.

While ad hoc approaches to facial obfuscation such as blurring and pixelization may be shown to be empirically effective in some cases, they do not formalize any attack model and offer no guarantee regarding their effectiveness in general. This lack of formalization leaves the potential for susceptibility to attacks that were not tested empirically. This deficiency is demonstrated clearly by the parrot attack [34], which handily defeats such ad-hoc methods by training machine learning models on obfuscated training data. Methods of face swapping (e.g., [62,63]) similarly provide no guarantee regarding potential attacks that might be launched against them. Furthermore, face swapping raises privacy concerns about the usage of real faces or components of real faces that are swapped into the released images.

Although the k-same family of approaches (e.g., [34] and subsequent developments) offers a formal privacy guarantee that limits the probability of re-identification to an upper bound of $\frac{1}{k}$, the framework remains susceptible to certain types of attacks. In particular, k-same obfuscation is vulnerable to composition attacks (i.e., attacks that combine information from multiple releases) and to attackers who already have partial knowledge about the context of the released images (e.g., individuals who could not be present in the image) [35].

Methods of redaction followed by in-painting (e.g., [40,41]) offer a guarantee of perfect privacy protection within a redacted area by destroying all information and allowing a machine learning approach to rebuild an appropriate image from contextual information. Since the process has no access to the original information from the redacted area, there can be no leakage of this information. Yet the contextual information around the redacted area (often the contour of the head including forehead, hair and ears) is likely to still reveal too much sensitive information. Increasing the area of redaction places a greater strain on the ability to in-paint the area in a manner that is both visually clean and able to preserve utility.

Existing approaches to differentially private obfuscation of images are able to provide a strong guarantee of formalized privacy but they fall short in utility. The application of differential privacy to pixel intensities [67] or to SVD matrices of images [42] leads to highly distorted images that no longer resemble human faces. Although differential privacy has also been applied via a generative neural network to achieve photo-realistic quality [35], the ability to handle variability in the facial images remains limited. This method has only been applied to images taken in a highly controlled setting and does not handle the great variation (e.g., image background, camera angle, pose, etc.) that is typically present in images. Furthermore, the lack of an encoder network in the model architecture greatly restricts the ability of the model to obfuscate novel instances of images, limiting its practical applicability.

To address the shortcomings outlined here, we employ the strong privacy guarantee of differential privacy and extend its application in generative models to a GAN. By carefully designing the model architecture and training process, we are able to achieve high visual quality and utility for unconstrained facial images. Our method obfuscates the full head and is able to preserve pose as well as selected facial attributes. Additionally, through the use of an architecture that includes an encoder network, our proposed model can obfuscate novel images of identities external to the training data.

[1] The two referenced works, [69] and [70], were posted to arxiv during the submission review process of our own work.

Table 1
Summary of the reviewed works that provide a formal guarantee of privacy.

Privacy guarantee	Paper	Generative model	Weaknesses
k-Same	Newton et al. 2005 [34]	None	Vulnerable to composition attacks and background knowledge.
	Gross et al. 2006 [39]	AAM [38]	
	Chi and Hu 2015 [54]	Custom autoencoder	
	Meden et al. 2017 [55]	DeconvFaces [46,74]	
l-Diversity/*t*-Closeness	Li and Lin 2019 [56]	StarGAN [52]	Only predefined features are protected, leaking identifiable information through the remainder of the encoding.
Redaction	Sun et al. 2018 [57]	Custom autoencoder	Ears and hair are not redacted, leaking identifiable information.
	Sun et al. 2018 [41]	Custom autoencoder	Parts of the image encoding are unaltered, leaking identifiable information.
	Hukkelås et al. 2019 [40]	Custom GAN	Ears and hair are not redacted, leaking identifiable information.
	Chen et al. 2021 [58]	DCGAN [48]	Poor preservation of utility.
Differential privacy	Fan 2018 [67]	None	Poor image quality.
	Fan 2019 [42]	None	Poor image quality. Parts of the SVD representation are unaltered, leaking identifiable information.
	Croft et al. 2019 [35], 2021 [68]	DeconvFaces [46,74]	Unable to handle varied image content.
	Liu et al. 2021 [69]	StyleGAN [75]	Privacy is not guaranteed due to an unbounded sensitivity. Uses very high ϵ values (18,000–72,000).
	Li and Clifton 2021 [70]	StyleGAN [75]	Uses very high ϵ values (9000–160,000).

4. Preliminaries

In this section, we provide details on background material of high relevance to our work. We begin with a review of the existing work on differential privacy applied to generative models. Following this, we provide a description of AttGAN, a network used for the modification of facial attributes in images. Our work applies the framework of differential privacy in generative models to a novel extension of AttGAN, adapting both the framework and the network to handle obfuscation of unconstrained facial images. Lastly, we review some details on PCA which we later draw on to better control the application of noise to the image encodings.

4.1. Differential privacy

Differential privacy [65] offers a formal and robust guarantee of privacy for the release of information about sensitive databases. This is achieved by allowing only for the release of responses to queries on the database that have been passed through a randomization mechanism. In the standard framework for differential privacy, the magnitude of noise is controlled by a privacy parameter ϵ and the query sensitivity. Query sensitivity is defined as the greatest possible difference between two noiseless query responses from any pair of adjacent databases (i.e., databases that differ by a single record).

Let the randomization mechanism be defined as a noise-adding function $K : \mathbb{D} \to \mathbb{R}^n$, where $\mathbb{D}$ is the set of valid database configurations and $n \in \mathbb{Z}^+$. Formally, the differential privacy guarantee states that such a mechanism is differentially private if for every pair of adjacent databases $D_1, D_2 \in \mathbb{D}$, the following holds:

$$\Pr\left(K\left(D_1\right) = R\right) \leq e^{\epsilon} \Pr\left(K\left(D_2\right) = R\right) \quad \forall R \in \mathbb{R}^n. \tag{1}$$

The repeated application of differentially private mechanisms leads to a composition of the privacy parameters used in each application [76]. Specifically, the use of a privacy parameter ϵ_1 followed by the use of a privacy parameter ϵ_2 results in a differential privacy guarantee that holds for $\epsilon = \epsilon_1 + \epsilon_2$. When multiple applications of mechanisms are expected, the largest acceptable sum of the privacy parameters is typically referred to as the privacy budget.

4.1.1. Differential privacy in generative models

Often, sensitive information exists in forms other than as records of a database. However, since the differential privacy guarantee is defined in terms of adjacency, a concept specific to the domain of databases, it cannot be directly applied outside of this context. A generalization of differential privacy [66] provides a privacy guarantee for secrets (structured data about individuals) by extending the concept of adjacency in databases to distances between secrets. The key intuition of the generalization is that an appropriate distance metric acts as a measure of distinguishability between secrets and thus takes the place of query sensitivity in the configuration of noise-adding mechanisms.

This generalized form of differential privacy has been applied to the obfuscation of facial images by considering the image encoding of generative models to be the secret [35]. By treating the image encoding as a vector $X \in \mathbb{R}^n$, the generalized privacy guarantee for images can be written using a distance function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and a randomization mechanism $K : \mathbb{R}^n \to \mathbb{R}^n$:

$$\Pr\left(K\left(X_1\right) = R\right) \leq e^{\epsilon d(X_1, X_2)} \Pr\left(K\left(X_2\right) = R\right) \qquad \forall X_1, X_2, R \in \mathbb{R}^n. \tag{2}$$

Since the encoding can be seen as a numeric representation of an individual depicted in an image, it is possible to measure distance between individuals and to alter the depicted identity through the injection of noise to the encoding. The distance measure used in [35] was an L_1 measure with the distance between each pair of elements scaled to the range [0, 1] and the overall vector distance similarly scaled. Letting $R_i = [min_i, max_i]$ be the range of elements in the ith position of the vector, the distance measure is defined as follows:

$$d\left(X_1, X_2\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{|X_{1i} - X_{2i}|}{max_i - min_i}. \tag{3}$$

Using this distance measure, the differential privacy guarantee can be achieved by independently adding Laplace noise to each element of the vector such that the element at the ith position uses a distribution with a scaling parameter of $\sigma_i = \frac{n(max_i - min_i)}{\epsilon}$ [35].

We make use of this privacy mechanism to protect the facial identities captured in GAN image encodings. We provide a discussion on the practical implications of this generalized privacy guarantee in the context of our proposed model in Section 7.1.

4.2. AttGAN

Manipulation of facial images is a common task in the machine learning community and has been treated quite successfully in recent works using GANs. In particular, AttGAN [53] has shown excellent results in the modification of high-level facial attributes (e.g., gender, hair colour, etc.) while preserving facial identity.

AttGAN uses an encoder/decoder pair which is trained alongside a discriminator in an adversarial manner. The encoder takes an image as input and passes it through a series of convolutional layers to generate an encoding which represents high-level features detected in the image. The decoder accepts an encoding along with a vector of supplementary facial attributes (e.g., gender) and passes the concatenated input through a series of transpose convolutional layers in order to produce a new image as output. The generated image is intended to resemble the original as closely as possible while differing only in the changes to facial attributes specified by the supplementary vector fed to the decoder. The discriminator is also a convolutional network which accepts an image as input. It serves two purposes. After matching the architecture of the encoder, it splits into two branches of fully connected layers. The first branch acts as the discriminator while the second branch acts as a classification network for the facial attributes present in the image.

We use the AttGAN architecture and training process as a starting point in our work and adapt them to suit our requirements for facial obfuscation. Full details on the adapted architecture and loss functions are given throughout Section 5.

4.3. Principal component analysis

Principal Component Analysis (PCA) is an approach used in machine learning to calculate an alternate basis within which to represent points of data (i.e., multivariate observations) such that the dimensions of the new basis are orthogonal and are arranged in descending order of explained variance over a training dataset. This technique is widely used as a method to greatly reduce dimensionality of data while preserving the most important information by dropping trailing dimensions in the new basis. This is typically done to assist in machine learning and data visualization tasks. We provide here a brief overview of how to extract and apply a PCA transformation matrix. For greater detail, we refer the reader to [77].

PCA involves computing a decomposition of the covariance matrix of a tabular dataset, either via eigendecomposition or SVD. The decomposition is used to calculate a transformation matrix which maps the original dataset into a new basis defined by the eigenvectors of the covariance matrix. The dimensions of the new basis are referred to as principal components. Each eigenvector has an eigenvalue associated with it which reflects the ratio of the variance from the original dataset that is expressed in the corresponding principal component. Since the ratios of explained variance typically decrease rapidly over the principal components, many of the later principal components (based on their order) can be dropped to reduce dimensionality while incurring very little information loss.

Let D be a matrix representing the original data in row format and let Z be a matrix made from the eigenvectors of the covariance matrix arranged in descending order of their eigenvalues. The transformed data is given by $D' = Z^T \times D$. Data that has been transformed to the principal component basis can be transformed back to its original basis by applying an inverse transformation. This is given as $D = Z^T \times D'$. This is possible to do even when dimensionality reduction has been applied. If all dimensions have been preserved, the inverse transformation is lossless. If trailing principal components have been dropped, information loss is minimized in the sense that the omitted principal components were those that expressed the least amount of variance in the original data.

4.3.1. PCA and differential privacy

When applying differential privacy to multi-variate data that has strong correlations between the variables, the independent application of noise to each variable can destroy a great deal of useful information [78]. As the magnitude of noise is increased, the correlations become less pronounced in the noisy data, leading to perturbed data points that are increasingly likely to lie outside of the original distribution.

The application of noise within a transformed PCA basis of reduced dimensionality can help to better preserve correlations in the data while simultaneously spending the privacy budget more efficiently [78,79]. The use of principal components that maximize explained variance within the training data ensures that the addition of noise along the axes of the transformed basis will produce noisy data that more closely resembles the original distribution. Furthermore, trailing axes with low explained variance ratios can be dropped without great loss of information, allowing the privacy budget to be focused on the axes with greater information content. We refer the reader to [78] for a visualization of these benefits realized on a two-dimensional example.

Most research (e.g., [79,80]) on this topic has focused on how to perform the PCA learning process in a differentially private manner since the training data is private. Once the data has been transformed to the PCA basis, noise can be added and the obfuscated results can then be transformed back to the original basis and released. Although less prominently addressed in the literature, the step of adding noise in the PCA basis requires care to ensure that the magnitude of the noise is controlled based on the sensitivity of the data in its new representation. In the context of databases, this sensitivity can be calculated using information about the PCA transformation and the queries posed on the data [78]. To the best of our knowledge, PCA has not been previously used in the context of distance-generalized differential privacy.

5. Differentially private obfuscation via a GAN

While GANs are an excellent tool for the representation and manipulation of facial images, the model must be carefully designed and trained in order to effectively execute its intended task. A pipeline to obfuscate images using an encoder-based GAN can be seen as a three-step process: (1) encode the facial image, (2) apply noise to the encoding, (3) decode the noisy encoding. With an appropriate application of noise, the decoding step will result in a novel image that differs sufficiently from the original to achieve effective obfuscation. However, a straight-forward application of differential privacy to image encodings leads to unrealistic distortions to images, resulting in poor-quality output as shown in Fig. 2-a.

Our work provides a method to address this concern and produce realistic obfuscated images via an encoder-based GAN architecture. Throughout this section, we cover important considerations for the structure of the image encoding and model architecture (covered in Section 5.1) as well as the model training process (covered in Sections 5.2–5.4). These design details do not change the structure of the obfuscation pipeline but rather the way in which the model will handle the steps of encoding and decoding. This is done to enable the model to treat the addition of noise as realistic modifications to facial identity as opposed to arbitrary visual distortions. We describe these details in the context of an extension and adaptation of AttGAN, however, the concepts can equally be applied to other models. We additionally propose to extend the pipeline with pre- and post-processing steps to first mask out and then later rejoin the image background (covered in Section 5.5). This is done to allow the GAN, and thus

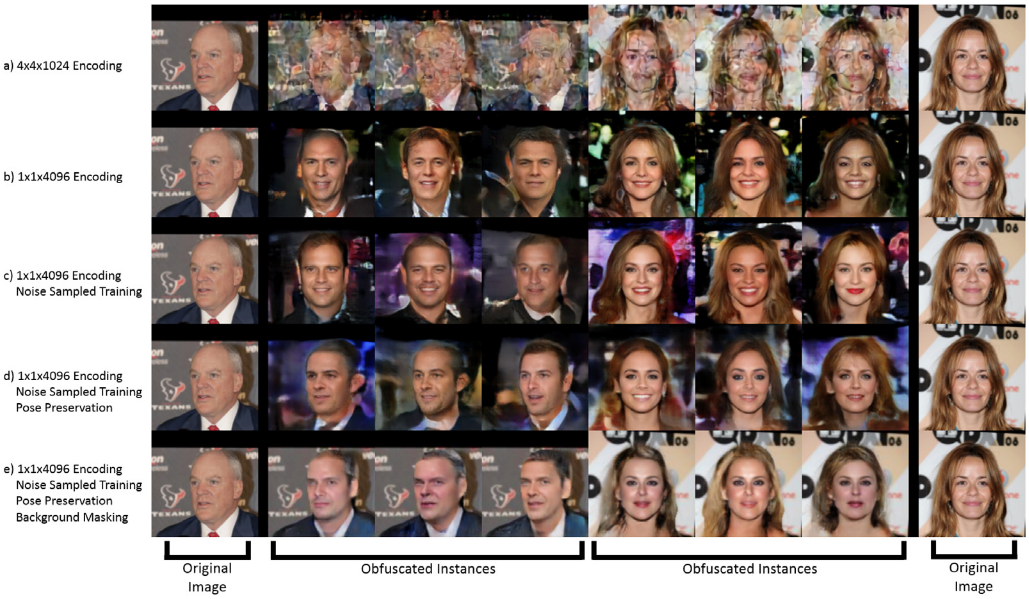


Fig. 2. An example of obfuscated instances of an image with each row showing results from a different architecture and training configuration. Row (a) uses the original 4 × 4 × 1024 encoding with no changes to the model. Each subsequent row incorporates one cumulative modification to the model. Row (b) substitutes a 1 × 1 × 4096 encoding (as described in Section 5.1) with no changes to the model training. Row (c) adds noise to encoding samples during training (as described in Section 5.2). Row (d) employs additional inputs to preserve pose (as described in Section 5.3). Row (e) applies background masking (as described in Section 5.5).

the obfuscation of its encoding, to focus on only the portion of the image that contains the facial identity (i.e., the head depicted in the image).

Samples of obfuscated images are shown in Fig. 2 to demonstrate the impact each of our major design considerations have on the obfuscated output. The complete pipeline is as follows:

1. Background masking (pre-processing)
2. Image encoding (GAN encoder)
3. Encoding noise addition (differential privacy mechanism)
4. Image decoding (GAN decoder)
5. Background merging (post-processing)

We define the following notation (extending that of the original AttGAN) to be used throughout this section. Let $x^{a,p}$ be an image depicting a face that has facial attributes matching the specification of a vector $a = [a_1, \dots, a_m]$ of binary attributes and a pose matching the specification of a vector $p = [p_1, \dots, p_k]$ of real-valued parameters. Let G_{enc} and G_{dec} be the encoder and decoder networks, respectively. Given an image $x^{a,p}$ as input, $z = G_{enc}(x^{a,p})$ denotes the encoding of $x^{a,p}$ and $x^{\hat{b},\hat{q}} = G_{dec}(z, b, q)$ denotes the image produced by the decoder when given an encoding z, a new vector $b = [b_1, \dots, b_m]$ of binary attributes and a new vector $q = [q_1, \dots, q_k]$ of real-valued pose parameters. Furthermore, a noise-adding function, N, may be applied to the encoding to produce an obfuscated image, denoted as $\tilde{x}^{\hat{b},\hat{q}} = G_{dec}(N(z), b, q)$.

5.1. Image encoding and model architecture

We begin by considering the design of the image encoding used by the model. Following the process of [35], we will be adding noise to the encodings of images in order to induce changes to high-level facial features as a means of obfuscation. The encoding must therefore be appropriately receptive to such noise. Regardless of the shape of the encoding, it is ultimately a multi-dimensional representation of numeric features to which noise can be added. However, the shape plays an important role in determining how visual quality is impacted by the addition of noise. Consider the 4 × 4 × 1024 encoding used in AttGAN. This can be interpreted as 1024 many 4 × 4 grids where each grid is a spatially compressed representation of the presence of some type of feature in the image. The 4 × 4 spatial dimensions present two problems. The first is that each cell of the grid relates to a local part of the image, meaning noise applied to a grid cell will have a localized effect rather than modifying the image at a global level. The second issue is that when convolutional filters are applied to overlapping areas of the output from the previous network layer, the resulting features (grid cells) describe overlapping areas and are thus highly correlated. Even when applications of a filter do not overlap (i.e., the convolutional stride matches the filter size), correlation between features remains likely due to the close spatial proximity of the filter applications. When noise is added to the encoding, the independent modification of features destroys their correlations with each other. Since this is done to features representing local areas in the image, the result is a highly distorted image which no longer resembles a human face. An example of this distortion can be seen in Fig. 2-a.

To avoid distortions induced by inconsistencies between previously correlated features representing local areas in the image, we propose that noise should instead be added to global facial features. Further compression of the encoding down to 1 × 1 spatial dimensions guarantees that each numeric value in the resultant encoding is a representation of a feature at the global level. This eliminates the potential for inconsistencies at the local level since the correlations are now compressed into a single value. Our proposed encoder/decoder architecture which implements this approach is given in Table 2. The improvement in visual quality from this change to the architecture can be seen in the difference between rows a and b of Fig. 2.

We note also that while the use of shortcut connections [81] is a common strategy in encoder/decoder pairs to help retain finer details in the image by bypassing the encoding bottleneck, we do not employ this strategy. Since our intent is to obfuscate images through the addition of noise to the encoding, any information that bypasses this bottleneck leads to leakage of sensitive information regarding the identity depicted in the image. This would render the differential privacy guarantee

Table 2
Encoder and decoder architectures. BN refers to batch normalization while LReLU refers to the leaky rectified linear units activation function.

Encoder	Decoder
Input: 128 × 128 × 3 image	Input 1 × 1 × 4096 encoding
Conv (4 × 4 × 64, Stride = 2), BN, LReLU	Concat binary attributes and pose parameters
Conv (4 × 4 × 128, Stride = 2), BN, LReLU	Trans Conv (2 × 2 × 2048, Stride = 1), BN, LReLU
Conv (4 × 4 × 256, Stride = 2), BN, LReLU	Trans Conv (4 × 4 × 1024, Stride = 2), BN, LReLU
Conv (4 × 4 × 512, Stride = 2), BN, LReLU	Trans Conv (4 × 4 × 1024, Stride = 2), BN, LReLU
Conv (4 × 4 × 1024, Stride = 2), BN, LReLU	Trans Conv (4 × 4 × 512, Stride = 2), BN, LReLU
Conv (4 × 4 × 2048, Stride = 2), BN, LReLU	Trans Conv (4 × 4 × 256, Stride = 2), BN, LReLU
Conv (2 × 2 × 4096, Stride = 1), BN, LReLU	Trans Conv (4 × 4 × 128, Stride = 2), BN, LReLU
	Trans Conv (4 × 4 × 3, Stride = 2), Tanh

meaningless. Although one could potentially apply noise to the information passed through the shortcut connections, this would present a number of additional challenges including the degradation of newly introduced correlations, obfuscation applied at a local level, and the requirement for a more complex privacy budget management scheme. We expect that this would not be a worthwhile venture but we leave the possibility open as a potential extension of our work.

5.2. Loss calculations for noisy encodings

In addition to using an encoding amenable to the application of noise, the decoder network must also be trained on how to handle noisy encodings. Networks such as DCGAN [48] have demonstrated that appropriate manipulation of encodings can lead to novel, realistic images within the targeted data distribution. However, since our modification of the encoding is stochastic by design, this makes the task of producing realistic images more difficult. Furthermore, despite the compression of features to a global level, each encoding feature is independently perturbed, leaving the possibility for unrealistic output images due to resulting combinations of noisy features that fall outside of the data distribution.

Specifically, the decoder must be trained to accept noisy encodings and produce realistic faces depicted in the output images. The adversarial nature of the training process can be leveraged to achieve this. To this end, we modify the training process by injecting noise into the encodings of training samples before passing the encoding to the decoder.

The adversarial loss objective function employs a Wasserstein critic with a gradient penalty [82]. Let D be the discriminator/critic, GP be the gradient penalty function and λ_{GP} be a hyperparameter that controls the weight of the gradient penalty. For a distribution p_{data} of training images, a distribution p_{attr} of binary attribute vectors and a distribution p_{pose} of real-valued pose parameter vectors, the discriminator and generator losses are defined as follows:

$$L_{adv_d} = -\mathbb{E}_{x^{a,p}\sim p_{data}} D(x^{a,p}) + \mathbb{E}_{x^{a,p}\sim p_{data}, b\sim p_{attr}, q\sim p_{pose}} D(\tilde{x}^{\hat{b},\hat{q}}) + \lambda_{GP}\mathbb{E}_{x^{a,p}\sim p_{data}, b\sim p_{attr}, q\sim p_{pose}} GP(x^{a}, \tilde{x}^{\hat{b},\hat{q}}), \quad (4)$$

$$L_{adv_g} = -\mathbb{E}_{x^{a,p}\sim p_{data}, b\sim p_{attr}, q\sim p_{pose}} D(\tilde{x}^{\hat{b},\hat{q}}). \quad (5)$$

The preservation of specific facial attributes in obfuscated images is also one of the goals for our model. We keep the attribute classification loss in the same form as originally used in AttGAN. Let C denote the classification network and $C_i(x)$ denote the application of the network to classify the ith element of a vector of binary facial attributes for an image x. The loss function is as follows:

$$\ell_a(x^{a,p}) = \sum_{i=1}^{m} -a_i \log C_i(x^{a,p}) - (1-a_i)\log(1 - C_i(x^{a,p})), \quad (6)$$

To ensure that the injection of noise does not alter the attributes we wish to preserve, we modify the classification loss function of the generator such that it is calculated for training samples to which noise has been injected:

$$L_{cls_g} = \mathbb{E}_{x^{a,p}\sim p_{data}, b\sim p_{attr}, q\sim p_{pose}} \ell_a(\tilde{x}^{\hat{b},\hat{q}}). \quad (7)$$

The loss for the classification network remains as it originally was with the exception that the pose parameters are now added to the notation:

$$L_{cls_c} = \mathbb{E}_{x^{a,p}\sim p_{data}} \ell_a(x^{a,p}). \quad (8)$$

The change from row b to row c in Fig. 2 demonstrates the improvements in the visual quality of the obfuscated faces using noisy samples during the training process. To more clearly discern the impact of training with noisy encodings in Fig. 2, we provide an additional example in Fig. 3 which demonstrates the difference between training with and without noisy encodings once all other changes proposed throughout the remainder of the section have also been applied. Here, the images produced by the model trained without noisy encodings show obvious distortions, particularly around the hair and clothing of the obfuscated individuals.

5.3. Pose preservation

Given that images in unconstrained settings capture individuals with a wide variety of poses, it is desirable to preserve this information in the obfuscated images. Unless care is taken in the design and training of the model to manage specific types of information, it will remain entangled in the encoding of the image. As a result, it is also subject to perturbation when noise is applied to the encoding. For example, if an image depicts an individual with their head turned to the side, the addition of noise to the encoding may alter the pose of the head, likely changing it to a forward-facing pose as this is typically the predominant orientation in training data. This produces an output that may look unnatural and destroys important contextual information in the image. To address this, we extend the model architecture and training process to accommodate for additional decoder inputs that capture pose information.

In order to extract pose information from images, we employ RingNet [83], a neural network which is able to regress 3D facial model parameters from 2D input (images). An important aspect of these parameters is a separation of identity-specific information, captured in shape parameters, from the non-identity-specific information of camera position, pose and facial expression. This separation allows us to target the non-sensitive information that is of interest to us. The RingNet model uses 3 parameters that capture axis-angle representations of the global rotation for the depicted head. Since we align all input images for our model (as described in our training details,

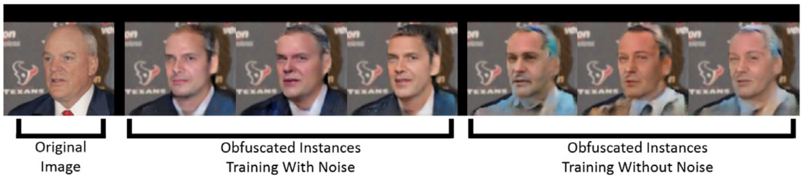

Original Image
Obfuscated Instances Training With Noise
Obfuscated Instances Training Without Noise

Fig. 3. A comparison of the impact of training with and without noise added to the sample encodings while using all other proposed modifications. The obfuscated instances from the model trained without noise show unnatural distortions of colour and texture.

Table 3
Discriminator, classifier and pose regressor architectures. IN refers to instance normalization, LReLU refers to the leaky rectified linear units activation function and FC refers to a fully connected layer.

Discriminator	Classifier	Pose regressor
	Input: 128 × 128 × 3 image	
	Conv (4 × 4 × 64, Stride = 2), IN, LReLU	
	Conv (4 × 4 × 128, Stride = 2), IN, LReLU	
	Conv (4 × 4 × 256, Stride = 2), IN, LReLU	
	Conv (4 × 4 × 512, Stride = 2), IN, LReLU	
	Conv (4 × 4 × 1024, Stride = 2), IN, LReLU	
FC (1024), LReLU FC (1)	FC (1024), LReLU FC (1)	FC (1024), LReLU FC (2)

Section 8.1), we require only the yaw and pitch angles. We use RingNet to annotate these two parameters for both training images as well as the images to be obfuscated.

We modify the decoder architecture to concatenate these parameters along with the vector of facial attributes. We modify the discriminator by adding an extra branch (parallel to the branches of the discriminator/critic and the classifier) with two fully connected layers to approximate the parameters. Details on the architecture of these networks can be found in Tables 2 and 3.

Let P denote the pose regression network embodied by the added branch and let $P_i(x)$ denote the output of P for the ith pose parameter of an image x. In order to train the model on these new parameters, we introduce an additional loss function to minimize of the mean squared error in the approximated values:

$$\ell_p(x^{a,p}) = \sum_{i=1}^{k} \frac{(P_i(x^{a,p}) - p_i)^2}{k}, \tag{9}$$

$$L_{pose_r} = \mathbb{E}_{x^{a,p}\sim p_{data}} \ell_p(x^{a,p}), \tag{10}$$

$$L_{pose_g} = \mathbb{E}_{x^{a,p}\sim p_{data}, b\sim p_{attr}, q\sim p_{pose}} \ell_p(\tilde{x}^{\hat{b},\hat{q}}). \tag{11}$$

The change from row c to row d in Fig. 2 demonstrates the impact of pose preservation in the obfuscated images.

5.4. Objective function

With the new loss calculations laid out, we are now able to put together the full objective function for the model training process. We leave the reconstruction loss essentially untouched, updating only the notation to reflect the use of unmodified pose parameters as input for the decoder:

$$L_{rec} = \mathbb{E}_{x^{a,p}\sim p_{data}} \left\| x^{\hat{a},\hat{p}} - x^{a,p} \right\|_1 . \tag{12}$$

While high fidelity reconstruction of images is not explicitly a goal in our setting, it is desirable for the model to learn that a noiseless image should resemble the original image for the sake of preserving utility in features that are not captured in the other loss calculations. In this way, under a low-privacy setting, the injected noise will only make minor adjustments to the depicted identity, which may be desirable in many cases.

The full objective function is given below. The definitions for each of the loss functions, L_{rec}, L_{cls_g}, L_{pose_g}, L_{cls_c}, L_{adv_d} and L_{pose_r}, are given in Formulae (12), (7), (11), (8), (4) and (10), respectively. Each of the corresponding λ coefficients denote hyperparameters used to configure the relative importance of the objectives captured by each of the loss functions.

$$\min_{G_{enc}, G_{dec}} \lambda_{rec} L_{rec} + \lambda_{cls_g} L_{cls_g} + L_{adv_g} + \lambda_{pose_g} L_{pose_g}, \tag{13}$$

$$\min_{D,C,P} \lambda_{cls_c} L_{cls_c} + L_{adv_d} + \lambda_{pose_r} L_{pose_r}. \tag{14}$$

To visualize how the network components fit together in the loss calculations, we provide a diagram depicting the flow of data during the model training process in Fig. 4.

5.5. Image background

While the focus of our model is on the face depicted in an input image, it is inevitable that portions of the image will also capture the background behind the individual. The image could be tightly cropped to the face, however this is undesirable since it excludes major portions of the head, such as hair and ears, which would leak identifiable features. Although accurate representation of the background can be handled by many GANs, allowing the background to appear in the input image leads to an entanglement of background-specific features in the image encoding. The primary concern when this happens is the undesirable distortion of the background upon injection of noise, which negatively impacts visual quality. Additionally, the inclusion of background details in the image encoding wastes encoding capacity on features that are irrelevant to the goal of facial obfuscation. This, in turn, leads to a wasteful use of the privacy budget which is intended only to be spent on the facial features.

To address these issues, we use Mask R-CNN [84] to mask the area covered by the head in each image, allowing us to subtract all background content. By pre-processing all training images in this way, the model learns a distribution in which all images depict a head in a space with a white background. This also avoids the need for features to capture complex patterns in the background of the image. When an image is to be obfuscated, we similarly apply masking and background subtraction prior to passing the image as input to the trained model. When noise is added to the

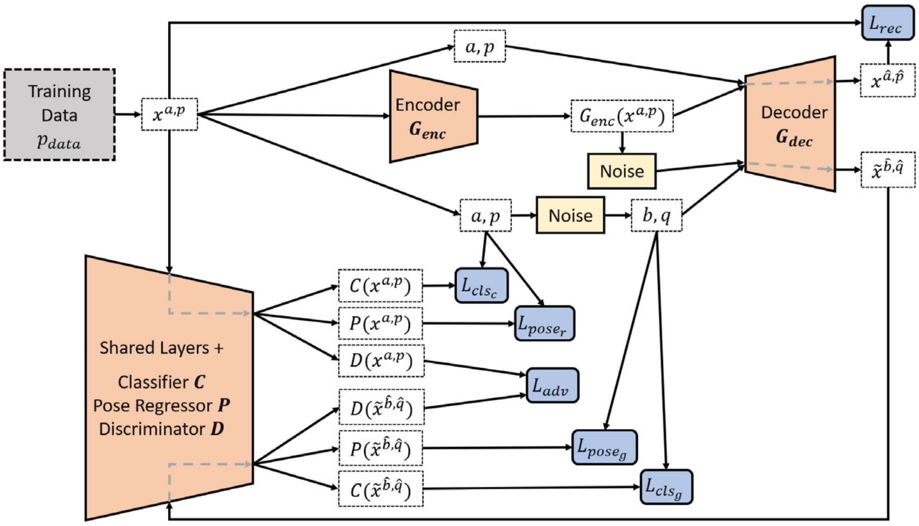


Fig. 4. Visual representation of the flow of data through the model during training. Dashed-edge boxes denote data, trapeze shapes denote network components and round-edged boxes denote loss calculations.

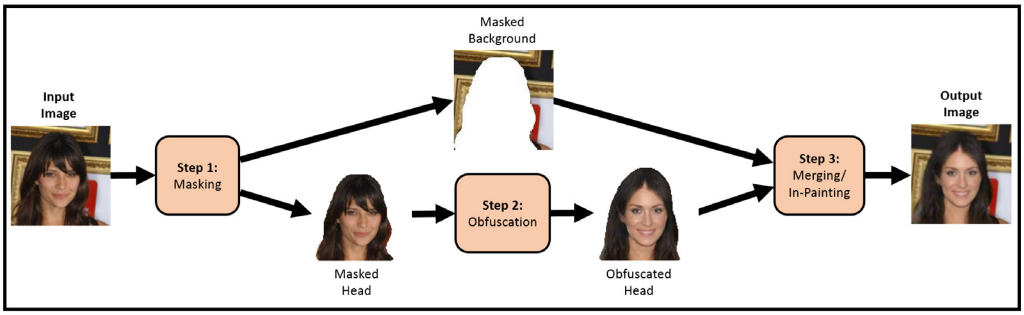


Fig. 5. Pipeline for the obfuscation process using masking. **Step 1:** the input image is masked to separate the head and the background. **Step 2:** Obfuscation is applied to the masked head. **Step 3:** The obfuscated head and masked background are recombined using an in-painting GAN.

encoding, the obfuscated image is likely to remain in the learned distribution, depicting a new identity over a white background.

Once an image has been masked and obfuscated, the new head must be recombined with the original image background. During the pre-processing step, we store the image background so that it can be later recombined. Some care must be taken to do this in a clean way. The area covered by the head, and in particular the hair, is likely to have changed slightly. A simple overlay of the obfuscated head onto the original background leaves the possibility for gaps to occur in the image where the new head no longer covers areas previously occupied by the original head. To fill in these gaps in a visually realistic manner, we apply Pluralistic Image Completion [85], a GAN-based approach to facial image in-painting.

Row e of Fig. 2 shows obfuscated images using the background masking and merging process. Unlike the obfuscated images in all other rows, the background remains untouched; only the depicted identity is changed. The pipeline for the process of masking, obfuscation and recombination is illustrated in Fig. 5.

6. Obfuscation within a PCA basis

While image encodings can be directly treated as the sensitive information to be obfuscated, better utility can be achieved by instead performing the obfuscation within a PCA basis. Since we employ distance-generalized differential privacy, we must carefully examine the measure of distance used in the new basis to ensure that a meaningful level of privacy is achieved. In this section, we explain how to manage the obfuscation process within a PCA basis and how to efficiently allocate the privacy budget. We then discuss how to update the model training process to accommodate these changes to the manipulation of image encodings such that we retain the ability to generate realistic obfuscated images. Lastly, we demonstrate the practical value of the proposed changes by comparing images obfuscated with and without the use of PCA.

6.1. Proposed usage of PCA

As discussed in Section 4.3, the application of differential privacy within a transformed PCA basis offers a number of benefits related to the preservation of utility and the management of the privacy budget. To obtain these benefits, we use the image encodings of our training data to learn a PCA transformation that is well-suited to the data we aim to obfuscate. When presented with a new image that requires obfuscation, we then encode it, transform the encoding to the PCA basis, apply noise, and transform the noisy data back to its original basis. This provides an obfuscated encoding that has been restored to a representation in its original basis, allowing for it to then be passed through the decoder as usual to generate the output image. This process is illustrated in Fig. 6.

Our usage of PCA with differential privacy differs from existing works on two counts. First, our training data is not private, therefore the task of learning the PCA transformation need not be done in a differentially private manner in our setting. Second, since

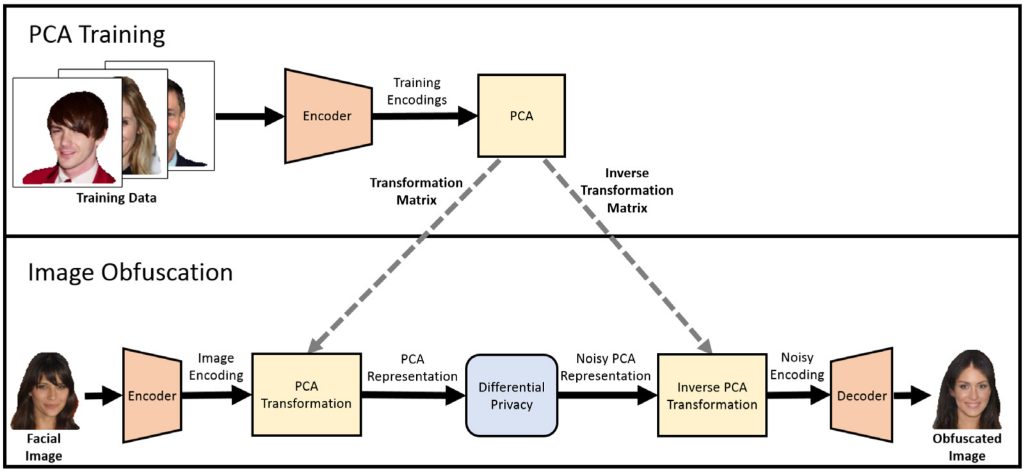


Fig. 6. Pipeline for the PCA training and image obfuscation process. The image obfuscation process can be seen as an expansion of Step 2 in Fig. 5.

we employ a distance-based generalization of differential privacy, we no longer deal with the concept of query sensitivity and cannot apply approaches such as [78] to determine the required magnitude of noise within the PCA basis.

In order to properly control the addition of noise to encodings represented in the PCA basis, we employ the strategy of [35] to set up a distance measure and a mechanism appropriate for the protection of secrets in the form of encodings of facial identities. Let $X = [x_1, \dots, x_n]$ be the representation of an image encoding in the PCA basis. We use the distance measure defined in Formula (3) to interpret distinguishability between the PCA basis encodings. Recall from Section 4.1 that an encoding can be made differentially private by independently adding Laplace noise to each element of the encoding vector such that the element at the ith position uses a distribution with a scaling parameter of $\sigma_i = \frac{n(max_i - min_i)}{\epsilon}$, where $(max_i - min_i)$ is a scaling factor used to ensure that the range of element is [0, 1]. The element-wise range scaling prevents features with large ranges from dominating the distance measurement, ensuring each element receives noise configured to the appropriate magnitude. Furthermore, the use of an L_1 distance measure is well-suited to representations in a PCA basis given that the principal components are linearly uncorrelated. We provide practical details for the interpretation of the resultant privacy guarantee in Section 7.1.

6.2. Privacy budget allocation

Beyond determining an appropriate distance measure, allocation of the privacy budget should also be considered. In a setting where each element of the secret carries roughly equal information, it is logical to evenly distribute the privacy budget across the elements. However, in the case of an encoding represented in a PCA basis, the information content of the elements decreases monotonically, and often quite significantly, from the first element to the last. We can take advantage of this property by selectively dropping trailing principal components in order to focus the privacy budget on the remaining components with higher information content.

Dropping components has the effect of reducing the dimensionality of the PCA basis representation. When data is transformed from the PCA basis back to its original basis, dropped components are treated as if they had assumed their mean value from the training dataset. For low variance components, this can have the effect of introducing relatively little error in the data while attaining perfect privacy protection since the original value of an encoding within the component is entirely removed form the released data. This is particularly valuable for dealing with components that have a low variance in relation to their overall range since the range-based scaling used in the configuration of the privacy mechanism would likely lead to a scaling parameter much higher than the standard deviation of the component. Given that such a high scaling parameter would induce heavy perturbation of the data, use of the mean value (calculated from a set of low variance points) is likely to induce a much smaller error on average. As a result, carefully choosing which components to drop can achieve lower expected error on these components while simultaneously providing them with perfect privacy. Further to this, the privacy budget that would have been spent on these components is freed up to instead be spent on components with greater information content.

To guide the process of determining which components to drop, we propose the use of a binary search. Given a user's privacy budget ϵ, the algorithm will search the range of $[1, n]$, where n is the number of features in the original encodings, to find the best choice for how many of the leading components to retain. The search is guided by a user-specified threshold on the acceptable ratio of each scaling parameter used by the mechanism to the standard deviation (within the training data) of the corresponding component. For example, the user may specify that the scaling parameter used for each retained element may be at most 1.5 times the standard deviation of its component. Since the dimensionality of the PCA basis appears in the numerator of the scaling parameters, a reduction in dimensionality leads to a reduction in the magnitude of required noise. If the ratio for any of the leading components exceeds the threshold, the privacy budget is spread too thin and more components must be dropped. If none of the ratios exceed the threshold, the budget can allow for a greater number of components to be retained. The search terminates when the boundary between these two cases is found. Recall that the principal components are ordered according to their explained variance in the training data. This approach therefore determines the selection of components that retains the maximum explained variance in the data subject to the constraint on the scaling parameter threshold. We provide the steps in Algorithm 1.

6.3. Model training

With the details for the mechanism configuration and budget allocation laid out, we must next ensure that the GAN is trained

Algorithm 1: Budget Allocation

```
Input: Encoding size $n$, Privacy budget $\epsilon$, Ratio threshold
       $\alpha$, Component standard deviations $s_1, ...s_n$,
       Component ranges $(min_1, max_1)$, ..., $(min_n, max_n)$
$c_{min} \leftarrow 1$
$c_{max} \leftarrow n$
$c_{mid} \leftarrow \lceil \frac{c_{min}+c_{max}}{2} \rceil$
while $c_{mid} \neq c_{max}$ do
    reduce $\leftarrow$ False
    for $i \leftarrow 1$ to $c_{mid}$ do
        if $\frac{c_{mid}(max_i - min_i)}{\epsilon} \geq \alpha s_i$ then
            reduce $\leftarrow$ True
            break
        end
    end
    if reduce then
        $c_{max} \leftarrow c_{mid}$
    else
        $c_{min} \leftarrow c_{mid}$
    end
    $c_{mid} \leftarrow \lceil \frac{c_{min}+c_{max}}{2} \rceil$
end
return $c_{mid}$
```

to handle encodings that have been accordingly perturbed. As per the loss functions given in Section 5, our training process requires the addition of noise to training samples in order to teach the decoder how to produce realistic images from noisy encodings. Yet the addition of noise requires the use of the PCA transformation matrix while the calculation of this matrix (refer back to Fig. 6) requires a training set of encodings from the target model. In short, there exists a circular dependence between the GAN and the PCA transformation. Although one could first train a model without noise and use this to generate the PCA training encodings, the model later trained with noise will differ in its learned parameters, rendering the PCA training data inaccurate with respect to the target distribution of encodings.

To handle this, we propose to compute the PCA transformation matrix at the end of each epoch of training using a snapshot of the current encoder. The transformation matrix can then be used during the next epoch of training. During the first epoch of training, we simply forgo the addition of noise and during each subsequent epoch, we add noise using the most recent transformation matrix. In this way, the noise-adding process is periodically updated throughout training to reflect the changes in the encoder. This allows for the model to always have access to a reasonable approximation of how it would currently be used to add noise in practice.

Since the mechanism is intended to allow for a configurable level of privacy based on the user selected parameter ϵ, we train the model for a range of different privacy budgets. Each time training data is sampled, we draw a privacy budget uniformly at random from a range of typical budgets. This helps to train the decoder for the preservation of utility at different levels of privacy. We provide more specific details on the training configuration in Section 8.1.

6.4. Obfuscation examples

We conclude the section with a brief comparison of obfuscation with and without PCA in order to demonstrate the practical importance of applying obfuscation within the PCA basis. We note that the obfuscated instances presented in Figs. 2 and 3 were produced using an ϵ value of 3000. While the values of ϵ used for our instantiation of distance-generalized differential privacy must be interpreted on a different scale from those of standard differential privacy (as we later explain in Section 7), 3000 is nonetheless a high value in our framework and offers a poor level of privacy. Yet reducing the value of ϵ while using only the modifications proposed in Section 5 results in obfuscated images with poor visual quality due to a lack of control in the application of the increased magnitude of noise. In Fig. 7, we demonstrate the improvements in visual quality obtained through the use of obfuscation within the PCA basis when using values of ϵ that result in a strong level of privacy.

7. Privacy interpretation

Although differential privacy offers a formalized guarantee of privacy, the implications of the guarantee are not always directly obvious. In particular, the interpretation of the distance-generalized guarantee we employ, and thus the selection of an appropriate privacy parameter, may be difficult for most users. In this section, we first provide information on how our privacy guarantee relates to other applications of differential privacy, one of which provides a bound on the change in knowledge (regarding the sensitive information) of attackers. We use this to offer some intuition on the meaning of the privacy guarantee in relation to the choice of ϵ. We then discuss, in terms agnostic to any particular method of obfuscation, how sensitive information can leak from portions of images that are left unobfuscated (e.g., the background of the image) and how an attacker can exploit this.

7.1. Differential privacy guarantee interpretation

The differential privacy guarantee in Formula (2) provides an upper bound on the allowable difference between the probabilities of different inputs to the mechanism producing the same obfuscated output. This allows us to make an assertion that similar inputs have similar probabilities of producing the same output and are thus difficult to distinguish between. Since this is strictly a property of the mechanism, it holds regardless of the attacker's background knowledge. This is a useful property since it allows for reasoning about the level of privacy in a manner that is largely agnostic to the attack model. However, this degree of abstraction prevents the calculation of an actual level of re-identification risk for an obfuscated release. As a result, the privacy guarantee may be difficult for most users to interpret, making the task of selecting a value for ϵ rather challenging.

To help interpret the privacy guarantee, we first highlight a strong parallel between the distance measure we use and the concept of query sensitivity in standard differential privacy. Recall that we configure the randomization mechanism using a scaling parameter of $\sigma_i = \frac{n(max_i - min_i)}{\epsilon}$ for the ith element of the image representation. The numerator of the scaling parameter acts as a measure of sensitivity for each element of the image representation, capturing its potential for influence in the overall distance measure. This serves as a surrogate for the concept of query sensitivity used in standard differential privacy for databases. Contrast this scaling parameter against that of the standard Laplace mechanism for databases, $\sigma = \frac{\Delta F}{\epsilon}$, where ΔF is the query sensitivity. Both scaling parameters are linear functions of a measure of sensitivity that captures the maximum amount by which a value can change. They are both, therefore, similarly sensitive to a worst-case interpretation of how much information could be revealed in the released data.

An important distinction in the interpretation of the distance-based guarantee is that without any notion of adjacency between

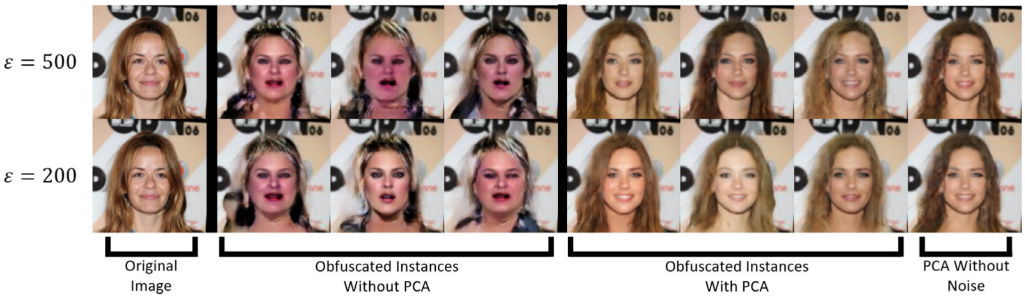

ε = 500
ε = 200
Original Image
Obfuscated Instances Without PCA
Obfuscated Instances With PCA
PCA Without Noise

Fig. 7. A comparison of obfuscation with and without the use of PCA to control the application of the noise. The instances which do not use PCA demonstrate darkened facial features and distortions in skin tone, hair style and facial expression. The final column is included as a point of reference to show the visual impact of the dimensionality reduction from PCA without the application of noise.

secrets, we are considering a worst-case over all possible image encodings. This is akin interpreting the standard differential privacy guarantee in terms of maximally distant databases as opposed to adjacent databases. As such, it is important to recognize that the values of ϵ applied to our mechanism will have an entirely different meaning than those commonly seen in differential privacy for databases. It is in fact possible to capture the standard differential privacy guarantee in the distance-generalized framework by (1), considering the domain of secrets to be the set of all possible database configurations, and (2) using the Hamming distance between pairs of databases as the distance measure [66]. Within this setting, the range of the distance measure is $[1, n]$, where n is the number of records in the database. The privacy guarantee for adjacent databases corresponds to pairs of secrets at distance 1, a small fraction of the total range of the distance measure. Although there is no precise correspondence between Hamming distance and our distance measure, it is clear that the guarantee would be most meaningfully interpreted in terms of a small fraction of the range of the distance measure.

Moving away from the concept of differential privacy for databases, we can also contrast our privacy guarantee against that of geo-indistinguishability [86]. Geo-indistinguishability is an application of distance-generalized differential privacy to protect location data, using a measure of physical distance to configure the privacy mechanism. A key intuition in geo-indistinguishability is that location data is desirable to reveal at an approximate level in order to allow for utility in the released information. Only pairs of locations that are nearby should remain highly indistinguishable. The same notion applies to our setting of facial obfuscation in the sense that distinguishability between "distant" pairs of facial identities is acceptable and desirable for the sake of utility. To help interpret the practical implication of this, we highlight a bound used in geo-indistinguishability that captures the maximum possible change (due to the release of obfuscated data via the privacy mechanism) in the probability that the secret has a particular numeric value. Let $B_r(X)$ be the set of secrets within distance r of X and let $\Pr(X|B_r(X))$ be the probability with which the attacker initially believes the secret to have the value X, while already possessing prior knowledge of the set $B_r(X)$. The probability after additionally observing the obfuscated release R is bounded as follows:

$$\Pr(X|R, B_r(X)) \leq e^{\epsilon r} \Pr(X|B_r(X)) \quad \forall r > 0 \quad \forall X, R \in \mathbb{R}^n. \tag{15}$$

The usage of this bound is necessarily dependent on assumptions regarding an attack model due to the use of probabilities representing the attacker's prior beliefs about the likelihood of candidate identities. However, we provide an example of a possible scenario. Consider an attacker with no prior knowledge about the secret beyond the set $B_r(X)$. The attacker's prior knowledge can therefore be represented by the following uniform probability distribution over $B_r(X)$:

$$\Pr(X'|B_r(X)) = \frac{1}{|B_r(X)|} \quad \forall X' \in B_r(X). \tag{16}$$

Since the guarantee is intended to offer strong protection for similar secrets while allowing for some distinguishability between distant secrets, we are interested in sets of candidates (i.e., instantiations of $B_r(X)$) with relatively small radii. Consider a scenario in which the set of candidates uses a distance of $r = 0.1$ (recall that the total range of the distance measure is [0,1]) and contains 5000 identities. The application of Formula (15) indicates that the posterior probability of re-identification on an image obfuscated using $\epsilon = 50$ is upper bounded by roughly 2.97%.

To provide further intuition on practical implications of the privacy guarantee, we show the distribution of inter-identity distances on a dataset of facial images in our experiments (Section 8.6).

7.2. Information leakage via auxiliary model inputs

In some generative models of facial obfuscation, such as ours, auxiliary inputs beyond the facial identity are provided to the model for the purpose of utility. These are inputs such as pose and gender which are considered useful pieces of information that are intentionally conveyed in the released image. Although some information about the identity is revealed via personal attributes such as gender, this is considered acceptable and the inputs are otherwise seen as being benign. However, there exists potential for unintended leakage of information if the auxiliary inputs are generated from other machine learning models. Recent work [87] has shown that representations learned by a neural network model can be re-purposed to perform unintended tasks that reveal sensitive information about the input data. For example, a binary classifier for gender in facial images will learn high-level facial features that are used to perform the gender classification task. The representation of these features in the penultimate layer of the network may also reveal information about other facial attributes such as ethnicity.

The attacks studied in [87] rely on access to an intermediate representation of the model input prior to its final classification output (i.e., hidden layer outputs). Such intermediate representations carry significantly more information about the input image than the final output of the model, which is highly specialized for a particular task. In the context of facial obfuscation, it is strictly the outputs of the final layer that an attacker could access. Yet, since a typical model enforces no formalized guarantee on the absence of information leakage in its outputs, it stands to reason that some minor degree of sensitive information may yet be present. This may occur due to imperfect disentanglement of features by the model or known bias in the model outputs. We do not study this aspect of information leakage but leave it as an open problem as to how much of a concern it may be for facial obfuscation.

We emphasize that this is not something unique to our application of differential privacy for facial obfuscation. Any method of obfuscation that relies on the use of automated generation of auxiliary inputs for utility-related objectives has the same potential for information leakage. Although we do not investigate these, we propose three different approaches to address this concern, should it be deemed necessary in practice:

- **Require manual entry of auxiliary inputs wherever possible.** This avoids the potential for inferences based on knowledge about models that generate the inputs. A practical example is the use of a user profile that requires entry of desired facial attributes only once and then applies the attributes each time an image of the user is to be obfuscated.
- **Apply differential privacy to the auxiliary inputs.** The degree of perturbation can be configured based on the expected risk associated with the inputs. Our approach to generalized differential privacy in a PCA basis can be applied to continuous values such as the RingNet parameters. For categorical values such as facial attributes, the exponential mechanism [88] could be applied.
- **Train a model to only use the image encoding without auxiliary inputs.** This ensures that all sensitive information is subject to the formalized privacy guarantee. In the context of differential privacy, the degree of utility preserved then becomes entirely dependent on the configuration of the mechanism.

8. Experiments

In this section, we present the results of an experimental evaluation of our proposed method of obfuscation and a comparison against other alternatives which offer a formal guarantee of privacy. We begin by providing the training details used for our model and the experimental setup. We then give an empirical evaluation of the level of privacy obtained through our model over a range of potential choices for the privacy budget ϵ. Lastly, we provide a comparison against other approaches in terms of a trade-off between re-identification risk and measures of utility.

The experiments were run on a machine using a GeForce GTX 1080 Ti graphics card and 24 GB of RAM. The Tensorflow and Pytorch libraries were used to train and execute all neural network models.

8.1. Training details

We train our model using the CelebA [89] dataset, a collection of 202,599 facial images annotated with identity, 40 binary facial attributes (e.g., gender, bearded, wearing makeup, etc.) and 5 facial landmark locations. The images contain a variety of poses and sizes of heads (in terms of their pixel coverage). We align and crop the training images to bounding boxes with ample space for the full head and hair using HD CelebA Cropper [90] and resize the cropped images to 128 × 128 pixels. We generate pose parameters for the processed images using RingNet [83] and standardize the pose parameters across the training data such that they have zero mean and unit variance.

To prevent the preserved vector of facial attributes from leaking too much information, we choose only to keep the attribute for gender. Following AttGAN, we uniformly perturb the facial attributes to generate the modified vector $b = [b_1, \ldots, b_m]$ given to the decoder during training. To generate the perturbed vector of pose parameters $q = [q_1, \ldots, q_k]$ we draw from a normal distribution using the original vector p as the location parameter and a scaling parameter of 1. We follow the AttGAN configuration of hyperparameters with the exception of λ_{cls_g} which we reduce to 0.5 due to the reduced number of attributes, and λ_{rec} which we increase to 200 for better preservation of visual quality when dealing with noisy encodings. We set our additional objective coefficients of λ_{pose_r} and λ_{pose_g} to 2 and 20, respectively.

When adding noise to the image encodings during training, we randomly draw a value of ϵ for each batch of images using a uniform distribution over the range [100,1000]. We use a threshold of 1.3 for the ratio of mechanism scaling parameters to standard deviation values. This exceeds the ratio of 0.9 that we use at the time of obfuscation in order to ensure that the model is trained to easily handle this magnitude of noise. If any elements of the noisy encodings fall outside of the range of values observed in the training data, we remap the out-of-bounds values to the boundary of the range. This is done to better preserve visual quality in the obfuscated images by restricting the encodings to remain closer to the distribution learned by the model.

8.2. Experimental setup

We aim to investigate three measures of success in the images obfuscated by our approach: (1) privacy, (2) visual quality, and (3) utility. To measure privacy, we use FaceNet [18], a neural network that has reported excellent results in the classification of facial identities. To measure visual quality, we use SSIM [91], a measure of structural similarity between images which is inspired by principles of the human visual system. Lastly, to measure utility in terms of preservation of specific information, we employ a neural network trained for the task of gender classification. For this, we use the DeepFace [92] framework which implements the DEX [93] gender classification model.

We run our experiments on the FaceScrub [94] dataset, a collection of 107,818 unconstrained setting facial images across 530 different identities. As with the CelebA dataset, the images contain a variety of poses and sizes of heads. However, annotations are provided only for identity and gender. Since not all identities have the same number of images we take 50 images per individual to avoid bias due to over or under-representation of some individuals. This leaves us with 506 identities which meet the minimum image requirement. We split the data using 40 images per identity for training FaceNet and the remaining 10 for testing the obfuscation. For all experiments, we pre-process the data with an identical process to that used for our training data, using Dlib [95] to generate the facial landmarks that are needed for the alignment process. In order to obtain the facial attribute annotations needed by our model, we use our classification network to generate the annotations. This is done to simulate a more realistic setting in which annotations cannot be expected to already exist for the images to be obfuscated.

We compare our proposed method of obfuscation against two alternative approaches which also offer a formalized privacy guarantee: k-same obfuscation and in-painting. We implement k-same obfuscation within the encoding space of our trained model using the clustering algorithm as specified for k-same-m [39]. As such, our implementation of k-same obfuscation gains all the benefits of our proposed GAN architecture and serves as a comparison strictly between the formal privacy guarantees of differential privacy and k-same obfuscation. For obfuscation via in-painting, we use the publicly available pre-trained DeepPrivacy [40] model. A visual comparison of obfuscated output from each of these methods is shown in Fig. 8.

During obfuscation, should any of the elements in the encodings of the test images fall outside of the bounds defined by the encodings of our training data, we remap the out-of-bounds

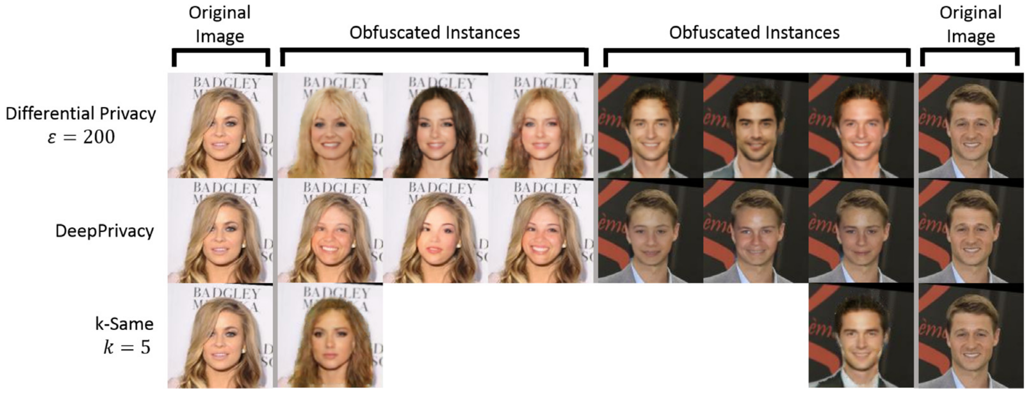


Fig. 8. Examples of output images from the three methods of obfuscation applied in our experiments. The top row shows our proposed method of differentially private obfuscation using $\epsilon = 200$. The middle row shows the in-painting approach of DeepPrivacy. The bottom row shows k-same obfuscation using $k = 5$. Only one obfuscated image per identity is shown for k-same as it is a deterministic process.

elements of the testing encodings to boundary of the training data range. This step ensures that the encodings are suitable for the distance measure that we use. The configuration of the mechanism is based on the assumption that we know the range of each element in the encoding in order to appropriately scale the magnitude of the noise applied to the elements. However, in practice, we can only approximate the range using our training data. The adjustment of out-of-bounds values in the test data ensures that all encodings we encounter in practice conform to the assumed range. With a sufficiently large training set, occurrences of new encodings with out-of-bounds elements are unlikely to be common. Furthermore, any adjustments made in this way will have only a minor impact on the original encoding and thus a negligible impact on the obfuscated output given that noise is to be subsequently added regardless. After the application of noise to the encoding, we perform the same process of adjusting out-of-bounds values as described for the handling of noisy training encodings.

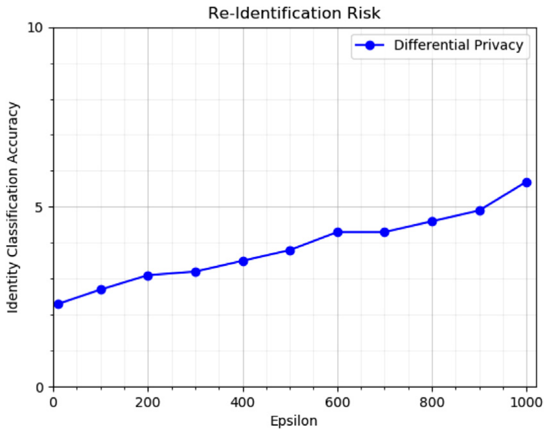


Fig. 9. Results on the impact of the privacy budget ϵ on re-identification risk. Identity classification accuracy is plotted as a function of ϵ. For a practical interpretation of the ϵ values shown in this plot, we refer the reader back to Section 7.1.

8.3. Re-identification risk

We begin with an evaluation of the impact the privacy budget ϵ has on re-identification risk. To calculate the level of re-identification risk, we pass obfuscated instances of images from the FaceScrub test partition to a trained FaceNet model and measure the identity classification accuracy over all instances. The classification accuracy is the percentage of obfuscated images that are correctly classified as their true identity by the model. Since differentially private obfuscation is stochastic, we generate 3 obfuscated instances per test image. We plot the accuracy as a function of ϵ to provide empirical results on the level of privacy attained at various choices of the privacy parameter. In other words, the experiment is repeated across a range of ϵ values. For each experiment, we obtain a measure of re-identification risk associated with the value of ϵ that was used.

We use FaceNet to execute a parrot attack [34] on the obfuscated images, training a different model for each value of ϵ we test in order to exploit the ability of a classification network to learn patterns in methods of obfuscation. To do so we take the set of training set of images for FaceNet and pass them through the obfuscation process (configured with the targeted value of ϵ), again generating 3 instances per image. The obfuscated output is used to train the FaceNet model such that it learns to classify images subject to the targeted obfuscation as best as it is able.

The results are shown in Fig. 9. The percentage value of the identity classification accuracy is plotted as a function of ϵ. Recall from Section 7.1 that the multiplicative bound in the differential privacy guarantee is given by $e^{\epsilon d(X_1,X_2)}$ where the distance $d(X_1, X_2)$ between a pair of identities represented by X_1 and X_2 is most meaningfully interpreted in terms of nearby identities (e.g., at distance 0.05). We do not show the other methods of obfuscation in this experiment since they cannot be plotted as a function of the privacy parameter used by our method. To provide a meaningful comparison against the other methods, the subsequent experiments examine the methods of obfuscation in terms of a privacy-utility trade-off.

The baseline classification accuracy achieved by FaceNet on the unobfuscated testing set is 92.6%. The plotted values show a significant drop in classification accuracy from this baseline. This demonstrates the ability of our proposed method of obfuscation to provide a strong level of privacy, even when subjected to a parrot attack. Although the classification accuracy does not drop as low as that of random guessing, this is to be expected if the mechanism is to retain useful information in the obfuscated output. The intention with differential privacy is not to guarantee that no identifiable information is leaked but rather to limit the distinguishability of the released information. Therefore, some of the released information can naturally be exploited by the parrot-trained classification model. However, the degree of success the model achieves at classification is greatly hampered, as we have shown by our results.

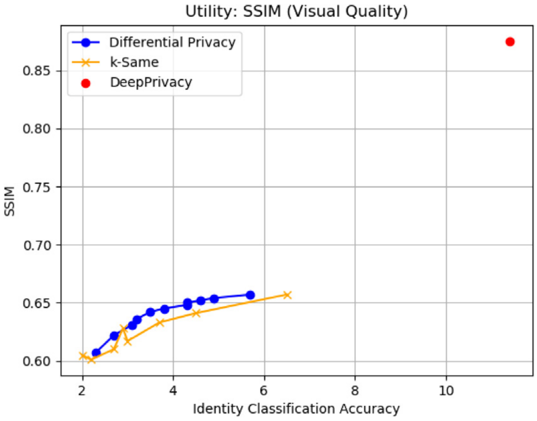


Fig. 10. A comparison of the trade-off between visual quality and re-identification risk. SSIM is plotted as a function of identity classification accuracy.

8.4. Visual quality

We next turn to the goal of preservation of visual quality in the obfuscated images. The measure of SSIM can be used to determine the similarity of an obfuscated image to its original instance. A measure of 1 indicates identical images whereas a measure of 0 indicates no structural similarity, therefore, higher values imply better visual quality. We plot the average SSIM over all obfuscated instances as a function of identity classification accuracy. This allows us to compare the methods of obfuscation in a manner that abstracts from approach-specific privacy parameters. The resulting plots can be interpreted as a representation of the trade-off between privacy and visual quality where high SSIM with low classification accuracy is desirable.

Each plotted data point is the result of applying a method of obfuscation with a particular privacy parameter configuration. For differential privacy, we apply values of 10, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 for the privacy budget. For k-same obfuscation, we apply values of 2, 3, 4, 5, 10, 20, 50 and 100 for k. Unlike differential privacy and k-same, DeepPrivacy has no configurable privacy parameter. Therefore, rather than producing a plot, it produces a single point of data. For each method of obfuscation, we calculate the identity classification accuracy when subjected to a parrot attack. As such, we train a separate FaceNet model for each plotted point of each obfuscation method. The SSIM values for each point are calculated as the average SSIM across all pairs of obfuscated images and their original counterparts.

As the approach of DeepPrivacy is stochastic in nature, we generate 3 obfuscated instances per test image as with differential privacy. Since k-same is deterministic, we generate a single obfuscated instance per test image. Due to a requirement of k-same obfuscation for each identity to appear at most once in the set of images to be obfuscated, we partition the k-same testing set into 10 subsets, each of which contains a single image of each identity. These galleries are then further subdivided based on equivalence classes for the gender attribute. This is necessary to ensure that clusters do not contain images from different equivalence classes as this would directly violate the k-same guarantee given that the obfuscation GAN is designed to reveal these attributes.

The results are shown in Fig. 10. Unsurprisingly, DeepPrivacy offers the highest level of SSIM in its obfuscated output, but at the cost of having the highest re-identification risk, since it modifies only a tightly cropped area containing the face rather than the whole head. In contrast, obfuscation via our proposed GAN modifies the hair, forehead, ears and neck as well. Furthermore, unlike DeepPrivacy which aims to match the new face to the unmodified facial contour, our GAN is able to make adjustments in skin tone. This leads to lower similarity to the original images and thus a lower SSIM score. However, the greater degree of modification is critical in achieving a sufficient level of privacy in the obfuscated output. A classification network trained for a parrot attack is able to recognize that the outer contour of the face remains invariant under tightly-cropped facial redaction and will thus focus on the features in the contour when attempting to classify identities. This is witnessed by the significantly higher identity classification accuracy measured for DeepPrivacy compared to the other methods of obfuscation. We argue that the drop in visual similarity to the original images under the more thorough modification of our GAN is a necessary sacrifice in order to achieve a reasonable degree of privacy.

The methods of differential privacy and k-same obfuscation appear to follow a similar trend to each other in the trade-off between re-identification risk and visual quality, likely due in part to the use of the same GAN for the generation of the output images. However, differential privacy demonstrates a slightly better trade-off.

8.4.1. Impact of PCA on visual quality

A related question to the impact of obfuscation on visual quality is the impact of the PCA dimensionality reduction on visual quality. In our proposed usage of PCA, stronger levels of privacy imply fewer preserved components, leading to a greater loss of visual detail from the original image. In order to examine the loss in visual detail purely in relation to the number of preserved components, we apply dimensionality reduction without the addition of any noise. By measuring the SSIM of the resulting decoded images in comparison to their original versions, we are able to empirically measure the loss in visual quality induced by the use of PCA. As before, we measure the overall visual quality as the averaged SSIM across all pairs of decoded images and their original counterparts. We repeat this experiment across a range of values, from 1 to 500, dictating how many of the PCA basis components are to be preserved out of the full 4096. The results are shown in Fig. 11, where the averaged SSIM value is plotted as a function of the number of preserved components.

The results above 100 preserved components show little variation in the average SSIM due to the fact that very little of the variance from the data is lost when dropping the majority of the principal components. This demonstrates how effectively the visual information is compressed into a small number of the leading components. From the plotted values, it is clear that the majority of the information pertinent to visual quality is preserved in the first 100 components. In Table 4, we provide information pertaining to the impact on visual quality from the dimensionality reduction in our privacy budget allocation scheme. For each value of ϵ used in our experiments, we list the number of preserved components and the corresponding explained variance across those components. Note that explained variance is a value in the range of [0,1] such that 0 indicates complete loss of all information and 1 indicates lossless preservation of information.

8.5. Utility

Beyond the generation of images with high visual quality, we also aim to preserve specific features in the obfuscated images. Here, we test for the ability to preserve the depicted gender. The FaceScrub dataset provides a gender label for each image and has an equal number of male and female identities. Using these labels as the ground truth, we measure the gender classification accuracy and plot this as a function of identity classification accuracy. The gender classification accuracy is calculated as the

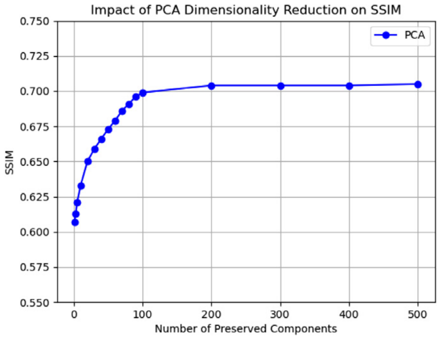


Fig. 11. A demonstration of the impact of PCA dimensionality reduction on the visual quality of the decoded images. SSIM is plotted as a function of the number of preserved components.

Table 4
A listing of the number of preserved principal components and the corresponding explained variance for a range of ϵ values.

ϵ	Preserved components	Explained variance
1000	86	0.7932
900	80	0.7759
800	72	0.7498
700	63	0.7148
600	54	0.6731
500	46	0.6294
400	37	0.5716
300	30	0.5186
200	20	0.4248
100	10	0.2963
50	5	0.2003
10	1	0.0642

percentage of obfuscated images classified as the correct gender by the DeepFace classifier. The setup with respect to privacy parameters and the calculation of identity classification accuracy is kept the same as in Section 8.4. Low identity classification accuracy with high gender classification accuracy is desirable. The results are shown in Fig. 12. For reference, we include the baseline gender classification accuracy on the unobfuscated set of test images.

Differential privacy and k-same obfuscation show a very similar trend in the preservation of gender. Both methods of obfuscation in fact provide accuracy above the baseline for most of the plotted points. This is due to the attribute preservation training objective of the GAN which aims to produce output images that clearly depict the specified attributes. In contrast to the methods that use our GAN, DeepPrivacy obfuscation suffers from a significant drop in gender classification accuracy. This is due to the fact that it does not use auxiliary inputs to selectively preserve certain types of information as we do in our proposed model. While methods of in-painting could be extended with such inputs, this would likely further exacerbate the issue of high identity classification accuracy.

8.6. Practical interpretation of inter-identity distances

To provide further insight into the practical implications of the privacy guarantee in the context of facial obfuscation, we examine the distribution of inter-identity distances in the FaceScrub dataset. To do so, we take one image per identity from the test partition and measure all pairwise distances. This produces a distribution consisting of roughly 250,000 inter-identity distances. It is important to note that the distribution of distances is impacted by the set of principal components over which distances are measured. Though we do not apply any obfuscation in this experiment (as we are interested in distances between unobfuscated identities), we do apply dimensionality reduction according a specified privacy budget. In Fig. 13, we show two examples of distance distributions, using privacy budgets of $\epsilon = 50$ and $\epsilon = 100$. The plotted distributions show that in practice, most identities have a relatively small distance between themselves and many others. This supports the intuition that the privacy guarantee can imply strong levels of protection in practical settings due to the large numbers of individuals that exist within relatively small radii of distance.

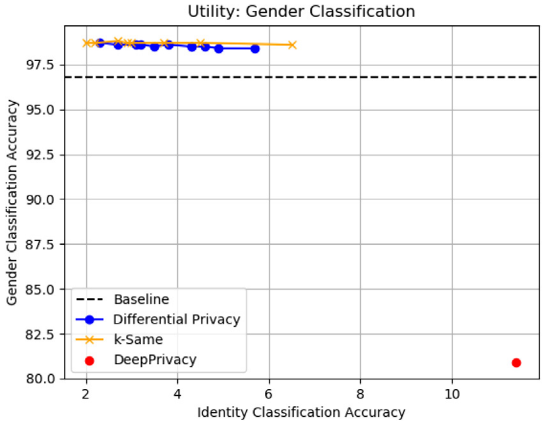


Fig. 12. A comparison of the trade-off between utility and re-identification risk. Gender classification accuracy is plotted as a function of identity classification accuracy.

To further assess the implications of the sampled distances, we extrapolate from the FaceScrub data to approximate the value of ϵ that would be required to achieve a maximum of 5% re-identification risk in a variety of scenarios. Let P be the full set of identities we are considering. We first measure P_r, the percentage of P covered by the set $B_r(X)$ averaged across all identities $X \in P$:

$$P_r = \frac{\sum_{X_i \in P} |B_r(X_i)|}{|P|^2}. \tag{17}$$

This calculation provides us with an approximation of the proportion of a population that is expected to fall within an instantiation of $B_r(X)$. With the FaceScrub data, we obtain $P_{0.1} = 0.1196$ and $P_{0.2} = 0.7481$ when using $\epsilon = 50$. We then rearrange the posterior probability bound of Formula (15) to calculate the value of ϵ required to achieve an upper bound of r_{post} re-identification risk on an obfuscated release:

$$\epsilon \leq \frac{\ln(|P| * P_r * r_{post})}{r}. \tag{18}$$

This form allows us to use our P_r values to consider scenarios involving other population sets. In Table 5, we provide examples for a variety of populations in which we require the re-identification risk to be upper bounded by 5% (i.e., $r_{post} = 0.05$). The population size $|P|$ is shown in the first column and the required value of ϵ is shown for radii of 0.1 and 0.2 in the second and third columns, respectively. The intent here is to provide intuition on how the generalized privacy guarantee can be understood in the context of facial obfuscation. We stress that this table is not intended to be used as a guideline in practice as it relies on assumptions regarding the attack model and extrapolation from a relatively small set of samples.

8.7. Discussion

As demonstrated through our experimental results, our proposed GAN is able to achieve a favourable trade-off between

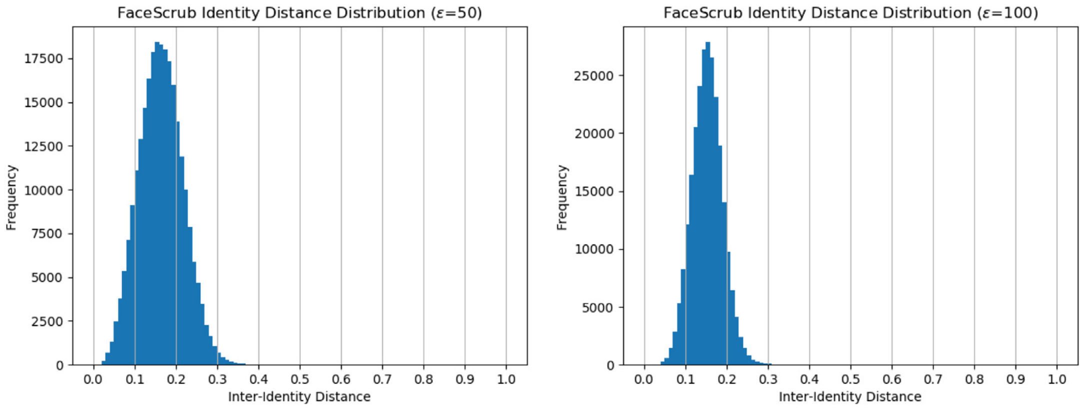


Fig. 13. Distribution over all pairwise inter-identity distances calculated on a set of images using one instance per identity from the FaceScrub test partition. The left distribution uses dimensionality reduction corresponding to a privacy budget of $\epsilon = 50$ while the right distribution uses a privacy budget of $\epsilon = 100$.

Table 5
Examples of the ϵ values required to achieve a maximum re-identification risk of 5% in various scenarios. Each row corresponds to a different population within which the attacker knows the targeted individual exists. Each column corresponds to a radius within which indistinguishability is required. Cell entries give the ϵ value required under the specified settings.

	$r = 0.1$	$r = 0.2$
7,900,000,000 (Global population)	$\epsilon \leq 176.71$	$\epsilon \leq 97.52$
1,000,000 (City population)	$\epsilon \leq 86.96$	$\epsilon \leq 52.65$
10,000 (Town population)	$\epsilon \leq 40.91$	$\epsilon \leq 29.62$
500 (Event attendance)	$\epsilon \leq 10.95$	$\epsilon \leq 14.64$

privacy and utility. Contrarily, the approach of redaction and in-painting used by DeepPrivacy falls short in the ability to sufficiently hamper re-identification risk. Without a configurable privacy parameter, methods of in-painting have no means to further reduce re-identification risk aside from widening the area of redaction. This type of all-or-nothing approach does not lend itself well to balancing privacy with utility.

The comparison between differential privacy and k-same obfuscation demonstrates similar levels of utility from the two methods of obfuscation with differential privacy performing slightly better for the measure of SSIM. Empirically, the two methods appear to perform very similarly when using the same GAN. Yet, differential privacy has been shown to achieve stronger privacy than k-same obfuscation against composition attacks and is better suited for protection against inferences in more practical scenarios where an attacker is likely to exploit background knowledge [35]. Given the stronger theoretical properties of differential privacy combined with comparable levels of utility to k-same obfuscation, we propose that our method of obfuscation offers a better overall privacy-utility trade-off.

9. Conclusions

In digital images depicting faces, privacy violations are often a concern. In this work, we propose a method of facial obfuscation that offers a provable guarantee of privacy while preserving utility in the obfuscated images. Our work allows us to provide a differentially private guarantee for GAN image encodings. Through careful design of the model architecture and training process, we achieve photo-realistic obfuscated images while preserving desired features such as gender and pose. We propose a novel combination of PCA with distance-generalized differential privacy to control the application of noise, allowing the privacy budget to be spent in an efficient manner. Through experimental comparisons, we demonstrate that our proposed approach achieves a strong level of privacy protection while preserving favourable levels of utility in the obfuscated images.

As the field of machine learning, and more specifically the study of GANs, is advancing rapidly, we expect that greater preservation of visual quality and utility in the obfuscated images can be attained by applying our methods to cutting-edge architectures and training methodologies. Improved handling of extreme variation in pose could also be achieved through the augmentation of the model training data with samples that capture a more balanced distribution over such types of variation. Furthermore, the investigation of alternate privacy budget allocation schemes within the PCA basis may prove fruitful in achieving improved privacy-utility trade-offs. Other interesting extensions of our work to explore include the application of differentially private obfuscation via GANs that handle full-body images or entirely different domains of sensitive information.

CRediT authorship contribution statement

William L. Croft: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Funding acquisition. **Jörg-Rüdiger Sack:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Wei Shi:** Conceptualization, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Facebook by the numbers: Stats, demographics & fun facts, 2020, https://www.omnicoreagency.com/facebook-statistics/, accessed: January 19, 2020.

[2] J.R. Padilla-López, A.A. Chaaraoui, F. Flórez-Revuelta, Visual privacy protection methods: A survey, Expert Syst. Appl. 42 (9) (2015) 4177–4195, http://dx.doi.org/10.1016/j.eswa.2015.01.041.

[3] S. Ribaric, A. Ariyaeeinia, N. Pavesic, De-identification for privacy protection in multimedia content: A survey, Signal Process., Image Commun. 47 (2016) 131–151, http://dx.doi.org/10.1016/j.image.2016.05.020.

[4] A. Cavailaro, Privacy in video surveillance [In the Spotlight], IEEE Signal Process. Mag. 24 (2) (2007) http://dx.doi.org/10.1109/MSP.2007.323270, 168–166.

[5] T. Winkler, B. Rinner, Security and privacy protection in visual sensor networks: A survey, ACM Comput. Surv. 47 (1) (2014) 1–42, http://dx.doi.org/10.1145/2545883.

[6] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, L. Vincent, Large-scale privacy protection in google street view, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 2373–2380, http://dx.doi.org/10.1109/ICCV.2009.5459413.

[7] K. Taylor, L. Silver, Smartphone Ownership Is Growing Rapidly Around the World, But Not Always Equally, Tech. rep., Pew Research Center, 2019, https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf.

[8] M. Duggan, Photo and Video Sharing Grow Online, Tech. rep., Pew Research Center, 2013, https://www.pewresearch.org/internet/2013/10/28/photo-and-video-sharing-grow-online/.

[9] Facebook, 2020, https://about.fb.com/company-info/, accessed: January 19, 2020.

[10] Facebook for business, 2020, https://www.facebook.com/business/marketing/instagram#, accessed: January 19, 2020.

[11] Instagram by the numbers: Stats, demographics & fun facts, 2020, https://www.omnicoreagency.com/instagram-statistics/, accessed: January 19, 2020.

[12] A. Oeldorf-Hirsch, S.S. Sundar, Social and technological motivations for online photo sharing, J. Broadcast. Electron. Media 60 (4) (2016) 624–642, http://dx.doi.org/10.1080/08838151.2016.1234478.

[13] S. Kairam, J.J. Kaye, J.A. Guerra-Gomez, D.A. Shamma, Snap decisions? how users, content, and aesthetics interact to shape photo sharing behaviors, in: 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 113–124, http://dx.doi.org/10.1145/2858036.2858451.

[14] X. Hu, D. Hu, S. Zheng, W. Li, F. Chen, Z. Shu, L. Wang, How people share digital images in social networks: A questionnaire-based study of privacy decisions and access control, Multimedia Tools Appl. 77 (14) (2018) 18163–18185, http://dx.doi.org/10.1007/s11042-017-4402.

[15] K. Martin, K. Shilton, Putting mobile application privacy in context: An empirical study of user privacy expectations for mobile devices, Inf. Soc. 32 (3) (2016) 200–216, http://dx.doi.org/10.1080/01972243.2016.1153012.

[16] J.M. Such, J. Porter, S. Preibusch, A. Joinson, Photo privacy conflicts in social media: A large-scale empirical study, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 3821–3832, http://dx.doi.org/10.1145/3025453.3025668.

[17] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in: British Machine Vision Conference, 2015.

[18] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815–823, http://dx.doi.org/10.1109/CVPR.2015.7298682.

[19] Face recognition in low quality images: A survey, 2018, CoRR abs/1805.11519. URL http://arxiv.org/abs/1805.11519.

[20] P. Li, L. Prieto, D. Mery, P.J. Flynn, On low-resolution face recognition in the wild: Comparisons and new techniques, IEEE Trans. Inf. Forensics Secur. 14 (8) (2019) 2000–2012, http://dx.doi.org/10.1109/TIFS.2018.2890812.

[21] K. Hill, The secretive company that might end privacy as we know it, 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html, accessed: 2020-06-03.

[22] J. Kröckel, F. Bodendorf, Customer tracking and tracing data as a basis for service innovations at the point of sale, in: 2012 Annual SRII Global Conference, 2012, pp. 691–696, http://dx.doi.org/10.1109/SRII.2012.115.

[23] P.L. Venetianer, Z. Zhang, A. Scanlon, Y. Hu, A.J. Lipton, Video verification of point of sale transactions, in: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, 2007, pp. 411–416, http://dx.doi.org/10.1109/AVSS.2007.4425346.

[24] E. Barnoviciu, V. Ghenescu, S. Carata, M. Ghenescu, R. Mihaescu, M. Chindea, GDPR compliance in video surveillance and video processing application, in: International Conference on Speech Technology and Human-Computer Dialogue, 2019, pp. 1–6, http://dx.doi.org/10.1109/SPED.2019.8906553.

[25] J. Hiller, M. Schuldes, L. Eckstein, Recognition and pseudonymization of data Privacy Relevant Areas in videos for compliance with GDPR, in: IEEE Intelligent Transportation Systems Conference, 2019, pp. 2387–2393, http://dx.doi.org/10.1109/ITSC.2019.8917267.

[26] Y. Li, N. Vishwamitra, B.P. Knijnenburg, H. Hu, K. Caine, Blur vs. Block: Investigating the effectiveness of privacy-enhancing obfuscation for images, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1343–1351, http://dx.doi.org/10.1109/CVPRW.2017.176.

[27] Y. Li, N. Vishwamitra, B.P. Knijnenburg, H. Hu, K. Caine, Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos, Proc. ACM Hum.-Comput. Interact. 1 (2017) 1–24, http://dx.doi.org/10.1145/3134702.

[28] R. Hasan, E. Hassan, Y. Li, K. Caine, D.J. Crandall, R. Hoyle, A. Kapadia, Viewer experience of obscuring scene elements in photos to enhance privacy, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–13, http://dx.doi.org/10.1145/3173574.3173621.

[29] X. Liu, N. Krahnstoever, T. Yu, P. Tu, What are customers looking at? in: 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, 2007, pp. 405–410, http://dx.doi.org/10.1109/AVSS.2007.4425345.

[30] C. Neustaedter, S. Greenberg, M. Boyle, Blur filtration fails to preserve privacy for home-based video conferencing, ACM Trans. Comput.-Hum. Interact. 13 (1) (2006) 1–36, http://dx.doi.org/10.1145/1143518.1143519.

[31] P. Sinha, B. Balas, Y. Ostrovsky, R. Russell, Face recognition by humans: Nineteen results all computer vision researchers should know about, Proc. IEEE 94 (2006) 1948–1962, http://dx.doi.org/10.1109/JPROC.2006.884093.

[32] S. Menon, A. Damian, S. Hu, N. Ravi, C. Rudin, PULSE: self-supervised photo upsampling via latent space exploration of generative models, 2020, CoRR.

[33] R. McPherson, R. Shokri, V. Shmatikov, Defeating image obfuscation with deep learning, 2016, CoRR. arXiv:1609.00408.

[34] E.M. Newton, L. Sweeney, B. Malin, Preserving privacy by de-identifying face images, IEEE Trans. Knowl. Data Eng. 17 (2) (2005) 232–243, http://dx.doi.org/10.1109/TKDE.2005.32.

[35] W.L. Croft, J. Sack, W. Shi, Differentially private obfuscation of facial images, in: Machine Learning and Knowledge Extraction, 2019, pp. 229–249, http://dx.doi.org/10.1007/978-3-030-29726-8_15.

[36] L. Harmon, B. Julesz, Masking in visual recognition: Effects of two-dimensional filtered noise, Science 180 (4091) (1973) 1194–1197, http://dx.doi.org/10.1126/science.180.4091.1194.

[37] L. Harmon, The recognition of faces, Sci. Am. 229 (5) (1973) 71–82.

[38] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685, http://dx.doi.org/10.1109/34.927467.

[39] R. Gross, L. Sweeney, F. de la Torre, S. Baker, Model-based face de-identification, in: 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), 2006, p. 161, http://dx.doi.org/10.1109/CVPRW.2006.125.

[40] H. Hukkelås, R. Mester, F. Lindseth, DeepPrivacy: A generative adversarial network for face anonymization, in: Advances in Visual Computing, 2019, pp. 565–578, http://dx.doi.org/10.1007/978-3-030-33720-9_44.

[41] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, B. Schiele, A hybrid model for identity obfuscation by face replacement, in: Computer Vision – ECCV 2018, 2018, pp. 570–586, http://dx.doi.org/10.1007/978-3-030-01246-5_34.

[42] L. Fan, Practical image obfuscation with provable privacy, in: 2019 IEEE International Conference on Multimedia and Expo (ICME), 2019, pp. 784–789, http://dx.doi.org/10.1109/ICME.2019.00140.

[43] S.J. Oh, M. Fritz, B. Schiele, Adversarial image perturbation for privacy protection a game theory perspective, in: IEEE International Conference on Computer Vision, 2017, pp. 1491–1500, http://dx.doi.org/10.1109/ICCV.2017.165.

[44] N. Raval, A. Machanavajjhala, L.P. Cox, Protecting visual secrets using adversarial nets, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1329–1332, http://dx.doi.org/10.1109/CVPRW.2017.174.

[45] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recognit. 7 (2018) 354–377, http://dx.doi.org/10.1016/j.patcog.2017.10.013.

[46] A. Dosovitskiy, J.T. Springenberg, M. Tatarchenko, T. Brox, Learning to generate Chairs, Tables and Cars with Convolutional Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 692–705, http://dx.doi.org/10.1109/TPAMI.2016.2567384.

[47] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A.C. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems 27, 2014, pp. 2672–2680.

[48] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, in: 4th International Conference on Learning Representations, 2016.

[49] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, CoRR abs/1411.1784.

[50] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, 2016, CoRR abs/1605.09782.

[51] G. Perarnau, J. van de Weijer, B. Raducanu, J.M. Álvarez, Invertible conditional GANs for image editing, 2016, CoRR abs/1611.06355.

[52] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797, http://dx.doi.org/10.1109/CVPR.2018.00916.

[53] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, Attgan: Facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (11) (2019) 5464–5478, http://dx.doi.org/10.1109/TIP.2019.2916751.

[54] H. Chi, Y.H. Hu, Face de-identification using facial identity preserving features, in: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015, pp. 586–590, http://dx.doi.org/10.1109/GlobalSIP.2015.7418263.

[55] B. Meden, Z. Emersic, V. Struc, P. Peer, K-Same-Net: Neural-network-based face deidentification, in: 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI), 2017, pp. 1–7, http://dx.doi.org/10.1109/IWOBI.2017.7985521.

[56] T. Li, L. Lin, AnonymousNet: Natural face de-identification with measurable privacy, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[57] Q. Sun, L. Ma, S.J. Oh, L.V. Gool, B. Schiele, M. Fritz, Natural and effective obfuscation by head inpainting, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5050–5059, http://dx.doi.org/10.1109/CVPR.2018.00530.

[58] Z. Chen, T. Zhu, P. Xiong, C. Wang, W. Ren, Privacy preservation for image data: A GAN-based method, Int. J. Intell. Syst. 36 (4) (2021) 1668–1685, http://dx.doi.org/10.1002/int.22356.

[59] Y. Wu, F. Yang, Y. Xu, H. Ling, Privacy-protective-GAN for privacy preserving face de-identification, J. Comput. Sci. Tech. 34 (1) (2019) 47–60, http://dx.doi.org/10.1007/s11390-019-1898-8.

[60] P. Nousi, S. Papadopoulos, A. Tefas, I. Pitas, Deep autoencoders for attribute preserving dace de-identification, Signal Process., Image Commun. 81 (2020) 115699, http://dx.doi.org/10.1016/j.image.2019.115699.

[61] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S.K. Nayar, Face swapping: Automatically replacing faces in photographs, ACM Trans. Graph. 27 (3) (2008) 39:1–39:8, http://dx.doi.org/10.1145/1360612.1360638.

[62] Y. Li, S. Lyu, De-identification without losing faces, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019, pp. 83–88, http://dx.doi.org/10.1145/3335203.3335719.

[63] Y. Nirkin, I. Masi, A.T. Tran, T. Hassner, G.G. Medioni, On face segmentation, face swapping, and face perception, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, 2018, pp. 98–105, http://dx.doi.org/10.1109/FG.2018.00024.

[64] H. Hao, D. Güera, A.R. Reibman, E.J. Delp, A utility-preserving GAN for face obscuration, 2019, CoRR abs/1906.11979.

[65] C. Dwork, Differential privacy, in: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II, 2006, pp. 1–12, http://dx.doi.org/10.1007/11787006_1.

[66] K. Chatzikokolakis, M.E. Andrés, N.E. Bordenabe, C. Palamidessi, Broadening the scope of differential privacy using metrics, in: Privacy Enhancing Technologies, 2013, pp. 82–102, http://dx.doi.org/10.1007/978-3-642-39077-7_5.

[67] L. Fan, Image pixelization with differential privacy, in: Data and Applications Security and Privacy XXXII, 2018, pp. 148–162, http://dx.doi.org/10.1007/978-3-319-95729-6_10.

[68] W.L. Croft, J. Sack, W. Shi, Obfuscation of images via differential privacy: From facial images to general images, Peer Peer Netw. Appl. 14 (3) (2021) 1705–1733, http://dx.doi.org/10.1007/s12083-021-01091-9.

[69] B. Liu, M. Ding, H. Xue, T. Zhu, D. Ye, L. Song, W. Zhou, DP-image: Differential privacy for image data in feature space, 2021, CoRR. URL https://arxiv.org/abs/2103.07073.

[70] T. Li, C. Clifton, Differentially private imaging via latent space manipulation, 2021, CoRR. URL https://arxiv.org/abs/2103.05472.

[71] L. Xie, K. Lin, S. Wang, F. Wang, J. Zhou, Differentially private generative adversarial network, 2018, CoRR. arXiv:1802.06739.

[72] C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, K. Ren, GANobfuscator: Mitigating information leakage under GAN via differential privacy, IEEE Trans. Inf. Forensics Secur. 14 (9) (2019) 2358–2371, http://dx.doi.org/10.1109/TIFS.2019.2897874.

[73] J. Jordon, J. Yoon, M. Van Der Schaar, PATE-GAN: Generating synthetic data with differential privacy guarantees, in: 7th International Conference on Learning Representations, 2019.

[74] M. Flynn, Generating faces with deconvolution networks, 2016, URL https://github.com/zo7/deconvfaces.

[75] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 4401–4410, http://dx.doi.org/10.1109/CVPR.2019.00453.

[76] C. Dwork, A. Roth, The algorithmic foundations of differential privacy, Found. Trends Theor. Comput. Sci. 9 (2014) 211–407, http://dx.doi.org/10.1561/0400000042.

[77] J. Shlens, A tutorial on principal component analysis, 2014, CoRR. URL http://arxiv.org/abs/1404.1100.

[78] X. Jiang, Z. Ji, S. Wang, N. Mohammed, S. Cheng, L. Ohno-Machado, Differential-private data publishing through component analysis, Trans. Data Priv. 6 (1) (2013) 19–34.

[79] C. Dwork, K. Talwar, A. Thakurta, L. Zhang, Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis, in: Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing, 2014, pp. 11–20, http://dx.doi.org/10.1145/2591796.2591883.

[80] K. Chaudhuri, A.D. Sarwate, K. Sinha, Near-optimal differentially private principal components, in: Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems, 2012, pp. 998–1006.

[81] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: 18th Medical Image Computing and Computer-Assisted Intervention, Vol. 9351, 2015, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[82] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A.C. Courville, Improved training of wasserstein GANs, in: Advances in Neural Information Processing Systems 30, 2017, pp. 5767–5777.

[83] S. Sanyal, T. Bolkart, H. Feng, M.J. Black, Learning to regress 3D face shape and expression from an image without 3D supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7763–7772, http://dx.doi.org/10.1109/CVPR.2019.00795.

[84] K. He, G. Gkioxari, P. Dollár, R.B. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.

[85] C. Zheng, T. Cham, J. Cai, Pluralistic image completion, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1438–1447, http://dx.doi.org/10.1109/CVPR.2019.00153.

[86] M.E. Andrés, N.E. Bordenabe, K. Chatzikokolakis, C. Palamidessi, Geo-indistinguishability: Differential privacy for location-based systems, in: 2013 ACM SIGSAC Conference on Computer &#38; Communications Security, 2013, pp. 901–914, http://dx.doi.org/10.1145/2508859.2516735.

[87] C. Song, V. Shmatikov, Overlearning reveals sensitive attributes, in: 8th International Conference on Learning Representations, 2020.

[88] F. McSherry, K. Talwar, Mechanism design via differential privacy, in: 48th Annual IEEE Symposium on Foundations of Computer Science, Providence, USA, 2007, pp. 94–103, http://dx.doi.org/10.1109/FOCS.2007.66.

[89] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015, pp. 3730–3738, http://dx.doi.org/10.1109/ICCV.2015.425.

[90] LynnHo, HD CelebA cropper, 2018.

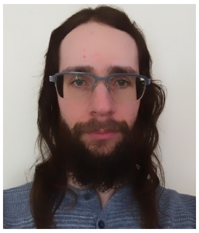
[91] Zhou Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[92] S.I. Serengil, Deepface, 2020, URL https://github.com/serengil/deepface.

[93] R. Rothe, R. Timofte, L.V. Gool, DEX: Deep expectation of apparent age from a single image, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2015, http://dx.doi.org/10.1109/ICCVW.2015.41.

[94] H. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, in: IEEE International Conference on Image Processing, 2014, pp. 343–347, http://dx.doi.org/10.1109/ICIP.2014.7025068.

[95] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.



William Croft is a Ph.D. graduate at Carleton University in the School of Computer Science. He completed his Ph.D. (2020), M.Sc. (2015) and B.Sc. (2013) in computer science at Carleton University. In 2017, he received the Natural Sciences and Engineering Research Council (NSERC) Alexander Graham Bell Canada Graduate Scholarship. His research focuses on the topic of data privacy.

Jörg-Rüdiger Sack is Chancellor's Professor at the School of Computer Science, Carleton University. He received a M.C.S. ("Diplom") degree from the University of Bonn, Germany, and a Ph.D. from McGill University, Montréal. His research interests include algorithms, data structures, geographic information systems and foremost computational geometry. He is editor-in-chief of the journal Computational Geometry-Theory and Applications, and serves on the Editorial Board of Journal of Spatial Information Science, and Computer Animation and Virtual Worlds. He is Chair of the Steering Committee of the Algorithms and Data Structures Symposium (WADS). He has been an industrial NSERC Chair in Applied Parallel Computing with a focus on spatial modelling. He served as Science Advisor to the German Embassy, Ottawa, to foster scientific collaboration between German and Canadian academic institutions and on the Excellence Initiative panels for the German Research Council (DFG) and the German Council of Science and Humanities. He chairs the Scientific Advisory Board of the Zuse Institute, the Scopus Content Selection & Advisory Board and is subject chair for Computer Science for Scopus.

Wei Shi is an associate professor in the School of Information Technology, Faculty of Engineering and Design at Carleton University. She is specialized in algorithm design and analysis in distributed systems, such as Distributed Data Center and Clouds, Edge Network, Mobile Agents and Actuator Systems and Wireless Sensor Networks. She has also been conducting research in data privacy and Big Data analytics. She holds a Bachelor of Computer Engineering from Harbin Institute of Technology in China and received her master's and Ph.D. of Computer Science degrees from Carleton University in Ottawa, Canada. Dr. Shi has published over 70 articles in reputable conferences and journals. She is also a Professional Engineer licenced in Ontario, Canada.