PrivacyGAN: robust generative image privacy

Mariia Zameshina, Marlene Careil, Olivier Teytaud, Laurent Najman

PrivacyGAN: robust generative image privacy

Mariia Zameshina
Univ Gustave Eiffel, CNRS, LIGM
F-77454 Marne-la-Vallee, France
mariia.zameshina@esiee.fr

Marlene Careil
LTCI, Telecom Paris,
Institut Polytechnique de Paris

Olivier Teytaud
TAO, CNRS - INRIA - LRI

Laurent Najman
Univ Gustave Eiffel, CNRS, LIGM
F-77454 Marne-la-Vallee, France

Abstract

Classical techniques for protecting facial image privacy typically fall into two categories: data-poisoning methods, exemplified by Fawkes, which introduce subtle perturbations to images, or anonymization methods that generate images resembling the original only in several characteristics, such as gender, ethnicity, or facial expression.

In this study, we introduce a novel approach, PrivacyGAN, that uses the power of image generation techniques, such as VQGAN and StyleGAN, to safeguard privacy while maintaining image usability, particularly for social media applications. Drawing inspiration from Fawkes, our method entails shifting the original image within the embedding space towards a decoy image.

We evaluate our approach using privacy metrics on traditional and novel facial image datasets. Additionally, we propose new criteria for evaluating the robustness of privacy-protection methods against unknown image recognition techniques, and we demonstrate that our approach is effective even in unknown embedding transfer scenarios. We also provide a human evaluation that further proves that the modified image preserves its utility as it remains recognisable as an image of the same person by friends and family.

1 Introduction

Individuals often share personal photos on various social media platforms, which facilitates communication and connection with family, friends, colleagues, and customers. Unfortunately, a significant drawback of this practice is that it can sometimes be possible to identify individuals social media accounts by taking their picture in public [22] or by comparing their dating app photos to their business-related social media profiles. This is often made possible by the existence of datasets collected by scraping social media platforms. While face detection systems can be used by the government for criminal identification purposes, they also present opportunities for both internal and external misuse [28], including enabling stalkers to track their victims [8]. Consequently, sharing real facial images publicly over the internet may compromise users privacy.

Multiple initiatives are dedicated to enhancing image and video privacy on the internet. One of the prominent groups is centred around **anonymization methods**, which involve altering users pictures to resemble those of other individuals. For instance, Kim et al. [17] proposed a privacy-preserving adversarial

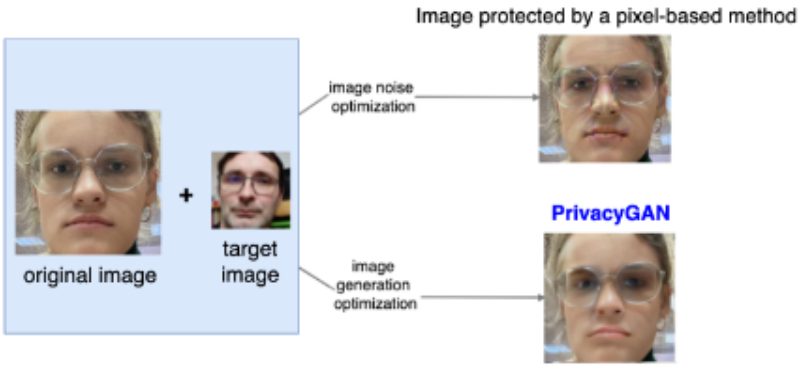

Figure 1: Schema for both data poisoning and generative privacy methods. We take the original image (OI) and create an image that is recognisable by human observers while being unlikely to be identified by image recognition methods using
(i) the classic approach by adding pixel noise such that it makes the modified image in the embedding space closer to the target image than to the original image,
(ii) (our approach, PrivacyGAN): generation of visually similar but distant images in the embedding space.

protector network (PPAPNet) as an image anonymization tool. PPAPNet transforms an image into another synthetic yet realistic image while remaining immune to model inversion attacks [27]. Anonymization techniques may also preserve key characteristics such as background, emotions, and facial feature movements [11] [14] [1] [15]. These techniques are valuable when the objective is to maintain realistic appearances without the need to recognise individuals in photos or videos. Such approaches are particularly useful, for example, for maintaining anonymity while expressing opinions on video-sharing platforms. For instance, [7] achieves the objective of decorrelating the identity while retaining the perception (pose, illumination, and expression). Some of these methods change only parts of the face. As an example, in their work [20], the authors suggest utilising generative techniques to enhance images that have been intentionally blurred or have had the subject's eyes obscured beforehand.

In the overview conducted by Wenger et al. [29], the authors tackle a challenge in the design of Anti-Facial Recognition (AFR) systems: finding a balance between **privacy, utility, and usability** . They categorise AFR systems based on their target components, ranging from data collection and model training to run-time inference, all with the shared objective of thwarting successful recognition by unauthorised or unwanted models. Moreover, the authors stress the user preference for privacy tools with minimal overhead, a concept underscored by studies such as Sharif et al. [25] and Dabouei et al. [4]. These findings highlight the significance of delivering protection against image recognition systems while mitigating any adverse effects on the user experience, a goal that many present anonymization methods struggle to attain. While certain attributes of images, such as gender, ethnicity, and facial expressions, can be retained through specific anonymization techniques, the resulting modified images frequently lack practicality for users. As a result, even though these images maintain crucial visual characteristics, the individual's identity within them may undergo substantial alterations, ultimately rendering them unidentifiable to acquaintances and family members.

In an effort to achieve a balance between utility and privacy protection, another area of research focuses on **altering or obscuring facial images to maintain human recognizability while creating difficulties**

for neural networks to decipher. Generally, these methods involve introducing precisely crafted pixel noise, causing the neural network to misclassify the image. These pixel-level perturbations have effectively challenged diverse image recognition neural networks. The dilemma of balancing privacy maintenance with recognition assurance of data-poisoning methods like Fawkes [24] and Lowkey [3] is discussed in detail in [21].

As an alternative approach to safeguarding images against unauthorised identification while preserving their utility for the users, one can consider **adversarial examples**. [20] demonstrated that makeup transfer can be an effective means of countering various face recognition systems. However, this method has limitations, as the model's performance may be inconsistent between male and female images due to an imbalance in the makeup transfer training dataset. This method is also ineffective in cases where the face cannot be found on an image, as it is not possible to transfer makeup in this case. Additionally, some individuals may find the use of makeup transfer images unacceptable.

In the present paper, we propose a general approach to the use of generative methods for privacy. We train our methods to be effective for embedding methods, on which pixel-based methods such as Fawkes used to fail. Our goal is to create images that resemble original photographs and are suitable for sharing on social media platforms, while also preventing identification by modern image recognition neural networks without using anonymization. To achieve this, we explore the effectiveness of two generative methods: a generative adversarial neural network StyleGAN [16], and the autoencoder VQGAN [6]. These generative methods are known for their realistic image generation capabilities, which adds an added layer of difficulty for neural network recognition.

By modifying facial images using generative methods, we aim to preserve their recognizability to human observers while rendering them unrecognisable to many existing image recognition neural networks. Inspired by pixel-based methods like Fawkes [24], we propose modifying the generated "private" images towards a different target image in the embedding space and evaluate the robustness of our approach against unknown image recognition neural networks. We validate our privacy methods on the Labeled Faces in the Wild (LFW) dataset [13], as well as introduce a new dataset of face crops extracted from Casual Conversations [9] to ensure their effectiveness in various environments. In Fig. 1 we present the schema of image modification using both pixel-based and generative methods.

Our proposed generative tools make subtle modifications to user images without adding pixel noise, so the resulting photos look natural and protect user privacy. Our algorithms operate within a black-box framework and demonstrate their efficacy against image recognition techniques they were not specifically trained on. We offer flexibility in selecting methods and privacy settings and conduct a comparison between our approach and existing state-of-the-art privacy protection methods, such as Fawkes.

To sum up the claims of the present paper, we

1. propose a novel approach to facial image privacy based on generative methods;
2. create a new privacy evaluation approach based on the percentage of dataset images that are closer in an embedding space to a modified "private" image than to an original image;
3. propose a new facial image dataset extracted from the Casual Conversations dataset [9] videos;
4. evaluate the privacy of the modified images against various embedding methods (including transfer to embeddings not used in our privacy method) and provide human evaluation of image quality for state-of-the-art and novel privacy methods.

2 Privacy Algorithms

2.1 A well-known pixel-based method: Fawkes.

Fawkes [24] is a data poisoning method that presents subtle image perturbations to the images. One of its main features is the concept of target image: by suggesting elements from a side target image, Fawkes ensures that the modified image will be recognised as another person by neural networks, thus ensuring privacy. Unlike just maximising the distance between the embedding of the original image and the modified image, this method

1. helps to keep the embedding of modified images within a valid range for a given dataset and
2. ensures that the embedding of the modified image does not stay close to its original version in the given dataset.

The idea of Fawkes is to pair each original image (**OI**) with a target image (**TI**). Then Fawkes associates to each original image a 'cloak image' (**CI**) which consists of noise obtained by optimising the following loss:

$$L_{Fawkes} = ||emb(TI) - emb(OI \oplus CI)||,$$

where: **OI** is an original image; $\oplus$ is capped addition; emb is an embedding method used for cloak optimisation in order to obtain a modified "private" version of an original image; **TI** is a target image; the private version of OI should be labelled the same as TI by the chosen image recognition system; ρ is a parameter that caps the noise strength; **CI** $< \rho$ is a cloak or a noise that should be added to OI in order for it ensure its' privacy; $OI \oplus CI$ is the published rendition of OI. In our experiments, by default, we use the "high" mode of Fawkes (as mentioned in [30] and [24]), since it provides a decent level of protection, and it is possible to compare this setting of Fawkes with our methods.

2.2 Our proposed generative methods based on VQGAN and StyleGAN

Our proposal involves utilising generative models for privacy protection, with a focus on generating an image that closely resembles the original in visual appearance while safeguarding users against image recognition attacks. Our objective is not to anonymize the image. Another generative method that transfers makeup in order to protect facial privacy (AMT-GAN) is reviewed in the supplementary material. In this paper, we expand on the idea of target images introduced in Fawkes. We use target images to ensure that the modified version of an image is closer to the target image than the original image in the chosen embedding space. To select target images, we choose from images in the dataset that have not been used in experiments. For each specific image, we select a target image based on its distance from the original image in the chosen embedding method used for optimization. The chosen target image should be far enough from the original image to ensure effective privacy protection.

We select the loss function L based on the goal of preserving the identity of the original image (OI) for humans while ensuring that the generated image (GI) embedding is as close as possible to the distant target image embedding. For this purpose, we use the Learned Perceptual Image Patch Similarity (LPIPS) distance for the preservation of OI identity and optimise the embedding distance between the generated image (GI) and the target image (TI) to achieve the closest possible embedding.

The loss for any generated image always consists of the sum of the following parts:

1. LPIPS distance between generated and original image

2. For each of the embeddings used for optimisation, coefficient K multiplied by the mean squared distance between the modified "private" image embedding and target image embedding.

$$L_{\substack{generative \\ privacy}} = LPIPS(OI, GI) + \\ + K \times \sum_{emb \in\, embeddings} || \, emb(GI) - emb(TI)||.$$

The hyperparameters of the loss described are the coefficient K (*i.e.*, weight compared to LPIPS) for the embedding distance, the learning rate, the batch size, and the number of iterations.

2.3 Generative Privacy Algorithm: PrivacyGAN

Here, we describe PrivacyGAN, an algorithm that we propose for creating private versions of facial images.

Algorithm 1 PrivacyGAN: Private image generation algorithm

Require: OI: Original Image
Ensure: GI: Generated Image
  $TI \leftarrow$ Target Image ▷ (distant from OI in the embedding space)
  $z \leftarrow$ random
  $G \leftarrow$ image generation method
  $K \leftarrow$ optimisation coefficient
  $chosen\_embeddings \leftarrow$ list of embedding methods for optimisation ▷ (we distance GI from OI in these embedding spaces)
  for i in range(0, num_iterations) **do**
    $GI \leftarrow G(z)$
    $emb\_dist \leftarrow 0$
    for emb in $chosen\_embeddings$ **do**
      $emb\_dist \; += ||emb(GI) - emb(TI)||$
    $lpips\_dist \leftarrow LPIPS(GI, OI)$
    $loss \leftarrow lpips\_dist + K \cdot emb\_dist$
    $z \leftarrow update(z, loss, \nabla_{\text{loss}})$
  return GI

As input to the algorithm, we use an original image **OI** and a target image **TI**. The algorithm aims to produce an image **GI** which would be a "private" version of OI, unrecognisable by many image recognition neural networks. In order to do that, we are using generative methods such as StyleGAN and VQGAN. TI is chosen randomly among the images furthest in embedding space from OI.

The algorithm consists of an iterative optimisation process, where **num_iterations** represents the number of iterations and **G** is the image generation method. In the latent space of G, we find a latent variable **z** and generate an image G(z), which we refer to as GI. For each of the embedding methods in the set **chosen_embeddings**, in each iteration of the algorithm, we compute the embedding distance between GI and TI, as well as the LPIPS distance [31] between images GI and OI. We use the computed distances to calculate the $L_{\substack{generative \\ privacy}}$ that we mention as 'loss' in the algorithm.

3 Evaluation Methods

3.1 Metrics for Privacy

In order to evaluate the privacy of generated images, it is important to determine how far the generated image is from the original in the dataset. After applying an image recognition neural network, attackers may choose to verify if the person on the image matches the top few possible results. That is why the method would work better for privacy protection if: i) the modified image would not be recognised as its original version; and ii) the original image would be far away from the modified one in the embedding space.

We ensure ii) not only by using existing evaluation methods such as Recall@k that help us to make sure that the modified "private" image is far from the original in absolute values but also by introducing a novel evaluation method that verifies the original and modified image being far in embedding space relative to the dataset size.

To measure the distance from the original image to its modified private version, we use the following privacy metrics: Recall@k and $Percentage$.

Recall@k for the set of query images L (which can either be original or modified images) and test images M, is defined as

$$Recall(L, M, k) = 100 \frac{\sum\limits_{q \in L} 1_{Id(q) \in Id(N(q,k,M))}}{||L||},$$

where function Id maps the set of people's images to the set of (unique) identities of the individuals present on these images, function $N(q, k, M)$ returns a set of k images from M that have the closest embedding to the one of the query image q.

We propose the use of a new metric, called the "Percentage", in addition to the Recall metric, to evaluate the effectiveness of our privacy methods. The reason for introducing this new metric is to ensure that the modified image is not only far from the original image in absolute terms, as ensured by Recall@k, but also in terms of the percentage of dataset size. This provides a common privacy metric that can be used to compare the effectiveness of our methods across different dataset sizes.

$Percentage$ is the proportion of images for each query image from the dataset L in between the query image and the closest image with the same identity from the dataset M:

$$Percentage(L, M, k) = 100 \sum_{q \in L} \frac{Between(q, N(q, 1, M))}{||L|| \times ||M||},$$

where the function $Between(q_1, q_2)$ returns the number of images in the dataset M that have a smaller distance to the embedding of q_1 than the distance in-between the embeddings of q_1 and q_2.

3.2 The problem of transfer

In practical scenarios, it is crucial that privacy methods are effective against various image recognition neural networks. We optimise our privacy methods to be effective for specific embeddings, and transferring to a different embedding method can be challenging as new methods are continually emerging. It is impossible to guarantee that privacy methods will be effective against future attacks, as some methods have been broken by newer recognition neural networks [21]. To evaluate the effectiveness of our proposed methods, we conducted two sets of optimisation experiments.

The first experiment involves optimising StyleGAN and VQGAN image generation to be effective against the FaceNet embedding method. We aim to make the FaceNet embedding of the generated image

	PrivacyGAN				Pixel-based
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes
Percentage	8.110	0.654	**14.696**	0.861	0.782
Recall@1:m.i.	1.754	22.085	**0.047**	19.242	20.521
Recall@1:o.i.	2.180	22.133	**0.332**	22.180	23.886
Recall@3: m.i.	6.114	61.564	**0.758**	54.597	56.398
Recall@3: o.i.	5.308	60.142	**0.900**	53.981	58.246
Recall@5: m.i.	8.815	77.678	**1.374**	70.711	74.502
Recall@5: o.i.	7.109	75.782	**1.327**	67.820	72.796
Recall@10: m.i.	13.365	86.256	**2.986**	79.668	83.412
Recall@10: o.i.	11.422	85.355	**2.512**	77.773	82.938
Recall@50: m.i.	32.417	94.408	**11.280**	92.227	92.986
Recall@50: o.i.	28.768	94.028	**10**	92.464	93.981
Recall@100: m.i.	43.697	96.303	**19.336**	95.166	95.592
Recall@100: o.i.	40.806	96.019	**17.062**	95.071	96.303

Table 1: Test on the LFW dataset. Evaluation for the same embedding that was used for training (no transfer): PrivacyGAN (based on VQGAN or StyleGAN) is optimised with FaceNet, tested with FaceNet, and compared to Fawkes in "high" mode (meaning: high privacy). We see that PrivacyGAN equipped with standard versions of StyleGAN and VQGAN obtains better privacy results compared to Fawkes.

distant from that of the original image during the optimisation process. We compare our proposed methods to Fawkes, which uses the same embedding method for optimisation, in Table 1. Additionally, we aim to test the transferability of our methods to embedding methods introduced after FaceNet, which Fawkes does not prove to be effective against [21].

The second experiment involves optimising StyleGAN and VQGAN image generation using MagFace and MobileFaceNet embedding methods, which have been shown to increase the robustness of generated images. We also compare them to the makeup transfer method AMT-GAN [12]

4 Datasets and embeddings

4.1 The Labelled Faces in the Wild

The dataset [13] contains multiple images for each person, with the number of images per person varying between 1 and 530. To ensure fairness, we extract a sub-dataset from the original dataset, which includes 5 randomly chosen images per person. We exclude images of people who have less than 5 photos present in the dataset. This sub-dataset is referred to as LFW in the following sections of this paper.

4.2 The Casual Conversations dataset

The Casual Conversations dataset [9] comprises 45186 videos, each of which features one person. We select 997 videos and extract 5 face crops of size 456×456 per person present in the dataset (in case the video contains face crops of a required size).

The process of selecting these face crops is as follows:

1. We select all the time frames from the video featuring a specific person.
2. We check if, among these time frames, there are at least 5 non-consecutive (± 10) time frames that satisfy the following conditions:
   - they contain a face crop of a size at least 456×456 with a margin of size 100;
   - average brightness of a time frame is at least 70. This condition is required since, among the videos in the CC dataset, there are many that were recorded in complete darkness, and it is not realistic to have such face crops as profile pictures.
3. If there are more than 5 time frames selected, we randomly choose 5 of them and add them to the dataset.

We make sure that we don't select successive frames of the video since they could contain identical face crops.

For the confounders set, we randomly choose different people's face crops that also satisfy conditions 1 and 2 and do not feature a person who was already selected for our primary dataset before.

The key difference between our novel dataset and LFW is that faces in our proposed dataset have similar backgrounds and are taken within a short timeframe, creating an additional challenge for privacy protection. That lack of variety makes this particular dataset very interesting for our research. By testing our methods on it, we are able to ensure that, even if there are many very similar photos of the same person in the dataset, the proposed privacy tools can still be effective. It is particularly important in cases where people publish their images from similar locations on different platforms over the internet. Later, we refer to this dataset as **CC**.

The complete code for the face crop dataset extraction will be provided at the time of publication.

4.3 Our proposed methods for transfer to unknown embeddings: optimising on multiple embeddings

The embedding methods that we are using in this paper are the following: FaceNet [23]; ArcFace [5]; SphereFace [18]; MagFace [19]; MobileFaceNet [2] with implementation from the FaceX-Zoo library [26]; and ResNet_152 [10] with implementation from the FaceX-Zoo library [26].

We evaluate the effectiveness of our proposed privacy methods in a black-box setting (*i.e.*, robustness to unknown image recognition methods not used in the privacy method). We optimise the generated image for one or two embeddings from the list and then check the generated image against all the other embeddings. Thus, we make sure that our privacy methods transfer well to unknown embeddings and can be used for the privacy protection of real photos published online.

5 Experiments and Results

Here we define the settings and notations that we use in our experiments.

We set the hyperparameters of the generative privacy loss ($L_{\substack{generative \\ privacy}}$) for our experiments in the following way: the learning rate to 0.01 and the batch size to 32. The only parameters that we modify from experiment to experiment are the coefficient K and the number of iterations.

We have introduced the notations "o.i" and "m.i" to represent the original image and the modified image generated by any privacy-preserving algorithm, respectively. For our recall evaluation, we select either the original image (o.i. context) or the modified image (m.i. context) as the query image, where the dataset used for recognition includes modified images and confounders for the former and includes the original images and confounders for the latter.

In all metric calculations, we also use a set of confounders, which are not used as queries in our experiments and are sourced from the same dataset as the original images. The number of confounders is always less than or equal to $\frac{1}{5}$ of the number of original images. To compare privacy protection methods, we use the average value of the transfer recall (*i.e.*, Recall@10) for all embeddings for which the algorithm was not optimized. Incorporating confounders in our experiments brings us closer to real-world scenarios, where datasets may contain unrelated images that can potentially affect the experiment results.

Moreover, in order to compare VQGAN, StyleGAN, and Fawkes to one another, we use different sets of parameters chosen to match the transfer recall results. Thus, we are able to compare the generated image quality and see which of the methods generates the best images in terms of image quality for a given privacy performance (measured by recall).

Later in this section, we use the following notations: By **standard version of StyleGAN** we mean PrivacyGAN equipped with StyleGAN optimized with a coefficient $K = 0.03$ for embedding distance in the loss and 128 iterations; By **standard version of VQGAN** we mean PrivacyGAN equipped with VQGAN optimized a coefficient $K = 0.03$ for embedding distance in the loss and 1000 iterations; By StyleGAN_x_y / VQGAN_x_y we mean PrivacyGAN equipped with StyleGAN/VQGAN optimized a coefficient $K = x$ for embedding distance in the loss and y iterations.

5.1 Experiment 1: Comparing Pixel-Based and Generative Methods Optimised for One Embedding on the LFW Dataset

In Table 1, we compare standard versions of VQGAN and StyleGAN and their versions StyleGAN_0.003_500 and VQGAN_0.005_128 that we prepare specifically to match the privacy results of Fawkes. With this parametrization, they have the same transfer recall score, which allows us to compare fairly the image quality of our proposed generative methods and the state-of-the-art pixel-based method Fawkes. The transfer recall values (average values of Recall@10 for other methods from the list) are 62.16% for StyleGAN, 89.28% for StyleGAN_0.003_500, 65.49% for VQGAN, 90.07% for VQGAN_0.005_128, 90.90% for Fawkes. In order to make sure that image privacy is robust against various facial recognition systems, we study a transfer to different embedding methods (ArcFace, MagFace, SphereFace, MobileFaceNet, ResNet_152). Some results are in Table 2 (SphereFace) and in Table 3 (MagFace).

More results can be found in the supplementary material.

Examples of original images from the LFW dataset and their modifications obtained by our methods and by Fawkes are presented in Fig. 2. More examples are in the supplementary material.

Overall, from Tables 1, 2 and 3, we note that with a standard set of parameters, VQGAN and StyleGAN are much better for privacy than Fawkes. In order to match Fawkes privacy results, we need to change the VQGAN and StyleGAN parameters tenfold. While these parameter changes decrease privacy significantly, generative methods still have a disruptive effect on image quality.

It is worth noting that optimising generative methods using a single embedding method is insufficient for adequate facial image privacy protection. In the case of transferring images to MagFace (Table 3), the correct identity for the modified image is often among the top 5 possibilities. Thus, in the next subsection, we use two different embedding methods in the optimisation process to generate private image versions. Furthermore, we have observed that combining Fawkes poisoning with our proposed methods can be advantageous

	PrivacyGAN				Pixel-based
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes
Percentage	**5.181**	1.388	5.104	1.271	1.213
Recall@1: m.i.	9.526	21.896	**8.768**	21.043	21.611
Recall@1: o.i.	9.431	22.891	**8.578**	23.223	23.791
Recall@3: m.i.	21.422	53.744	**20.142**	54.929	55.877
Recall@3: o.i.	19.621	53.744	**17.678**	54.360	56.588
Recall@5: m.i.	27.915	67.109	**26.066**	68.294	69.668
Recall@5: o.i.	24.834	66.682	**23.128**	66.777	69.100
Recall@10: m.i.	35.261	74.739	**33.175**	76.161	77.014
Recall@10: o.i.	32.322	75.308	**30.521**	74.455	77.393
Recall@50: m.i.	55.592	86.493	**53.602**	87.867	87.536
Recall@50: o.i.	53.507	87.773	**51.422**	86.967	88.673
Recall@100: m.i.	65.308	91.232	**63.697**	91.469	91.754
Recall@100: o.i.	63.744	91.896	**61.327**	91.611	91.943

Table 2: Evaluation in the case of transfer to another embedding on the LFW dataset: PrivacyGAN (with VQGAN or StyleGAN) are optimised with FaceNet and tested with SphereFace. Generative methods do obtain better privacy results than Fawkes, except for the versions specifically created (weakened) to have privacy results similar to Fawkes (these versions are created for comparing image quality in Table 7 in a context with equal privacy performance).

for facial image privacy protection. We expand on that in supplementary material.

5.2 Experiment 2: Comparing StyleGAN and VQGAN Optimised with 2 Embedding Methods on the LFW Dataset

We now compare standard versions of VQGAN and StyleGAN together with other specific versions:

1. StyleGAN_0.02_500 and VQGAN_0.04_128;
2. StyleGAN_0.02_1000 and VQGAN_0.03_512.

These versions are proposed so that they have similar transfer recall scores in each pair. Specifically, average transfer recall scores for different methods are: 20.12% for StyleGAN, 32.33% for StyleGAN_0.02_500, 36.23% for StyleGAN_0.02_1000, 42.41% for VQGAN, 36.99% for VQGAN_0.03_512 and 33.07% for VQGAN_0.04_128. We create these specific versions of VQGAN and StyleGAN so that we can fairly compare the quality of the generated private images produced by the generative methods. We want to know which method produces the best image quality for a given threshold of our privacy metric. An example of an evaluation result without transfer is presented in Table 4 and for MobileFaceNet in the supplementary material.

We also study the transfer to different embedding methods. One of the results of this study (for the embedding method SphereFace) is presented in Table 5. Transfer results for other embedding methods can be found in the supplementary material.

	PrivacyGAN				Pixel-based
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes
Percentage	**0.767**	0.361	0.627	0.422	0.408
Recall@1: m.i.	**20.758**	24.028	24.028	24.265	24.882
Recall@1: o.i.	**21.611**	25.972	22.701	25.545	25.403
Recall@3: m.i.	**60.237**	71.848	66.493	73.744	75.403
Recall@3: o.i.	**60.142**	73.602	65.261	73.507	73.507
Recall@5: m.i.	**78.294**	96.967	86.209	98.389	98.768
Recall@5: o.i.	**76.777**	97.156	83.744	98.436	98.863
Recall@10: m.i.	**85.687**	97.962	91.043	98.863	99.194
Recall@10: o.i.	**84.313**	98.152	89.431	98.910	99.005
Recall@50: m.i.	**93.223**	99.005	95.545	99.005	99.194
Recall@50: o.i.	**92.512**	98.957	95.450	99.052	99.194
Recall@100: m.i.	**95.308**	99.052	97.062	99.100	99.194
Recall@100: o.i.	**94.976**	99.052	96.825	99.147	99.194

Table 3: Evaluation on the LFW dataset in the case of transfer to another embedding: PrivacyGAN (with VQGAN or StyleGAN) is optimised with FaceNet and tested with MagFace. Generative methods do obtain better privacy results than Fawkes, except for the versions specifically created (weakened) to have privacy results similar to Fawkes (these versions are created for comparing image quality in Table 7 in a context with equal privacy performance). However, both Fawkes and generative methods optimised with one embedding do not transfer well to the novel embedding methods such as MagFace, while they transfer better to some other embedding methods such as SphereFace, as in Table 2.

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512	VQGAN _0.04_128
Percentage	**15.049**	7.909	7.264	4.910	7.472	8.307
Recall@1: m.i.	**0.095**	0.806	0.900	0.521	0.284	**0.095**
Recall@1: o.i.	**3.555**	8.768	10.521	11.185	6.588	6.919
Recall@3: m.i.	**3.270**	10.711	11.327	15.071	9.147	8.057
Recall@3: o.i.	**6.493**	17.583	20.332	23.981	14.360	14.218
Recall@5: m.i.	**5.118**	16.919	19.479	26.682	17.583	14.834
Recall@5: o.i.	**8.863**	21.943	25.071	30.664	19.147	17.867
Recall@10: m.i.	**9.242**	25.261	28.057	37.488	24.787	22.180
Recall@10: o.i.	**12.275**	28.626	32.133	40	25.972	24.929
Recall@50: m.i.	**23.649**	44.550	46.919	57.678	44.692	42.085
Recall@50: o.i.	**26.114**	48.436	50.521	60.806	46.967	44.028
Recall@100: m.i.	**31.706**	53.555	57.204	66.730	54.597	51.991
Recall@100: o.i.	**33.649**	57.393	60.332	69.052	55.450	54.028

Table 4: Evaluation of various PrivacyGAN variants on the LFW dataset, case without transfer: PrivacyGAN (equipped with VQGAN and StyleGAN, including variants) are optimised with MagFace and MobileFaceNet and tested with MagFace. Lower recall means better privacy. Compared to Fawkes, results in Table 3: generative methods do get better privacy results. However, we did use MagFace in the algorithm, whereas Fawkes does not, hence the need for further validation (*i.e.*, testing in the case of transfer to embeddings not used in the privacy algorithm), which is done, for example, in Table 5.

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512
Percentage	**12.273**	8.157	7.715	6.211	7.387
Recall@1: m.i.	**2.749**	5.118	5.735	6.682	4.787
Recall@1: o.i.	**5.261**	7.536	9.431	10.047	7.204
Recall@3: m.i.	**6.256**	11.801	12.701	14.408	12.512
Recall@3: o.i.	**9.573**	14.218	18.199	17.630	14.408
Recall@5: m.i.	**8.578**	15.924	17.109	19.289	17.536
Recall@5: o.i.	**11.848**	17.725	22.227	22.464	19.005
Recall@10: m.i.	**13.033**	20.711	22.891	26.398	24.360
Recall@10: o.i.	**15.877**	24.218	27.441	29.858	26.303
Recall@50: m.i.	**26.209**	39.858	41.943	47.441	42.749
Recall@50: o.i.	**30.379**	43.554	45.403	51.659	46.161
Recall@100: m.i.	**35.024**	49.336	50.995	57.820	52.749
Recall@100: o.i.	**39.716**	52.654	55.024	61.754	56.256

Table 5: Evaluation of various PrivacyGAN variants in the case of transfer to another embedding on the LFW dataset: PrivacyGAN equipped with VQGAN or StyleGAN is optimised with MagFace and MobileFaceNet and tested with SphereFace. In this case, generative methods optimised with two embeddings obtain better results than generative methods optimised with only one embedding method in Table 2.

From the table 5 compared to 2 we can see that, in general, generative methods optimised with two embeddings transfer better to other embedding methods than generative methods optimised with only one embedding. For instance, for the unused in an optimisation process embedding SphereFace, the percentage score for StyleGAN with standard parameters optimised with 2 different embedding methods is more than 15 while it was just around 5 for one embedding method. Examples of images produced by the methods of experiment 2 are presented in Fig. 3. More of the examples can be found in the supplementary material.

		PrivacyGAN		Pixel-based	PrivacyGAN	Adversarial
	VQGAN	VQGAN	VQGAN	Fawkes	StyleGAN	AMT-GAN
	_0.003_128		_0.04_4096		_0.02_1000	
Percentage	9.424	13.399	16.379	7.519	**17.024**	14.067
Recall@1: m.i.	17.854	9.529	5.998	23.952	**5.416**	9.328
Recall@1: o.i.	18.034	11.214	**6.800**	24.092	6.841	8.445
Recall@3: m.i.	40.702	21.364	13.561	52.979	**12.197**	19.980
Recall@3: o.i.	41.765	23.149	14.483	53.420	**14.443**	17.593
Recall@5: m.i.	48.245	26.439	17.051	61.244	**15.727**	24.293
Recall@5: o.i.	49.629	27.924	**17.994**	61.224	18.134	21.344
Recall@10: m.i.	53.220	31.515	21.143	65.055	**20.100**	29.228
Recall@10: o.i.	54.945	33.621	**22.768**	65.135	23.531	25.436
Recall@50: m.i.	64.253	44.835	34.343	72.979	**32.618**	42.086
Recall@50: o.i.	65.537	47.442	**36.068**	72.738	36.911	36.409
Recall@100: m.i.	68.726	51.675	41.484	76.108	**39.478**	49.509
Recall@100: o.i.	70.030	54.363	44.152	75.928	45.176	**42.467**

Table 6: Evaluation in the case of a transfer to another embedding on the CC dataset: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet and tested with SphereFace. PrivacyGAN basically outperforms Fawkes while the comparison with AMT-GAN (which could be used on top of our method) depends on criteria and parameters.

5.3 Experiment 3: Comparing StyleGAN, VQGAN, and Fawkes on the CC dataset

Here we choose the specific versions of VQGAN, StyleGAN, and Fawkes that have similar transfer recall scores in each group for the dataset CC:

1. VQGAN_0.003_128 and Fawkes;
2. StyleGAN_0.02_1000 and VQGAN_0.04_4096.

In addition, we have compared our results to those obtained with AMT-GAN. However, it is important to note that a direct comparison between our proposed methods and AMT-GAN is not possible, as AMT-GAN is unable to transfer makeup to faces that were not detected. Therefore, in cases where faces were not detected, we had to replace them with the original images, which may affect the comparability of the results. Transfer recalls for the proposed methods are the following: AMT-GAN: 49.57%, Fawkes: 79.21%, StyleGAN_0.02_1000: 24.71%, VQGAN_0.003_128: 74.76%, VQGAN_0.04_4096: 26.09%, VQGAN: 40.45%. We compare how well proposed generative and pixel-based approaches protect privacy against different embedding methods. One of the results of this study (for the embedding method SphereFace) is presented in

Table 6. Transfer results for other embedding methods can be found in the supplementary material. Examples of images produced by the methods of experiment 3 are presented in Fig. 4. More examples can be found in the supplementary material.

Using table 6, we can conclude that, despite using the CC dataset instead of LFW, generative methods prove to be effective for privacy preservation and tend to outperform both the pixel-based method Fawkes and the generative makeup transfer method AMT-GAN.

5.4 Human preferences for similar transfer recall

In Figs 4 and 3, we can see that, in some cases, modifying the number of iterations in optimisation and the coefficient K affects the quality of an image and its privacy protection. Therefore, to evaluate the modified image quality for different privacy methods with similar transfer recall, we conducted a human preference study. Three human raters were presented with 40 pairs of images generated by the methods discussed in section 5.3. Given two images generated by two different methods, the human rater could choose "I prefer the left one as an avatar," "I prefer the right one as an avatar," or "No preference." To provide context, the human raters involved in the experiment were not paid and were not authors of the current paper. They were selected using the snowball principle, and their task was to assess the quality and similarity of the modified image to its original version. Without this assessment, we could end up with a black square instead of a privacy-protected image. The human raters did not have a degree in computer science and were not informed that the experiment related to privacy. However, the instructions provided to them, which included presenting the original image at the centre and emphasising that all images were reasonably close to it, as well as providing examples of potential use cases such as social networks, news articles, and dating websites, made it clear that assessing image similarity was an integral part of the task. The human-assessment results are presented in Table 7.

Transfer recall	Avatar method 1	Avatar method 2	Human preference: success rate of 1 vs 2
High privacy (low recall), LFW dataset			
32.6%	(StyleGAN_0.02_500, MagFace + MobileFaceNet)	**(VQGAN_128_0.04, MagFace + MobileFaceNet)**	43.75±5%
36.7%	(StyleGAN_0.02 1000, MagFace + MobileFaceNet)	**(VQGAN_0.03_512, MagFace + MobileFaceNet)**	35.37% ± 5%
Low privacy (high recall), LFW dataset			
90%	**(StyleGAN_0.003_500 FaceNet)**	Fawkes	87.2% ± 2%
High privacy, (low recall), CC dataset			
26.09% - 24.71%	**(VQGAN_0.04_4096, MagFace + MobileFaceNet)**	(StyleGAN_0.02_1000, MagFace + MobileFaceNet)	55.2 % ± 3.59%
Low privacy, (high recall), CC dataset			
74.76% - 79.21%	**(VQGAN_0.003_128) MagFace + MobileFaceNet)**	Fawkes	51.5 % ± 3.06%

Table 7: We modify the strength of different privacy-protection-algorithm perturbations until we get to similar target recall levels. We compare the quality of images, for each recall level. Text in bold font refers to human preference, for each recall level (see rightmost column).

We compared 5 different pairs of privacy-preserving methods with similar target recall values. In the low

privacy (high transfer recall) setting for both LFW and CC datasets, we were able to compare, in terms of quality, the Fawkes method with generative methods specifically modified to match Fawkes transfer recall values. When we choose FaceNet as an embedding for generative methods (StyleGAN) optimisation, the quality of the images generated by StyleGAN appears to be worse than that of Fawkes. However, when we use MagFace and MobileFaceNet as embeddings for generative methods optimisation, we obtain similar image quality results for VQGAN and Fawkes ($51.5\% \pm 3.06\%$), while VQGAN has a better transfer recall (74.76% compared to 79.21%). In the high privacy (low transfer recall) setting for both the LFW and CC datasets, we were able to compare different modifications of the generative methods (VQGAN and StyleGAN) with similar transfer recall. In every case, it appears that human raters preferred VQGAN-generated images over StyleGAN-generated images.

5.5 Human Identification of Same-Person Images

In this section, our objective is to evaluate the effectiveness of various privacy preservation methods for their applicability on social media platforms and to determine the extent to which people can identify the person after privacy-preservation modifications by different methods, namely: VQGAN_0.005_128, AMT-GAN, and Fawkes, and the anonymization method Deep Privacy 2 [15].

The experiment is structured as follows: we begin with the original image, referred to as the "original" in the filename. We then examine privacy-preserved versions of different images of the same individual, followed by random images of different people. The central question for each image in a given set is, "Is this the same person as in the original?".

To execute this experiment, we established a setup illustrated in Figure 5. This schema visually represents the process, illustrating the different image types involved in the human identification experiment.

We recruited 5 human evaluators, aged from 15 to 43 years, to assess pairs of images, comprising the original image and constructed using privacy-preserving methods for the same individual, as outlined in the schema in Figure 5. The results of the human study are presented in Table 8.

Model	Accuracy for humans	99% confidence interval
PrivacyGAN using VQGAN_0.005_128	0.796	$\pm$ 0.04
AMT-GAN	0.829	$\pm$ 0.038
Fawkes	0.842	$\pm$ 0.037
Deep Privacy 2	0.187	$\pm$ 0.039

Table 8: Human face identification for various privacy preservation models. The table demonstrates that, while our method VQGAN_0.005_128, together with AMT-GAN and Fawkes, generate recognisable images, the anonymization method Deep Privacy 2 often produces images that cannot be recognised as the same person. The purpose and limitations of this experiment are further discussed in the text.

From the findings presented in Table 8, we deduce that VQGAN_0.005_128, AMT-GAN, and Fawkes generate images that are similarly identifiable by human evaluators, with approximately 80% of images generated by these methods being successfully recognised as the same as the original. Conversely, the anonymization method Deep Privacy 2 frequently produces images that cannot be identified as those of the same individual. This further substantiates that while anonymization methods such as Deep Privacy 2 preserve certain attributes of images, they may fail to preserve their utility. In contrast, our method, along with other utility-focused methods, while not providing absolute protection against facial recognition, effectively

safeguards human facial images against prevalent face recognition techniques and maintains image utility for social media use.

This experiment was designed specifically to showcase that the recognisability of our method, along with AMT-GAN and Fawkes, is higher than that of anonymisation methods such as Deep Privacy 2. This explains our focus on testing just one version of PrivacyGAN. In future experiments, we intend to incorporate multiple versions of PrivacyGAN, and to involve a larger pool of human raters. This approach aims to determine which versions of PrivacyGAN perform best in terms of privacy/recall balance.

6 Limitations and Future Work

While generative methods are effective in safeguarding image privacy against various embedding methods, they cannot be compared to anonymization techniques. As [21] argues, it is always possible for new recognition attacks to be effective against provided data poisoning methods. However, our approach is different from anonymization, as we do not aim to provide a privacy guarantee against future attacks. Instead, our objective is to protect users against stalkers and unauthorized identification using current state-of-the-art recognition methods, while still enabling them to share their photos online, and be recognised by their family and friends. In the future, we would also like to use face enhancement as a tool for or against privacy methods, and check how much PrivacyGAN can be combined with AMT-GAN.

7 Conclusions

Our contributions are

1. A new approach to privacy based on inspirational generation, namely PrivacyGAN, using generative models for generating faces close to a given target. This method is orthogonal to the principles of AMT-GAN, so that our method could be used as a first step before AMT-GAN.

2. A comparison between these methods and traditional pixel-based methods, including transfer to unknown embeddings (a.k.a. robustness to unknown embeddings used for identifying people) and human raters for validating image quality.

3. A new privacy evaluation method based on the percentage of dataset images that are closer in an embedding space to a modified "private" image than to an original image.

4. A new dataset extracted from CC (more details are provided in the supplementary material). At the end, we recommend generative methods (Alg. 1), with several embeddings so that robustness and transfer to new methods are properly tested.

According to the human ratings study, Fawkes might be better than StyleGAN for generating high-quality images in the category "low privacy" (recall rate of 90%) on LFW. However, VQGAN and Fawkes have similar results in a low-privacy (74.76-79.21% as a transfer recall) setting, while VQGAN provides better privacy protection. Among the proposed generative methods, VQGAN is better than StyleGAN overall in terms of quality for a given privacy threshold (see Table 7). By the human identification study (section 5.5) we further show that, in contrast to anonymization methods, our method, along with other utility-focused methods, effectively safeguards facial image privacy against prevalent face recognition techniques while maintaining image utility for social media use.

In comparison to AMT-GAN, our method demonstrates superior privacy outcomes depending on the parameter settings, although it doesn't necessarily enhance human recognizability. While AMT-GAN excels in scenarios where recognizability is high and privacy is low, PrivacyGAN offers a broader spectrum of applications, particularly in cases where definitive facial detection is challenging, or where makeup contradicts the user's personal beliefs.

Figure 2: Experiment 1: Examples of original images from the LFW dataset and their counterparts modified by different privacy methods: StyleGAN_0.003_500, VQGAN_0.005_128, StyleGAN, VQGAN (using FaceNet as an embedding method for optimization), and Fawkes. Here we can see that, while generative methods in general add more modification to an image than Fawkes, generative methods produce realistic images and do not add pixel noise. Study based on human ratings in Table 7.

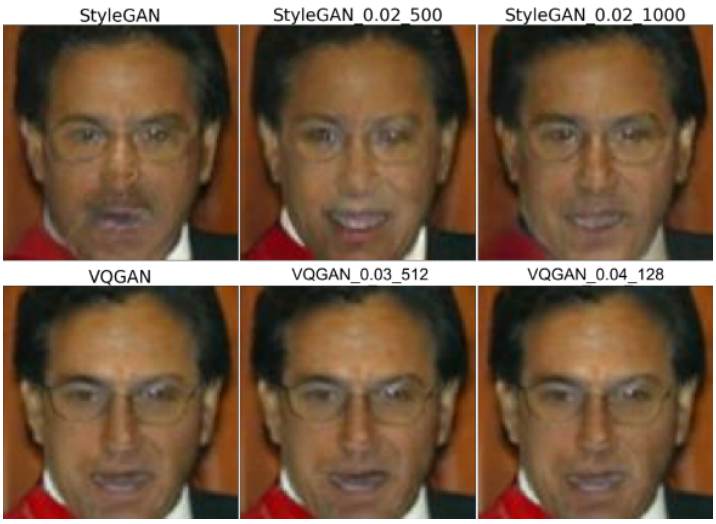

Figure 3: Experiment 2: Examples of images from the LFW dataset modified by different privacy methods: StyleGAN, StyleGAN_0.02_500, StyleGAN_0.02_1000, VQGAN, VQGAN_0.03_512, VQGAN_0.04_128 with embedding methods MagFace and MobileFaceNet. Original and modified Fawkes image versions can be seen in Fig. 2. These images have different privacy levels and different qualities: the human rating experiment performs comparisons between images produced by methods with similar recall, *i.e.*, similar privacy results.

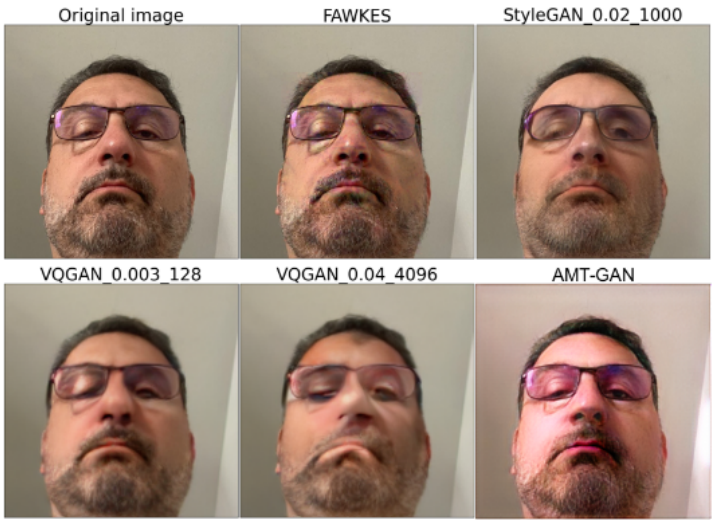

Figure 4: Experiment 3: Examples of images of volunteers modified by various privacy methods, including AMT-GAN, Fawkes and StyleGAN_0.02_1000, VQGAN_0.003_128, VQGAN_0.04_4096 optimised with embedding methods MagFace and MobileFaceNet. The different methods and parametrizations lead to different image quality/privacy results; the human rating experiments will compare the quality for methods with similar recall.

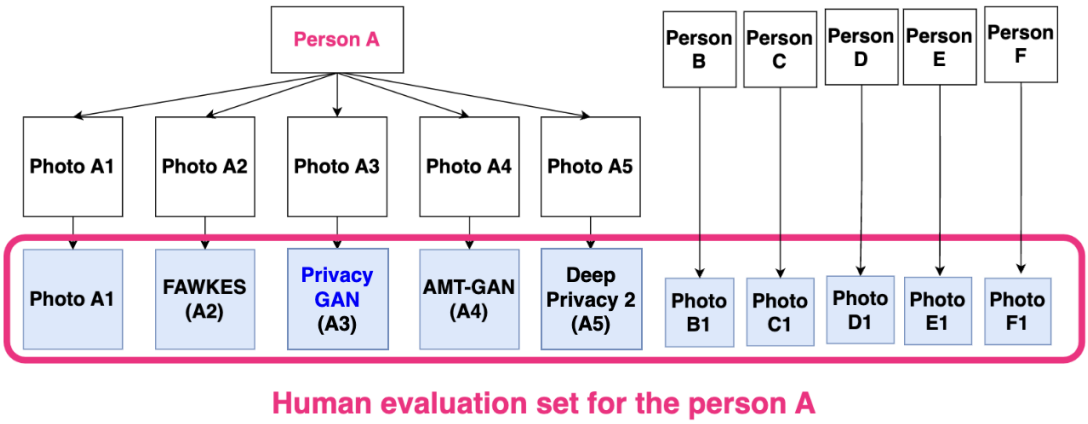

Figure 5: Schema of the human identification experiment setup. The experiment involves comparing the original image with privacy-preserved versions of images of the same person and random images of different people to determine whether privacy-preservation methods could preserve the utility of the images modified by them: we expect human raters to recognize the original face for privacy-preserving methods, and not for anonymization methods.

References

[1] Simone Barattin, Christos Tzelepis, Ioannis Patras, and Nicu Sebe. *Attribute-preserving face dataset anonymization via latent code optimization*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8001–8010, 2023.

[2] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. *Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices*. In Chinese Conference on Biometric Recognition, pages 428–438. Springer, 2018.

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

[4] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. *Fast geometrically-perturbed adversarial faces*. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1979–1988. IEEE, 2019.

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. *Arcface: Additive angular margin loss for deep face recognition*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4690–4699, 2019.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. *Taming transformers for high-resolution image synthesis*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12873–12883, 2021.

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

[8] Drew Harwell. *This facial recognition website can turn anyone into a cop—or a stalker*. In Ethics of Data and Analytics, pages 63–67. Auerbach Publications, 2022.

[9] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. *Towards measuring fairness in ai: the casual conversations dataset*. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep residual learning for image recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[11] Fabio Hellmann, Silvan Mertes, Mohamed Benouis, Alexander Hustinx, Tzung-Chien Hsieh, Cristina Conati, Peter Krawitz, and Elisabeth André. *Ganonymization: A gan-based face anonymization framework for preserving emotional expressions*. arXiv preprint arXiv:2305.02143, 2023.

[12] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. *Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15014–15023, 2022.

[13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. In Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

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

[15] Håkon Hukkelås and Frank Lindseth. *Deepprivacy2: Towards realistic full-body anonymization*. In 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1329–1338, 2023.

[16] Tero Karras, Samuli Laine, and Timo Aila. *A style-based generator architecture for generative adversarial networks*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019.

[17] Taehoon Kim and Jihoon Yang. *Latent-space-level image anonymization with adversarial protector networks*. IEEE Access, 7:84992–84999, 2019.

[18] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. *Sphereface: Deep hypersphere embedding for face recognition*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 212–220, 2017.

[19] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. *Magface: A universal representation for face recognition and quality assessment*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14225–14234, 2021.

[20] Yuying Qiu, Zhiyi Niu, Biao Song, Tinghuai Ma, Abdullah Al-Dhelaan, and Mohammed Al-Dhelaan. *A novel generative model for face privacy protection in video surveillance with utility maintenance*. Applied Sciences, 12(14):6962, 2022.

[21] Evani Radiya-Dixit and Florian Tramèr. *Data poisoning won't save you from facial recognition*. arXiv preprint arXiv:2106.14851, 2021.

[22] Ian Sample. *What is facial recognition-and how sinister is it*. The Guardian, 29, 2019.

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

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

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

[26] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. *Facex-zoo: A pytorch toolbox for face recognition*. In Proceedings of the 29th ACM International Conference on Multimedia, pages 3779–3782, 2021.

[27] Kuan-Chieh Wang, Yan Fu, Ke Li, Ashish Khisti, Richard Zemel, and Alireza Makhzani. *Variational model inversion attacks*. Advances in Neural Information Processing Systems, 34:9706–9719, 2021.

[28] Thibaut Weise, Hao Li, Luc Van Gool, and Mark Pauly. *Face/off: Live facial puppetry*. In Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer animation, pages 7–16, 2009.

[29] Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y Zhao. *Sok: Anti-facial recognition technology*. In 2023 IEEE Symposium on Security and Privacy (SP), pages 864–881. IEEE, 2023.

[30] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In FAWKES
github: https://github.com/Shawn-Shan/fawkes, 2020

[31] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. *The unreasonable effectiveness of deep features as a perceptual metric*. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

A Appendix

Facial image privacy protection is a multi-objective problem combining image quality preservation and privacy robustness against various image recognition systems:

- In Section A.1, we provide the description of generative makeup transfer method AMT-GAN
- In Sections A.2-A.4, we present quantitative results (table of recall/percentage, showing privacy performance).
- Then, in Sections A.5.1-A.5.3, we present images, showing the image quality.

A human evaluation study that combines both privacy and image quality aspects is available in the main paper (Table 7, Table 8).

A.1 Privacy algorithms: generative makeup transfer method AMT-GAN

In this paper, we compare the results of our method PrivacyGAN based on generative techniques such as StyleGAN [16] and VQGAN[6] to a method for generative makeup transfer known as AMT-GAN [12]. The objective of AMT-GAN is to produce adversarial images that incorporate the makeup style of reference images. Although AMT-GAN introduces more alterations to the original image, it confines these modifications to the makeup application areas, thus resulting in visually natural images, as demonstrated by the FID results. The authors of the method employ LPIPS [31] loss to retain image similarity to the original, which we also utilise in our paper.

In the main body of the paper, we mention that while photographs of individuals wearing makeup may look appropriate and natural to some people, there are certain drawbacks to this approach. Firstly, publishing photos with makeup may be deemed unsuitable for certain groups of individuals. Secondly, if the face in the photograph is not clearly discernible, AMT-GAN may not recognise it and may not generate a private version of the image, unlike Fawkes [24] and our method.

A.2 Experiment 1: additional and extended tables of results

In this subsection, we present transfer results for Experiment 1. All the tables here are similar to the Table 2 except for the different choice of transfer embeddings used in recognition.

To be precise, we evaluate the transfer results of PrivacyGAN equipped with generative methods VQGAN and StyleGAN optimised with FaceNet[23] embedding and compare them to the transfer results of Fawkes. The criterion is the transfer to other embeddings than FaceNet, namely ArcFace[5], MobileFaceNet[2], and Resnet_152[10].

We also present the results of Fawkes combination with generative methods. We note that combining Fawkes poisoning with our methods can be beneficial for facial image privacy protection.

Table 9 presents results of the transfer to ArcFace embedding; Table 10 presents results of the transfer to MobileFaceNet embedding; and Table 11 presents the results of transfer to Resnet_152 embedding method.

From the results of this experiment, we conclude that generative methods optimised with one single embedding do not provide strong privacy protection, but their results are still better than the results of Fawkes.

A.3 Experiment 2 (comparing PrivacyGAN equipped with StyleGAN and PrivacyGAN equipped with VQGAN optimised with 2 embedding methods on the LFW dataset): additional tables of results

In the main part of the article, we presented the results of Experiment 2 without transfer for embedding method MagFace in Table 4 and with transfer for embedding method SphereFace[18] in Table 5. Here we present another example without transfer for embedding method MobileFaceNet in Table 15 and with transfer for embedding methods FaceNet, ArcFace, and ResNet_152 in Tables 12, 13 and 14.

From the results of the experiment 2, we conclude that generative methods optimised with two different embeddings provide stronger privacy protection than those optimised with a single embedding method (as in experiment 1).

A.4 Experiment 3 (comparing PrivacyGAN, AMT-GAN, and Fawkes on CC dataset): additional tables of results

In this section, we present the results of experiment 3. All the tables here are similar to Table 6 of the main paper, except for the differences in choice of transfer embeddings.

We present results for generative methods with a criterion based on transfer from MobileFaceNet and MagFace (used in our privacy algorithm) to embeddings FaceNet, ArcFace and ResNet_152 (used in the recognition) in Tables 16, 18 and 20 as well as results without transfer for embeddings MagFace and MobileFaceNet in Tables 17 and 19.

Overall, results for the CC dataset are similar to those for LFW, and generative methods remain preferable for privacy protection.

A.5 Image examples for original and modified images using both pixel-based and generative methods

The examples in this section show that generative methods modify image features more than the pixel-based methods. Nonetheless, they have less artificial pixel noise, which is common for images protected by Fawkes. Pixel noise can be more detrimental in terms of visual quality.

A.5.1 Modified image examples: experiment 1

In this section, we present original images and their private versions that were obtained in the course of experiment 1 (similarly to Fig. 2). In addition, we also provide image examples for combinations of Fawkes and generative methods, namely Fawkes + StyleGAN (F+S), Fawkes + VQGAN (F+V), StyleGAN + Fawkes (S+F) and VQGAN + Fawkes (V+F). We see that adding Fawkes on top of generative methods improves image privacy, as in methods StyleGAN + Fawkes (S+F) and VQGAN + Fawkes (V+F). Tthe mentioned image examples can be found in Figs 6, 7, 8 and 9.

A.5.2 Modified image examples: experiment 2

In this section, we present the original images and their private versions that were obtained during experiment 2, in addition to the images that were presented in the main paper in Fig 3. These images can be found in Figs 10, 11, 12 and 13.

	PrivacyGAN				Pixel-based	Combinations			
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes	F + S	F + V	S + F	V + F
Percentage	5.052	0.917	4.931	0.984	0.936	7.612	7.855	**12.456**	10.397
Recall@1: m.i.	7.962	23.981	8.531	22.464	23.602	5.782	5.118	**2.464**	3.791
Recall@1: o.i.	8.246	24.976	10.095	24.692	24.313	6.445	7.062	**2.749**	4.408
Recall@3: m.i.	18.152	56.019	18.720	57.062	59.953	12.749	11.848	**7.346**	10.095
Recall@3: o.i.	17.536	56.730	19.858	58.578	59.431	11.991	13.175	**6.019**	8.768
Recall@5: m.i.	24.076	69.905	24.265	72.986	76.209	16.919	16.114	**10**	13.081
Recall@5: o.i.	22.038	70.379	24.929	72.559	75.877	15.782	16.445	**8.199**	11.611
Recall@10: m.i.	32.227	79.052	33.365	79.953	83.175	23.649	23.791	**13.886**	19.194
Recall@10: o.i.	30.190	78.389	33.318	79.905	83.033	21.754	22.607	**12.038**	16.635
Recall@50: m.i.	53.791	90.379	55.829	90.995	92.464	42.559	43.223	**28.436**	34.882
Recall@50: o.i.	52.701	91.185	55.924	91.469	92.227	41.706	42.607	**26.398**	31.848
Recall@100: m.i.	63.934	93.602	64.882	93.697	94.929	51.991	52.938	**37.488**	44.313
Recall@100: o.i.	63.318	94.076	65.877	94.123	94.692	51.754	53.602	**35.782**	42.607

Table 9: Evaluation on LFW dataset in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with FaceNet, and all tests are performed with ArcFace. Lower recall and a higher percentage mean better privacy. As shown in Table 14 methods that are optimised for two embeddings have a better transfer recall. We see that adding Fawkes on top of generative methods improves image privacy, as in the methods StyleGAN + Fawkes (S+F) and VQGAN + Fawkes (V+F).

A.5.3 Modified image examples: experiment 3

Here, we present more image examples obtained by the procedure described in experiment 3. They are obtained the same way as images in Fig. 4. These examples are presented in Figs 14, 15 and 16.

We would also like to note that all necessary approvals were obtained for the use of images in the present paper.

Figure 6: Experiment 1: Examples of original images from the LFW dataset and their counterparts modified by different privacy methods: StyleGAN_0.003_500, VQGAN_0.005_128, StyleGAN, VQGAN (using FaceNet as an embedding method for optimisation), Fawkes, and combinations of Fawkes with generative methods. Here we can see that while generative methods in general add more modification to an image than Fawkes, generative methods produce realistic images and do not add pixel noise.

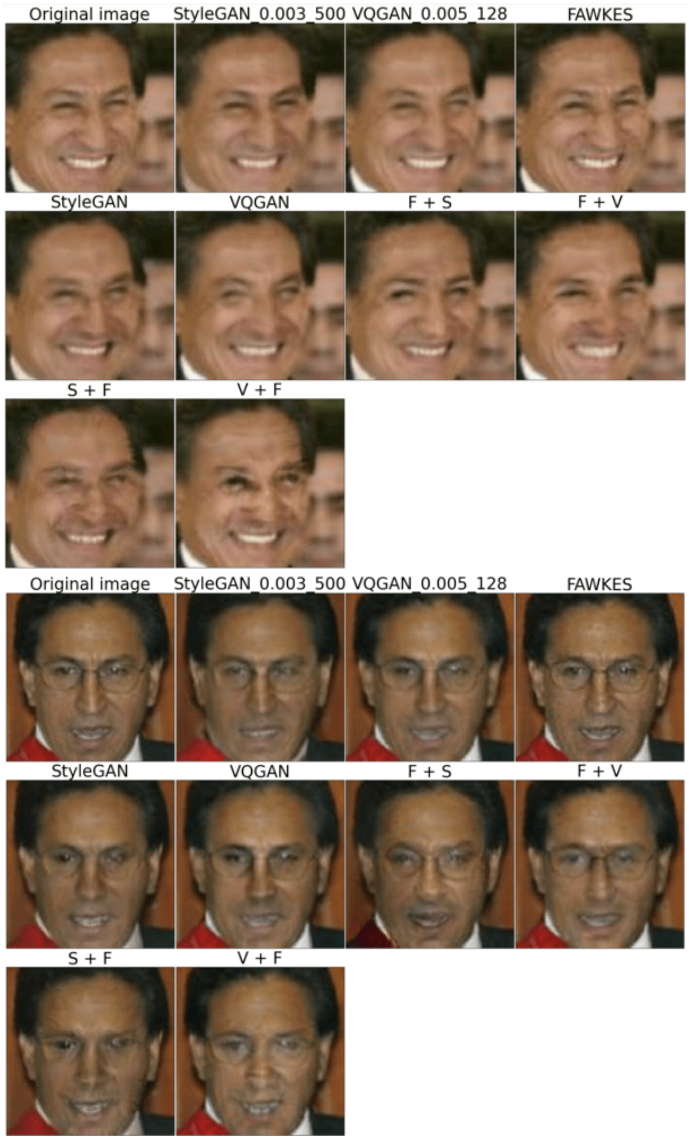

Figure 7: Experiment 1: Examples of original images from the LFW dataset and their counterparts modified by different privacy methods: StyleGAN_0.003_500, VQGAN_0.005_128, StyleGAN, VQGAN (using FaceNet as an embedding method for optimisation), Fawkes, and combinations of Fawkes with generative methods. Here we can see that while generative methods in general add more modification to an image than Fawkes, generative methods produce realistic images and do not add pixel noise.

Figure 8: Experiment 1: Examples of original images from the LFW dataset and their counterparts modified by different privacy methods: StyleGAN_0.003_500, VQGAN_0.005_128, StyleGAN, VQGAN (using FaceNet as an embedding method for optimisation), Fawkes, and combinations of Fawkes with generative methods. Here we can see that while generative methods in general add more modification to an image than Fawkes, generative methods produce realistic images and do not add pixel noise.

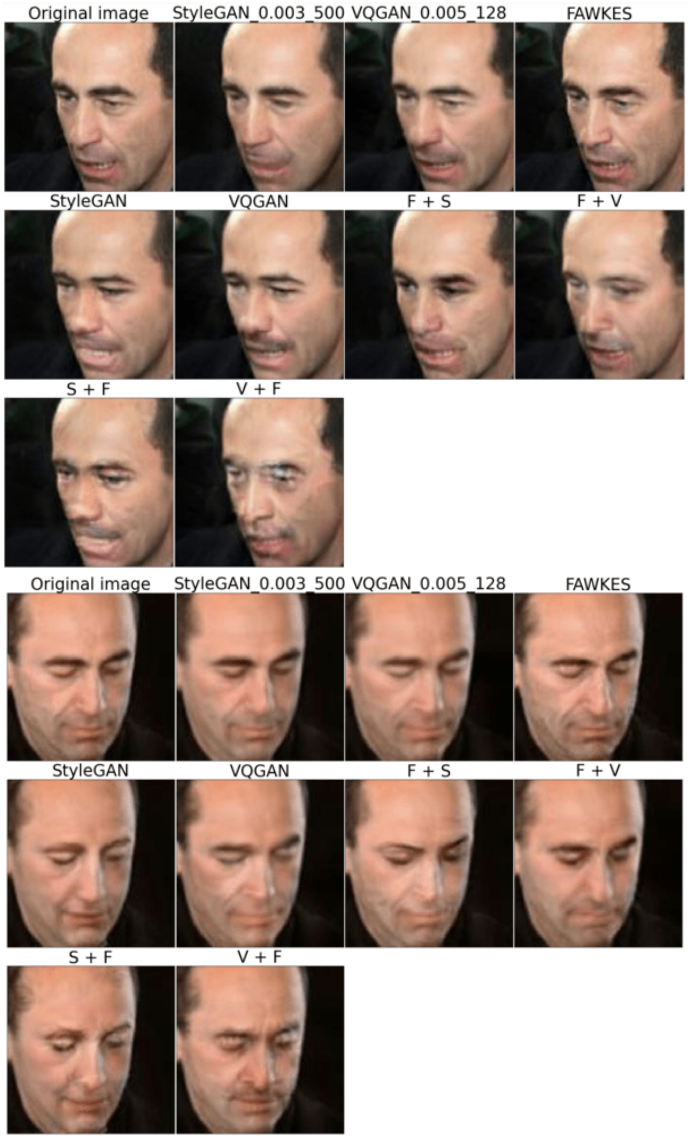

Figure 9: Experiment 1: Examples of original images from the LFW dataset and their counterparts modified by different privacy methods: StyleGAN_0.003_500, VQGAN_0.005_128, StyleGAN, VQGAN (using FaceNet as an embedding method for optimisation), Fawkes, and combinations of Fawkes with generative methods. Here we can see that while generative methods in general add more modification to an image than Fawkes, generative methods produce realistic images and do not add pixel noise.

Figure 10: Experiment 2: Examples of images from the LFW dataset: the original image and the image modified by different privacy methods: StyleGAN, StyleGAN_0.02_500, StyleGAN_0.02_1000, VQGAN, VQGAN_0.03_512, VQGAN_0.04_128 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection.

Figure 11: Experiment 2: Examples of images from the LFW dataset: the original image and the image modified by different privacy methods: StyleGAN, StyleGAN_0.02_500, StyleGAN_0.02_1000, VQGAN, VQGAN_0.03_512, VQGAN_0.04_128 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection.

Figure 12: Experiment 2: Examples of images from the LFW dataset: the original image and the image modified by different privacy methods: StyleGAN, StyleGAN_0.02_500, StyleGAN_0.02_1000, VQGAN, VQGAN_0.03_512, VQGAN_0.04_128 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection.

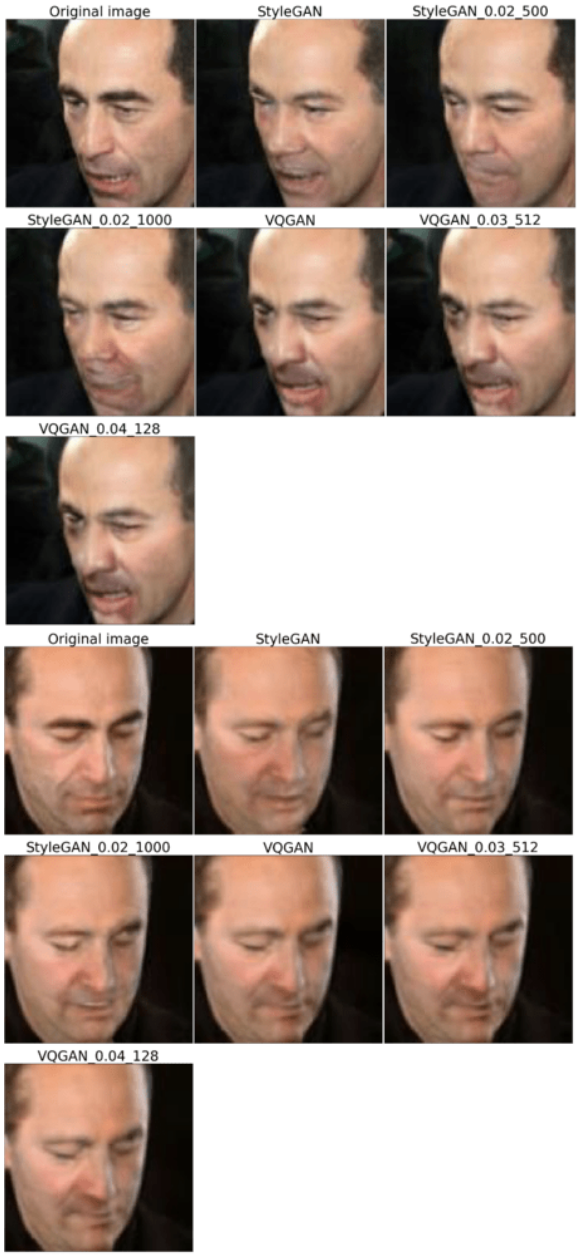

Figure 13: Experiment 2: Examples of images from the LFW dataset: the original image and the image modified by different privacy methods: StyleGAN, StyleGAN_0.02_500, StyleGAN_0.02_1000, VQGAN, VQGAN_0.03_512, VQGAN_0.04_128 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection.

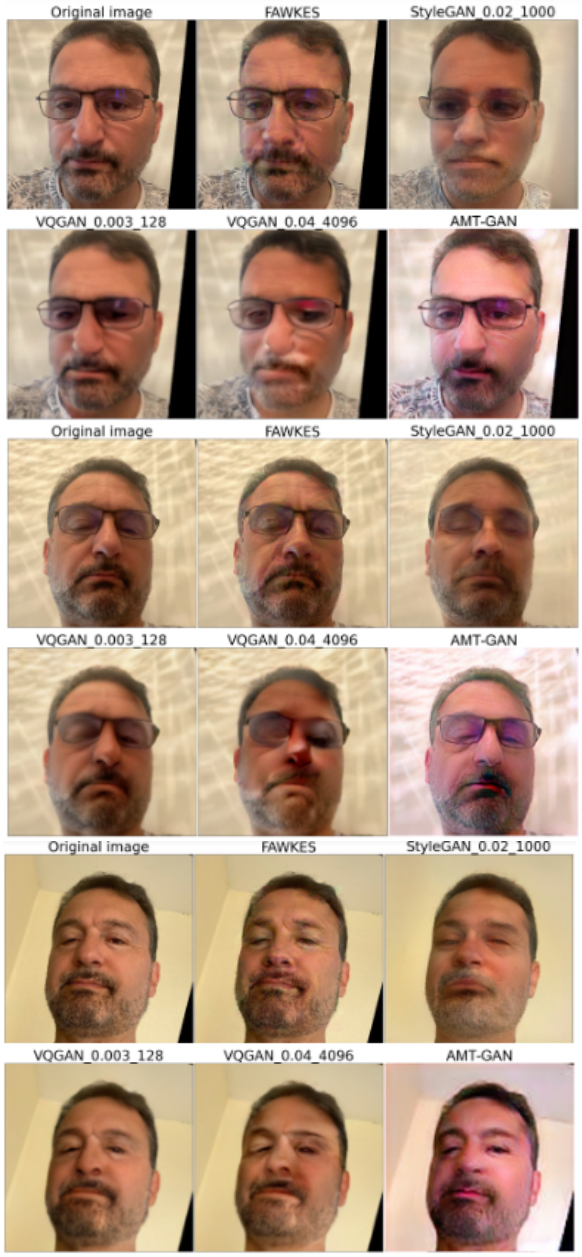

Figure 14: Experiment 3: Examples of images of volunteers modified by various privacy methods, including AMT-GAN, Fawkes, StyleGAN_0.02_1000, VQGAN_0.003_128, VQGAN_0.04_4096 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection. We can also see that some images are modified more than others after applying privacy-protection methods.

Figure 15: Experiment 3: Examples of images of volunteers modified by various privacy methods, including AMT-GAN, Fawkes, StyleGAN_0.02_1000, VQGAN_0.003_128, VQGAN_0.04_4096 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection. We can also see that some images are modified more than others after applying privacy-protection methods.

Figure 16: Experiment 3: Examples of images of volunteers modified by various privacy methods, including AMT-GAN, Fawkes, StyleGAN_0.02_1000, VQGAN_0.003_128, VQGAN_0.04_4096 with embedding methods MagFace and MobileFaceNet. Here, we can see that in some cases, increasing the number of iterations in optimisation and modifying the coefficient K of the embedding method can affect the quality of an image while improving its privacy protection. We can also see that some images are modified more than others after applying privacy-protection methods.

	PrivacyGAN				Pixel-based	Combinations			
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes	F + S	F + V	S + F	V + F
Percentage	1.226	0.446	0.947	0.454	0.454	1.875	4.359	**8.645**	5.586
Recall@1: m.i.	19.479	26.872	21.991	26.588	25.592	17.773	**1.185**	7.062	10.332
Recall@1: o.i.	19.100	26.303	21.943	26.493	26.351	19.147	12.938	**5.071**	8.910
Recall@3: m.i.	55.166	72.275	58.815	73.981	72.986	43.033	18.436	**16.825**	23.555
Recall@3: o.i.	51.090	74.692	56.682	72.701	73.697	41.517	26.161	**11.232**	19.005
Recall@5: m.i.	70.616	94.360	74.265	96.209	94.408	55.261	30.853	**22.275**	30.444
Recall@5: o.i.	64.882	93.934	71.896	95.498	93.934	51.896	31.801	**14.929**	24.550
Recall@10: m.i.	77.678	96.303	82.227	97.536	96.351	65.118	42.227	**29.052**	38.341
Recall@10: o.i.	75.024	96.066	80.379	97.156	96.019	61.848	41.185	**21.801**	33.033
Recall@50: m.i.	88.626	98.436	91.517	98.626	97.962	79.763	62.370	**48.483**	58.104
Recall@50: o.i.	87.867	98.389	90.284	98.389	97.725	79.242	61.943	**41.280**	53.460
Recall@100: m.i.	91.659	98.815	94.313	98.815	98.436	84.929	71.090	**55.735**	66.446
Recall@100: o.i.	91.754	98.673	93.223	98.768	98.152	84.408	69.716	**51.043**	62.844

Table 10: Evaluation on LFW dataset in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with FaceNet, and recognition (for all methods) is tested with MobileFaceNet. Lower recall and a higher percentage mean better privacy. Generative methods do obtain better results than Fawkes. We can also see that adding Fawkes on top of generative methods improves image privacy, as in the methods StyleGAN + Fawkes (S+F) and VQGAN + Fawkes (V+F).

	PrivacyGAN				Pixel-based	Combinations			
	StyleGAN	StyleGAN _0.003_500	VQGAN	VQGAN _0.005_128	Fawkes	F + S	F + V	S + F	V + F
Percentage	0.670	0.342	0.564	0.395	0.408	1.273	2.573	**6.823**	3.309
Recall@1: m.i.	20.900	25.071	23.270	26.019	25.403	17.488	8.436	**7.536**	11.896
Recall@1: o.i.	20.616	26.398	22.607	25.972	27.488	22.938	17.204	**8.957**	15.071
Recall@3: m.i.	59.526	73.128	65.592	75.545	76.351	49.573	30.000	**18.957**	34.597
Recall@3: o.i.	58.199	73.886	63.886	72.559	75.498	49.431	38.009	**17.251**	31.706
Recall@5: m.i.	78.673	97.773	87.583	98.626	98.578	66.303	45.498	**27.299**	45.924
Recall@5: o.i.	73.791	97.441	83.081	98.531	98.389	62.512	48.720	**21.706**	40.758
Recall@10: m.i.	85.972	98.483	91.801	98.863	99.005	74.976	58.483	**35.545**	56.493
Recall@10: o.i.	82.891	98.389	89.668	98.863	98.815	72.370	59.479	**30.332**	51.754
Recall@50: m.i.	94.028	99.005	97.014	99.052	99.194	87.915	78.152	**55.735**	74.123
Recall@50: o.i.	93.128	99.147	95.924	99.194	99.147	87.583	77.393	**51.469**	72.464
Recall@100: m.i.	95.877	99.100	97.867	99.100	99.194	91.659	83.744	**63.934**	80.095
Recall@100: o.i.	95.308	99.194	97.299	99.194	99.147	91.422	82.938	**60.900**	78.531

Table 11: Evaluation on LFW dataset in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with FaceNet and tested with ResNet_152. Lower recall and a higher percentage mean better privacy. The generative methods with two embeddings (see Table 13) do obtain better results, showing that using multiple embeddings increases robustness and transfer. We can also see that adding Fawkes on top of generative methods improves image privacy, as in the methods StyleGAN + Fawkes (S+F) and VQGAN + Fawkes (V+F).

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512	VQGAN _0.04_128
Percentage	**7.849**	4.494	4.107	2.626	3.470	3.807
Recall@1: m.i.	**2.844**	6.066	6.730	8.152	6.919	6.019
Recall@1: o.i.	**4.597**	8.294	10.047	11.706	9.479	9.431
Recall@3: m.i.	**9.242**	15.640	18.009	22.701	19.242	17.536
Recall@3: o.i.	**9.668**	17.156	19.479	25.545	20.995	19.431
Recall@5: m.i.	**13.175**	22.322	25.166	32.275	27.062	24.787
Recall@5: o.i.	**13.081**	22.701	25.118	32.749	27.109	25.592
Recall@10: m.i.	**18.578**	31.564	35.118	43.175	37.725	34.408
Recall@10: o.i.	**17.725**	30.758	34.360	43.412	36.019	33.460
Recall@50: m.i.	**37.441**	54.028	57.062	67.725	60.142	57.441
Recall@50: o.i.	**35.213**	52.464	55.118	66.019	58.910	55.592
Recall@100: m.i.	**48.531**	64.882	68.436	77.109	71.137	68.104
Recall@100: o.i.	**46.351**	63.934	65.735	75.640	69.147	67.014

Table 12: Evaluation on LFW dataset in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with FaceNet. Lower recall and a higher percentage mean better privacy. In this case, generative methods optimised with two embeddings obtain worse results than generative methods optimised with only one embedding method in Table 1: this is, however, not a fair comparison because we do not use FaceNet for optimisation in this experiment, while we do in Table 1.

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512	VQGAN _0.04_128
Percentage:	**5.656**	2.671	2.296	1.549	2.097	2.605
Recall@1: m.i.	**4.976**	8.294	11.422	12.322	9.526	7.630
Recall@1: o.i.	**9.289**	15.450	18.152	18.957	16.635	15.403
Recall@3: m.i.	**15.829**	28.578	32.986	37.488	31.043	28.057
Recall@3: o.i.	**19.763**	34.597	39.100	44.455	38.863	35.545
Recall@5: m.i.	**23.507**	41.327	46.445	55.687	46.161	41.469
Recall@5: o.i.	**26.114**	43.175	49.858	56.303	48.246	43.981
Recall@10: m.i.	**34.218**	52.796	58.957	68.673	59.716	54.929
Recall@10: o.i.	**34.313**	52.796	60.616	68.531	60.284	55.071
Recall@50: m.i.	**56.398**	74.123	78.104	84.692	79.763	76.540
Recall@50: o.i.	**54.929**	73.507	79.005	84.218	79.242	77.109
Recall@100: m.i.	**64.882**	81.611	84.360	89.242	85.118	82.986
Recall@100: o.i.	**63.839**	80.853	84.550	88.863	85.118	83.081

Table 13: Evaluation on LFW dataset in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet and recognition is tested with ResNet_152. Lower recall and a higher percentage mean better privacy. In this case, generative methods optimised with two embeddings obtain better results than generative methods optimised with only one embedding method, as in Table 10.

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512	VQGAN _0.04_128
Percentage	**11.993**	8.305	7.534	6.067	7.164	8.181
Recall@1: m.i.	**2.038**	4.123	4.550	6.967	4.882	4.739
Recall@1: o.i.	**3.555**	5.687	7.488	7.867	6.540	6.351
Recall@3: m.i.	**6.588**	11.754	12.370	15.640	13.270	11.232
Recall@3: o.i.	**6.919**	12.417	14.787	16.398	13.791	12.038
Recall@5: m.i.	**9.005**	16.209	17.109	20.758	19.052	16.588
Recall@5: o.i.	**9.526**	16.493	18.673	21.611	18.578	16.066
Recall@10: m.i.	**13.081**	22.844	24.976	29.384	26.066	22.417
Recall@10: o.i.	**14.171**	22.938	25.450	29.810	25.450	22.322
Recall@50: m.i.	**27.299**	41.374	43.270	51.280	44.455	41.611
Recall@50: o.i.	**28.863**	41.991	45.024	51.754	46.730	42.464
Recall@100: m.i.	**36.445**	50.806	52.986	61.137	54.360	51.232
Recall@100: o.i.	**38.910**	53.128	56.161	62.796	57.867	52.796

Table 14: Evaluation on LFW dataset, in the case of transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with ArcFace. Lower recall and a higher percentage mean better privacy. In this case, generative methods optimised with two embeddings obtain better results than generative methods optimised with only one embedding method, as in Table 9.

	StyleGAN	StyleGAN _0.02_500	StyleGAN _0.02_1000	VQGAN	VQGAN _0.03_512	VQGAN _0.04_128
Percentage	**6.925**	3.290	3.079	2.583	3.555	4.461
Recall@1: m.i.	**0.569**	2.986	2.749	2.227	0.900	0.664
Recall@1: o.i.	**6.635**	14.739	14.976	15.782	12.654	10.379
Recall@3: m.i.	**11.090**	22.417	24.408	26.066	20.190	17.062
Recall@3: o.i.	**13.365**	30.095	32.180	34.834	26.351	22.085
Recall@5: m.i.	**18.531**	36.635	38.152	42.607	31.517	28.246
Recall@5: o.i.	**18.009**	38.578	40.379	45.308	34.550	29.289
Recall@10: m.i.	**26.682**	48.389	49.953	55.735	43.223	39.052
Recall@10: o.i.	**24.739**	48.720	49.100	54.739	44.645	38.720
Recall@50: m.i.	**47.014**	69.384	71.611	74.171	65.687	59.668
Recall@50: o.i.	**46.872**	69.716	71.991	74.597	66.919	60.427
Recall@100: m.i.	**56.682**	77.583	78.957	80.237	73.981	68.578
Recall@100: o.i.	**56.967**	76.588	79.336	80.948	75.308	69.431

Table 15: Evaluation on LFW dataset: VQGAN and StyleGAN (and their variants) are optimised with MagFace and MobileFaceNet, and recognition is tested with MobileFaceNet. Generative methods do obtain better privacy than Fawkes (Table 10).

	VQGAN _0.003_128	VQGAN	VQGAN _0.04_4096	Fawkes	StyleGAN _0.02_1000	AMT-GAN
Percentage	0.697	3.975	8.312	0.715	**12.268**	1.633
Recall@1: m.i.	23.571	10.712	5.055	23.831	**3.450**	15.868
Recall@1: o.i.	24.012	15.125	8.064	24.534	**6.239**	16.429
Recall@3: m.i.	70.271	26.359	12.778	66.680	**8.947**	45.256
Recall@3: o.i.	71.274	32.177	15.527	66.800	**12.237**	46.319
Recall@5: m.i.	91.775	36.169	17.934	87.964	**13.340**	60.522
Recall@5: o.i.	91.555	39.860	20.040	84.875	**15.787**	60.923
Recall@10: m.i.	94.965	45.537	24.835	91.575	**18.495**	68.867
Recall@10: o.i.	94.905	48.265	26.239	89.950	**20.963**	69.829
Recall@50: m.i.	97.733	65.597	42.508	96.309	**34.483**	83.470
Recall@50: o.i.	97.553	66.941	45.436	95.125	**36.349**	84.554
Recall@100: m.i.	98.235	72.457	51.775	97.151	**42.989**	88.465
Recall@100: o.i.	98.195	75.125	55.486	96.670	**46.098**	89.749

Table 16: Evaluation on CC dataset in the case of a transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with FaceNet. Lower recall and a higher percentage mean better privacy. We can see that generative methods' evaluation results for the CC dataset are very similar to the ones for the LFW dataset.

	VQGAN _0.003_128	VQGAN	VQGAN _0.04_4096	Fawkes	StyleGAN _0.02_1000	AMT-GAN
Percentage	2.880	14.337	**27.089**	2.507	22.651	4.201
Recall@1: m.i.	23.390	2.086	0.562	24.674	**0.301**	19.097
Recall@1: o.i.	21.926	5.918	**1.043**	24.995	1.886	17.031
Recall@3: m.i.	62.708	5.998	**1.143**	67.462	1.484	52.277
Recall@3: o.i.	63.129	10.973	**2.187**	67.683	3.992	49.107
Recall@5: m.i.	78.335	9.468	**1.805**	84.393	2.768	64.654
Recall@5: o.i.	78.154	13.561	**2.949**	83.952	5.296	63.149
Recall@10: m.i.	81.846	14.845	**3.230**	86.399	4.975	70.351
Recall@10: o.i.	82.046	17.553	**4.794**	86.219	7.763	68.706
Recall@50: m.i.	87.182	30.211	**9.007**	89.709	14.624	80.100
Recall@50: o.i.	87.603	31.033	**11.013**	89.749	16.309	78.495
Recall@100: m.i.	89.087	38.736	**14.203**	90.853	20.702	84.092
Recall@100: o.i.	89.348	39.238	**15.226**	90.913	22.287	82.247

Table 17: Evaluation on CC dataset, in the case without transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with MagFace. We can see that generative methods' evaluation results for the CC dataset are very similar to the ones for the LFW dataset.

	VQGAN _0.003_128	VQGAN	VQGAN _0.04_4096	Fawkes	StyleGAN _0.02_1000	AMT-GAN
Percentage	2.339	7.232	12.125	1.739	**14.657**	4.074
Recall@1: m.i.	21.184	9.188	5.115	24.453	**4.173**	15.266
Recall@1: o.i.	21.966	12.217	**6.640**	24.955	7.442	10.532
Recall@3: m.i.	55.426	22.628	11.635	63.71	**9.910**	37.733
Recall@3: o.i.	57.392	25.436	**13.260**	64.975	14.002	28.706
Recall@5: m.i.	69.468	28.445	15.165	78.034	**12.839**	48.004
Recall@5: o.i.	70.973	32.016	**16.830**	78.656	17.212	37.813
Recall@10: m.i.	75.165	35.426	19.960	81.825	**17.854**	55.366
Recall@10: o.i.	76.209	39.699	**22.487**	82.648	23.049	44.433
Recall@50: m.i.	84.072	52.879	35.807	88.104	**30.993**	70.090
Recall@50: o.i.	85.035	58.295	40.040	88.546	**38.716**	61.765
Recall@100: m.i.	87.442	61.204	44.714	90.913	**38.837**	76.670
Recall@100: o.i.	88.185	66.419	48.967	90.973	**46.620**	69.087

Table 18: Evaluation on CC dataset in the case of a transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with ArcFace. We can see that generative methods' evaluation results for the CC dataset are very similar to the ones for the LFW dataset.

	VQGAN _0.003_128	VQGAN	VQGAN _0.04_4096	Fawkes	StyleGAN _0.02_1000	AMT-GAN
Percentage	6.513	12.668	**20.632**	5.200	15.798	7.708
Recall@1: m.i.	20.542	7.763	**1.224**	24.313	3.972	14.664
Recall@1: o.i.	19.719	12.638	**5.155**	24.875	8.004	12.919
Recall@3: m.i.	53.440	17.673	**3.852**	60.100	9.850	36.871
Recall@3: o.i.	53.902	24.433	**10.251**	59.880	16.108	27.663
Recall@5: m.i.	63.952	22.648	**5.436**	70.812	13.159	44.835
Recall@5: o.i.	64.614	30.451	**12.417**	70.933	20.662	32.397
Recall@10: m.i.	67.643	29.027	**8.666**	74.022	18.415	50.973
Recall@10: o.i.	68.826	37.051	**16.510**	74.223	27.061	36.429
Recall@50: m.i.	74.564	43.771	**19.278**	79.539	33.280	64.092
Recall@50: o.i.	75.206	53.039	**29.027**	79.819	43.290	45.456
Recall@100: m.i.	77.733	51.013	**26.820**	82.006	41.204	69.649
Recall@100: o.i.	78.134	59.980	**37.131**	82.247	51.715	50.291

Table 19: Evaluation on CC dataset without transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace and MobileFaceNet, and recognition is tested with MobileFaceNet. Lower recall and a higher percentage mean better privacy. We can see that generative methods evaluation results for the CC dataset are very similar to the ones for the LFW dataset.

	VQGAN _0.003_128	VQGAN	VQGAN _0.04_4096	Fawkes	StyleGAN _0.02_1000	AMT-GAN
Percentage	3.025	4.970	6.915	2.652	**7.650**	4.518
Recall@1: m.i.	20.582	12.758	**7.783**	24.012	8.345	16.790
Recall@1: o.i.	20.602	14.584	11.033	24.754	**10.732**	15.908
Recall@3: m.i.	56.951	32.217	**20.361**	61.685	21.846	39.398
Recall@3: o.i.	56.971	33.079	**22.949**	62.347	24.092	34.303
Recall@5: m.i.	69.930	40.943	**27.021**	75.165	28.646	49.328
Recall@5: o.i.	69.629	41.083	**28.686**	74.905	30.150	41.805
Recall@10: m.i.	74.463	49.328	**35.065**	78.816	36.489	56.289
Recall@10: o.i.	74.183	49.589	**36.189**	78.696	37.232	47.141
Recall@50: m.i.	82.608	66.379	**53.521**	85.496	54.644	71.414
Recall@50: o.i.	82.508	67.041	**54.664**	85.216	56.851	59.920
Recall@100: m.i.	85.657	73.561	62.628	87.823	**62.327**	77.553
Recall@100: o.i.	85.537	73.400	**63.591**	87.803	64.754	65.236

Table 20: Evaluation on CC dataset in the case of a transfer to another embedding: VQGAN and StyleGAN are optimised with MagFace + MobileFaceNet, and recognition is tested with ResNet_152. Lower recall and a higher percentage mean better privacy. We can see that generative methods' evaluation results for the CC dataset are very similar to the ones for the LFW dataset.