Attribute-preserving Face Dataset Anonymization via Latent Code Optimization

Simone Barattin*[1] Christos Tzelepis*[2] Ioannis Patras[2] Nicu Sebe[1]

[1]University of Trento
simone.barattin@studenti.unitn.it, niculae.sebe@unitn.it

[2]Queen Mary University of London
{c.tzelepis, i.patras}@qmul.ac.uk

Abstract

This work addresses the problem of anonymizing the identity of faces in a dataset of images, such that the privacy of those depicted is not violated, while at the same time the dataset is useful for downstream task such as for training machine learning models. To the best of our knowledge, we are the first to explicitly address this issue and deal with two major drawbacks of the existing state-of-the-art approaches, namely that they (i) require the costly training of additional, purpose-trained neural networks, and/or (ii) fail to retain the facial attributes of the original images in the anonymized counterparts, the preservation of which is of paramount importance for their use in downstream tasks. We accordingly present a task-agnostic anonymization procedure that directly optimizes the images' latent representation in the latent space of a pre-trained GAN. By optimizing the latent codes directly, we ensure both that the identity is of a desired distance away from the original (with an identity obfuscation loss), whilst preserving the facial attributes (using a novel feature-matching loss in FaRL's [48] deep feature space). We demonstrate through a series of both qualitative and quantitative experiments that our method is capable of anonymizing the identity of the images whilst–crucially–better-preserving the facial attributes. We make the code and the pre-trained models publicly available at: https://github.com/chi0tzp/FALCO.

1. Introduction

The ubiquitous use of mobile devices equipped with high-resolution cameras and the ability to effortlessly share personal photographs and videos on social media poses a significant threat to data privacy. Considering that modern machine learning algorithms learn from vast amounts of data often crawled from the Web [18, 38], it has become increasingly important to consider the impact this has on the privacy of those individuals depicted. Motivated by privacy concerns, many societies have recently enacted strict legislation, such as the General Data Protection Regulation (GDPR) [7], which requires the consent of every person that might be depicted in an image dataset. Whilst such laws have obvious benefits to the privacy of those featured in image datasets, this is not without costly side effects to the research community. In particular, research fields such as computer vision and machine learning rely on the creation and sharing of high-quality datasets of images of humans for a number of important tasks including security [24], healthcare [1], and creative applications [18, 35].

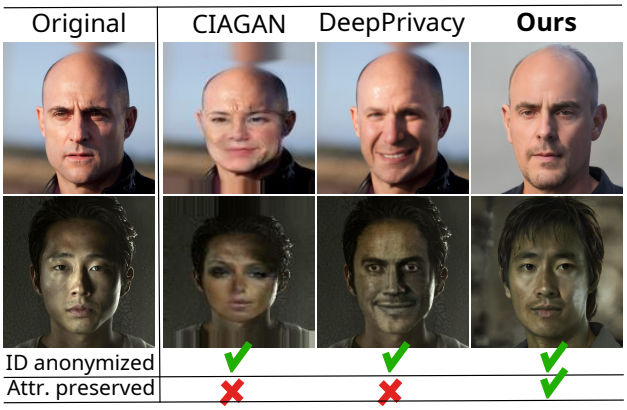

Figure 1. Comparison of the proposed method to CIAGAN [27] and DeepPrivacy [16] in terms of identity anonymization and attribute preservation.

A recent line of research focuses on overcoming this issue by *anonymizing* the identity of the individuals in image datasets. Through this approach, the machine learning com-

*These authors contributed equally. This work has been conducted during a research exchange visit of S. Barattin in QMUL in the framework of the EU H2020 project AI4Media.

munity can still benefit from the wealth of large datasets of high-resolution images, but without cost to privacy.

This research field has seen several developments throughout the last few years. Early methods proposed by the computer vision community attempt to solve this problem with simple solutions based on blurring [10] or other masking techniques, such as pixelation [12]. The result of this masking process succeeds in anonymizing the images by completely hiding the identity-related components, but as a consequence renders the facial attribute information such as a person's pose, expression, or skin tone (from which many computer vision tasks learn) indecipherable. Another problem with these methods is that, whilst the resulting images may not be re-identifiable by humans, they can often be reversed by deep learning models [28, 32].

Another line of work leverages the power of Generative Adversarial Networks (GANs) [13], which have recently been used for discovering controllable generation paths in their latent or feature spaces [2, 33, 34, 42, 43]. Towards face anonymization, GANs have been incorporated in order to synthesize new images in order to obtain photos that maintain most of the image while changing the face of the subject of interest. In particular, these approaches use techniques like image inpainting [16], conditional generation [27], attribute manipulation [21], or adversarial perturbation [39]. These works are able to obtain anonymized images that can still be used for computer vision tasks such as tracking and detection, with very good results in terms of privacy preservation. However, many of these works lack the ability to generate natural-looking faces and often fail to preserve the original facial attributes in the anonymized images (or, on the occasions in which such methods do preserve the facial attributes, they fail to demonstrate this quantitatively). This is critical for many applications which rely on the attributes of the inner face, such as expression recognition [20], or mental health affect analysis [11]. To further complicate the picture, a fundamental problem often found with existing works is the way in which the anonymized images copy not just the original image's background, but also more identifiable features [16, 27], such as the clothes of an individual, or their hair (see examples of this in Fig. 1). We argue that leaving such structure of the images unchanged constitutes a glaring privacy vulnerability, as one can re-identify the original image from the anonymized counterpart by comparing the image background or person's clothes.

Motivated by these concerns, in this work we propose to de-identify individuals in datasets of facial images whilst *preserving* the facial attributes of the original images. To achieve this, in contrast to existing work [16, 21, 27, 44, 45] that train custom neural networks from scratch, we propose to work directly in the latent space of a powerful pre-trained GAN, optimizing the latent codes directly with losses that explicitly aim to retain the attributes and obfuscate the identities. More concretely, we use a deep feature-matching loss [48] to match the high-level semantic features between the original and the fake image generated by the latent code, and a margin-based identity loss to control the similarity between the original and the fake image in the ArcFace [9] space. The initialisation of the latent codes is obtained by randomly sampling the latent space of GAN, using them to generate the corresponding synthetic images and finding the nearest neighbors in a semantic space (FARL [48]). In order to preserve texture and pose information of the original image, we perform inversion of the original image and retain the parts that correspond to the properties we want to preserve in the final code. This results in a latent code that yields a high-resolution image that contains a new identity but retains the same facial attributes as the original image.

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to address the problem of identity anonymization whilst also explicitly retaining facial attributes.
- We propose a novel methodology and loss functions working with *pre-trained* GANs capable of generating high-resolution anonymized datasets.
- We show through a series of thorough experiments on both Celeba-HQ [25] and LFW [15] that our method competes with the state-of-the-art in obfuscating the identity, whilst better-retaining the facial attributes under popular quantitative metrics.

2. Related Work

Face obfuscation The first privacy-preserving approaches proposed were based on obfuscating the face of the person. This means that different techniques, like blurring, masking, or pixelating [3, 5, 30, 40] are used to completely remove the personally identifiable information (PII). In the masking approach, the face region is simply covered with a shape such that the body or face of the person is completely covered, with pixelation the resolution of the face region is reduced, and blurring uses Gaussian filters with varying standard deviation values, allowing different strengths of the blurring. Tansuriyavong et al. [40] de-identifies people in a room by detecting the silhouette of the person, masking it, and showing only the name to balance privacy protection and the ability to convey information about the situation, Chen et al. [5] obscures the body information of a person with an obscuring algorithm exploiting the background subtraction technique leaving only the body outline visible. Naive de-identification techniques that maintain little information about the region of interest, such as pixelation and blurring, may seem to work to the human eye, but there exist approaches able to revert

the anonymized face to its original state [28, 32]. To improve the level of privacy protection, techniques defined as k-Same have been introduced [31], where, given a face, a de-identified visage is computed as the average of the k closest faces and then used to replace the original faces from the ones used in the calculation. This set of techniques works very well in removing privacy-related information, however, the result of the process completely removes the information related to the facial region, resulting in samples that are impossible to use in applications that need to use face detectors, trackers, or facial attributes. To solve these issues, our method instead leverages the generation capability of the state-of-the-art StyleGAN2 [19] to obtain realistic-looking face images, which are still detectable and that retain the facial attributes present in the original image.

Generative face anonymization After the advent of GANs [13], several lines of work have been proposed to tackle the problem of anonymization leveraging the generative power of these networks. Prior to this, [6, 26] proposed auto-encoder based methods, in particular Cho et al. [6] used such networks to learn to disentangle the identity information from the attributes given a vector representation of an image, allowing then to tweak the identity part of the vector to obtain an anonymized subject. This work reported emotion preservation results, however in this case we are concerned with the preservation of facial attributes more generally. The main weakness of such methods is the lack of sharpness of the generated images. Given the power of StyleGAN2 to synthesize sharp, high-resolution images, our framework avoids the problem of generating blurry images. Several methods [8, 16, 21, 27, 39, 44, 45] also employ the use of GAN networks to similar ends, given their incredible ability to capture the distribution of the training samples and then generate similar looking images. DeepPrivacy [16] extracts the face region along with sparse facial key points from a face image, removes the face from the image using the bounding box coordinates and sets its region to a constant value. This is passed to a conditional GAN [29], along with background and pose information, which inpaints a randomly generated face from StyleGAN2's [19] generator, while maintaining contextual and pose information. CIAGAN [27] also uses a conditional GAN, performing a form of conditional ID-swapping. The model uses an identity discriminator to force the generated image, conditioned on the landmark information and masked image, to display a different identity from the source one with similar features to the borrowed identity. These two methods obtain great privacy preservation results, but still lack the ability to generate natural-looking faces for images of large resolution. Even if contextual knowledge is injected under the form of the masked original image, i.e., without the face region, there is no guarantee that facial attributes are retained after the anonymization. In the case of CIAGAN in particular, the results are visually similar to the original subject only when the conditional ID happens to share the same features such as gender, age or skin tone. Using our attribute preservation loss, and thanks to "prior knowledge" of textural information that we gain from the original inverted image latent code, these issues are solved. Techniques of ID-swapping, like [27], present also another issue: since *real* images' identity features are used to condition the anonymized face, it is unclear if this truly solves the privacy problem. These issues are avoided in our framework, as the pre-trained generator outputs random, non-existing faces and thus no information about the original identities is retained. One of the most recent works, namely IdentityDP [44], tackles the problem of anonymization in a three-fold process: first an attribute encoder and an identity encoder are used to extract the corresponding features, which are then injected into a GAN to reconstruct the original image. In this way the encoders learn to disentangle attribute and identity information. Then, the identity vector is perturbed with Laplacian noise and, finally, it is passed to the generator, along with the attributes vector to obtain the de-identified image. In the absence of any publicly available implementation of [44], we do not provide quantitative comparisons with [44]. We can, however, comment on their qualitative results, stating that, indeed, the facial attributes are maintained, but the resulting image can still be recognized by a human observer. Our proposed method does not suffer from such a problem, since the controllable similarity allows us to obtain face images of completely different persons that share the same attributes as the original subject.

3. Proposed Method

In this section, we present our method for anonymizing the identity of faces in a given real face dataset by optimizing the representations of the dataset's images in the latent space of a pre-trained StyleGAN2 [19]. More specifically, given a real dataset $\mathcal{X}_R$, we first create a fake dataset $\mathcal{X}_F$ by randomly generating a large set (i.e., such that $|\mathcal{X}_F| > |\mathcal{X}_R|$) of fake images and obtaining the corresponding latent codes in the $\mathcal{W}^+$ space of StyleGAN2, namely, $\mathcal{W}^+_F$. Additionally, we obtain the latent codes of the real dataset in the $\mathcal{W}^+$ space by inverting its images using e4e [41], arriving at a set of latent codes $\mathcal{W}^+_R$. In order to obtain meaningful initial values for the latent codes that will be optimized to create the anonymized version of the real dataset, namely $\mathcal{X}_A$, we first pair the real images from the original set (i.e., $\mathcal{X}_R$) with fake ones from the generated dataset (i.e., $\mathcal{X}_F$) in the feature space of the ViT-based FaRL [48] image encoder and use their latent codes for initializing the aforementioned trainable codes. The latent codes of the anonymized dataset are then optimized under the following objectives via two novel loss functions: (a) to be similar to the corresponding real ones, up to a certain margin, using the proposed iden-

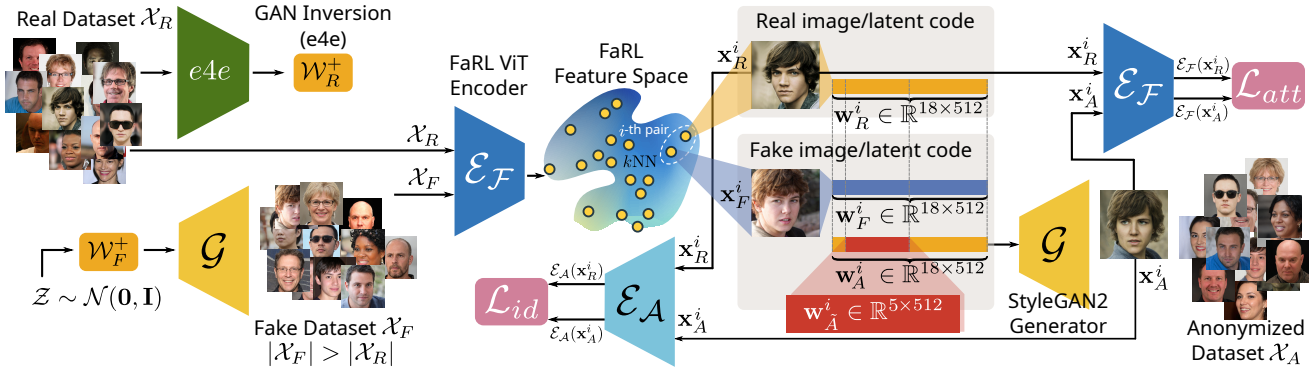

Figure 2. Overview of the proposed method: optimizing the trainable portion of the latent code $\mathbf{w}_{\tilde{A}}^i \in \mathbb{R}^{5\times512}$ to obfuscate the identity of the resulting synthetic image $\mathbf{x}_A^i$ with $\mathcal{L}_{id}$ whilst preserving the facial attributes with $\mathcal{L}_{att}$.

tity loss ($\mathcal{L}_{id}$), and (b) to preserve the facial attributes of the corresponding real ones by being pulled closer in the feature space of the pre-trained FaRL [48] image encoder using the proposed attribute preservation loss ($\mathcal{L}_{att}$). In this way, in contrast to state-of-the-art works [16, 27], the anonymized images are optimized to inherit the labels of the original ones. An overview of the proposed method is given in Fig. 2.

The rest of this section is organized as follows: in Sect. 3.1 we briefly introduce the pre-trained modules of our framework, in Sect. 3.2 we discuss the initialization of the latent codes that will generate the anonymized version of the real dataset, and in Sect. 3.3 we present the proposed optimization process and losses.

3.1. Background

StyleGAN2 We use the StyleGAN2's [19] generator $\mathcal{G}$, pre-trained on the FFHQ [19] dataset and, in particular, its $\mathcal{W}^+$ latent space. In this case, we operate on the latent codes $\mathbf{w} \in \mathbb{R}^{18\times512}$, where the first 9 layers are responsible for coarse- and medium-grained attributes (such as the head pose and facial texture details), while the rest correspond to more fine-grained attributes (such as the hair colour, or the skin tone, as first identified in [18]).

e4e For the inversion of real images onto the $\mathcal{W}^+$ space of StyleGAN2, we use e4e [41], which has been trained in order to preserve a good trade-off between fidelity of the inversion and editability in $\mathcal{W}^+$.

ArcFace For measuring the similarity of the identities of two face images, we use ArcFace [9], which represents images in a 512-dimensional identity-related feature space using which we optimize the GAN latent codes to generate images to maximize the cosine similarity between features corresponding to the same face identity.

FaRL For the representation of images in a meaningful and rich semantic feature space, we use FaRL [48], a universal facial representation scheme trained in a contrastive manner in 20 million face images-text pairs. Specifically, we use the ViT image encoder of the FaRL framework in order to represent images in a 512-dimensional feature space and find meaningful initial values for the latent codes that will be optimized to anonymize the real dataset.

3.2. Fake dataset generation and pairing with real images

Given a dataset $\mathcal{X}_R$ of real face images, we incorporate the generator $\mathcal{G}$ of a StyleGAN2 [19] pre-trained on the FFHQ [19] dataset, in order to generate a set of fake face images $\mathcal{X}_F$, where $|\mathcal{X}_F| > |\mathcal{X}_R|$. We do so by sampling from the $\mathcal{Z}$ latent space of StyleGAN2, i.e., the Standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and by then obtaining the corresponding $\mathcal{W}^+$ latent codes (using the input MLP of $\mathcal{G}$), i.e., the set $\mathcal{W}_F^+$. At the same time, we calculate the latent representations of the face images in the real dataset $\mathcal{X}_R$ by inverting them using e4e [41]. This assigns $\mathcal{X}_R$ with the set of corresponding $\mathcal{W}^+$ latent codes, i.e., the set $\mathcal{W}_R^+$. This is illustrated in the left part of Fig. 2.

For pairing the real images in $\mathcal{X}_R$ with fake ones in $\mathcal{X}_F$ we use the pre-trained FaRL [48] ViT-based image encoder $\mathcal{E}_\mathcal{F}$ and we represent all images of each dataset using the class (i.e., CLS) token representation, i.e., in a 512-dimensional features space. By doing so, we obtain a powerful feature representation of both datasets, which we subsequently use in order to train a kNN classifier and obtain, for each real image, the closest fake one in terms of the Euclidean distance. More formally, after the aforementioned process of generation and pairing, the images/latent codes of the real dataset $\mathcal{X}_R$ are paired with images/latent codes in the fake dataset $\mathcal{X}_F$ forming the following set of pairs

$$\{\big((\mathbf{x}_R^i, \mathbf{w}_R^i), (\mathbf{x}_F^i, \mathbf{w}_F^i)\big) : \mathbf{x}_R^i \in \mathcal{X}_R, \mathbf{w}_R^i \in \mathcal{W}_R^+, \quad (1)$$
$$\mathbf{x}_F^i \in \mathcal{X}_F, \mathbf{w}_F^i \in \mathcal{W}_F^+,$$
$$i = 1, \ldots, |\mathcal{X}_R|\}.$$

In order to initialize the latent codes that will be opti-

mized to anonymize the real images, we use the above pairs of real-fake latent codes as follows. Given the i-th real image, we first modify the latent code of the corresponding fake one (i.e., its nearest neighbor), $\mathbf{w}_F^i \in \mathbb{R}^{18\times512}$, and replace layers 3-7 with a trainable vector $\mathbf{w}_{\tilde{A}}^i \in \mathbb{R}^{5\times512}$, while we set its first three layers (i.e., layers 0-2) and the last layers (i.e., layers 8-17) equal to the corresponding layers of the real latent code $\mathbf{w}_R^i$. By doing so, we arrive at a latent code $\mathbf{w}_A^i \in \mathbb{R}^{18\times512}$, initialized so as a) we retain information that is crucial for generating anonymized face images with head pose and other coarse geometric details same as the corresponding real ones (layers 0-2), b) we maintain the color distribution and background information (layers 8-17) of the real ones, and c) optimize information that is critical for the identity characteristics of a face (layers 3-7). This is illustrated in the centre part of Fig. 2.

3.3. Latent code optimization

In order to create an anonymized version $\mathcal{X}_A$ of the real dataset $\mathcal{X}_R$, we use the pairs of real-fake images obtained and initialized by the process discussed in the previous section and shown in (1), i.e., pairs of real and fake images that are semantically close to each other in terms of the FaRL image representation scheme. More specifically, the real image of each pair, $\mathbf{x}_R^i$, along with the corresponding anonymized image, $\mathbf{x}_A^i$, generated by the modified latent code, $\mathbf{w}_A^i$, are used for calculating the proposed losses. That is, the identity loss $\mathcal{L}_{id}(\mathbf{x}_A^i, \mathbf{x}_R^i)$ so as $\mathbf{x}_A^i$ retains a similar identity to $\mathbf{x}_R^i$, up to a desired margin, and the attribute preservation loss $\mathcal{L}_{att}(\mathbf{x}_A^i, \mathbf{x}_R^i)$ that imposes that the facial attributes of the original image are preserved in the anonymized image.

Given a pair consisting of a real image $\mathbf{x}_R^i$ and its anonymized version $\mathbf{x}_A^i$, we estimate the learnable parts of its latent code $\mathbf{w}_{\tilde{A}}^i \in \mathbb{R}^{5\times512}$, for $i = 1, \ldots, |\mathcal{X}_R|$ by optimizing the following losses:

Identity loss The identity loss is defined as follows

$$\mathcal{L}_{id}(\mathbf{x}_A^i, \mathbf{x}_R^i) = \left|\cos\left(\mathcal{E}_{\mathcal{A}}(\mathbf{x}_A^i), \mathcal{E}_{\mathcal{A}}(\mathbf{x}_R^i)\right) - m\right|, \quad (2)$$

where $\cos(\cdot,\cdot)$ denotes the cosine distance, $\mathcal{E}_{\mathcal{A}}$ denotes the ArcFace [9] identity encoder, and m denotes a hyperparameter that controls the dissimilarity between the real and the anonymized face images. When $m = 0$, the proposed identity loss imposes orthogonality between the features of the real and the anonymized face images, leading to anonymized faces with large identity difference compared to the corresponding real ones. By contrast, when $m = 1$, the proposed identity loss imposes high similarity between the features of the real and the anonymized face images. That is, the hyperparameter m controls the trade-off between data utility and privacy preservation.

Attribute preservation loss The attribute preservation loss is defined as follows

$$\mathcal{L}_{att}(\mathbf{x}_A^i, \mathbf{x}_R^i) = \left\|\mathcal{E}_{\mathcal{F}}(\mathbf{x}_A^i) - \mathcal{E}_{\mathcal{F}}(\mathbf{x}_R^i)\right\|_1, \quad (3)$$

where $\mathcal{E}_{\mathcal{F}}$ denotes the FaRL [48] ViT-based image encoder. It is worth noting that we found empirically that using the patch-level features of the ViT (i.e., the 14×14 512-dimensional features, flattened as $14 \cdot 14 \cdot 512$-dimensional vectors, leads to better attribute preservation than using the features at the class (CLS) token. We argue that maintaining the raw representation allows for better results compared to using only the class token, as this encodes a class contextual representation of the image, while the untouched patches' features contain a higher degree of information.

4. Experiments

In this section we evaluate the performance of our anonymization framework against other state-of-the-art anonymization works, evaluating our results on privacy-related metrics in Sect. 4.2.1, and – in contrast to other works – attribute classification metrics in Sect. 4.2.2. Finally, we show in Sect. 4.3 the impact of the identity loss margin involved in our method through an ablation study.

Datasets We perform anonymization on the following datasets: (i) **CelebA-HQ** [25], which contains 30000 1024×1024 face images of celebrities from the CelebA dataset with various demographic attributes (e.g., age, gender, race) and where each image is annotated with 40 attribute labels related to the inner and outer regions of the face, and (ii) **LFW** [15], which contains over 13000 images collected from the Web (5749 identities with 1680 of those identities being pictured in at least 2 images).

State of the art We compare our anonymization framework with two state-of-the-art anonymization methods, namely CIAGAN [27] and DeepPrivacy [16].

4.1. Evaluation metrics

We evaluate our method by quantifying privacy preservation, image quality, and attribute preservation. We briefly introduce the metrics we use below:

Image quality and identity anonymization We quantify the ability to anonymize images by measuring the "re-identification rate" (defined as the number of images whose identity is still detected in the anonymized version, over the total number of images) using FaceNet [37], pre-trained on two large-scale face datasets (CASIA WebFace [46] and VGGFace2 [4]). Moreover, we measure the "detection rate" as the number of anonymized images for which a valid

face is successfully detected over the total number of images in the dataset. By quantifying how recognisable a face is to a machine learning algorithm, this metric is an important measure of the quality of the facial image. To measure this, we use the MTCNN [47] face detector. An ideal anonymization method would retain a valid face in all anonymized images (100% detection rate), but anonymize all the particular identities (0% re-identification rate). Finally, we report the Fréchet Inception Distance (FID) [14] for all generated images as a measure of the quality of the anonymized datasets.

Attribute preservation Unlike other works [16, 27], we propose a protocol to quantify how well each method can retain the attributes of the original images. More specifically, the evaluation is posed as a standard classification task and the metric used to quantify this is the accuracy of classifiers on the real test sets when trained on the anonymized training set. In this way, one can quantify how well the anonymized training data has retained the original attributes in the images. The train/test split structure followed is the one provided by the official CelebA dataset in the case of CelebA-HQ [25], while the images from LFW [15] are randomly shuffled and then split with an 80-20 ratio. We use a MobileNetV2 [36] to perform multi-label classification, trained with a focal loss [23] to handle class imbalance.

4.2. Comparison to state-of-the-art (SOTA)

In this section, we report the evaluation performance of our method compared to two other SOTA methods (CIAGAN [27] and DeepPrivacy [16]) using the evaluation metrics introduced earlier. Finally, in Sect. 4.2.3 we conduct a qualitative comparison to the SOTA.

4.2.1 Image quality and De-identification

In Tables 1, 2 we show the results for FID, face detection, and face re-identification for the two considered datasets. We see that our method excels at producing the most realistic-looking images under the FID metric for CelebA-HQ in Tab. 1, and also outperforms the baselines for the FID metric on LFW [15] in Tab. 2 when considering the CelebA-HQ [25] dataset as the "target" distribution*. We argue this success is due to the way in which we design our method to operate in the latent space of a well-trained GAN, capable of producing high-resolution, sharp images. On the other hand, the existing SOTA involves techniques such as image inpainting, which we find have a tendency to introduce small artifacts in the anonymization procedure. The realism of our generated images is further attested to by the perfect face detection scores in Tables 1, 2 – indicating the images contain recognisable faces readily usable for downstream machine learning tasks.

However, our images are not just of high quality, but also successfully anonymize the identity–we also see from the last columns of Tables 1, 2 that our anonymization results are competitive with the SOTA. However, it is important to note that whilst the baselines excel under this metric, they fail to preserve the attributes to the extent of our method, which we detail in the next section.

*Given that CelebA-HQ is of much higher quality than LFW, we report both cases to demonstrate that our images can better match the distribution of high-resolution data.

	FID↓	Detection↑		Face re-ID↓	
		dlib	MTCNN(%)	CASIA(%)	VGG(%)
Randomly generated	18.09	100	99.99	3.61	1.08
CIAGAN [27]	37.94	95.10	99.82	**2.19**	**0.37**
DeepPrivacy [16]	32.99	92.82	99.85	3.61	1.05
Ours (ID)	44.12	98.58	97.99	3.28	0.58
Ours (ID+attributes)	44.11	100	**100**	3.06	2.06
Ours	**29.93**	100	**100**	2.80	1.67

Table 1. CelebA-HQ [25] privacy and image quality results.

	FID↓	FID (C-HQ)↓	Detection↑		Face re-ID↓	
			dlib	MTCNN(%)	CASIA(%)	VGG(%)
CIAGAN [27]	**22.07**	85.23	98.14	99.89	**0.17**	**0.91**
DeepPrivacy [16]	23.46	123.67	96.70	99.57	2.74	1.52
Ours	27.45	**68.88**	100	**100**	2.07	1.58

Table 2. LFW [15] privacy and image quality results.

4.2.2 Attribute preservation

In this section we quantify the attribute preservation of the anonymization methods:

CelebA-HQ For CelebA-HQ [25], which provides images annotated according to 40 facial attributes, we first train a MobileNetV2 [36] on the anonymized training sets to predict the attributes of the images, and evaluate its performance on the untouched test set–as a proxy measure for how well the anonymized images have retained the original expected facial attribute labels. Tab. 3 shows the performances of our framework compared to the other methods and also when training using the original dataset.

As can be seen, our method's images result in a classifier capable of almost the same accuracy as when training on the original labels, demonstrating the ability of our method to retain the original facial features. Whilst the other two baselines also produce reasonable results under this *combined* accuracy metric, we argue this is because of the way in which they preserve the image outside the region of the inner face of the images – out of 40 attributes, 17 correspond to the "outer face" region. As shown in Tab. 5 with the accuracy breakdown for the individual attributes, the face inpainting methods excel at preserving the "outer face" attributes as expected, whereas we often outperform the baselines for attributes related to the "inner" region of the face, such as "eyeglasses" or "smile".

	Inner face	Outer face	Combined
Original	0.8409	0.8683	0.8539
CIAGAN [27]	0.7277	0.8372	0.7852
DeepPrivacy [16]	0.7658	0.8511	0.8135
Ours	**0.7817**	**0.8518**	**0.8181**

Table 3. Attribute classification results on CelebA-HQ [25].

	FID	Detection	Face re-ID		Accuracy
		MTCNN(%)	CASIA(%)	VGG(%)	
Ours (m=.0)	29.93	**100**	**2.80**	**1.67**	0.8181
Ours (m=.9)	**27.58**	**100**	3.41	1.76	**0.83**

Table 4. Ablation study on the margin m on CelebA-HQ [25].

LFW Since no official annotations regarding facial attributes are provided for the LFW [15] dataset, we instead use two classifiers [17, 22] pretrained on CelebA-HQ [25]. The model of [22] predicts all the 40 attributes officially provided by CelebA, while [17] predicts only 5 of them, namely *Bangs, Eyeglasses, No_Beard, Smiling* and *Young* (more details on these two classifiers can be found in the supplementary material). For the original LFW [15] dataset, we first predict "pseudo-labels" to approximate the ground-truth facial attribute labels using these two models, and then proceed with the same classification procedure as above. As we can see from the first column of Tab. 6, the accuracy results when training on pseudo-labels on CelebA-HQ [25] are close to those using the real labels in Tab. 3, validating the reliability of the classifiers to generate accurate pseudo-labels. Furthermore, we see in the last two columns of Tab. 6 that our method is able to generate images that much better preserve the facial attributes of the original images than the existing SOTA anonymization methods, through being able to train more accurate attribute classifiers.

4.2.3 Qualitative evaluation

In this section, we make a qualitative comparison to our method and the SOTA. We show in Figs. 3, 4 the original and anonymized images from the various methods on the two studied datasets. As can be clearly seen, our method is capable of retaining the facial attributes of the image to a much greater extent than the baselines – [16] often changes the expression and [27] often modifies the makeup of the images. Crucially, our method also succeeds in removing the background (and identifiable traces of the original image, such as particular clothing choices), which we have argued is of vital importance to true anonymization–removing any ability to infer the original image from the anonymized counterpart. More qualitative results can be found in the supplementary material.

	CIAGAN [27]	**DeepPrivacy** [16]	**Ours**
Outer face region			
Bald	**0.9778**	0.9772	0.9769
Bangs	0.8127	**0.85**	0.8241
Black_Hair	**0.7927**	0.7794	0.7864
Blond_Hair	0.8497	**0.8708**	0.8707
Brown_Hair	**0.7626**	0.7615	0.7593
Double_Chin	**0.9377**	0.9362	0.9364
Gray_Hair	**0.9603**	0.9569	0.9587
Receding_Hairline	**0.9168**	0.9126	0.9117
Sideburns	0.9197	**0.9228**	0.9186
Straight_Hair	0.53	**0.7738**	0.7702
Wavy_Hair	0.6433	0.6603	**0.6652**
Wearing_Earrings	0.6972	0.6721	**0.7123**
Wearing_Hat	0.9636	**0.9641**	0.9595
Wearing_Necklace	**0.822**	0.8017	0.81
Wearing_Necktie	**0.9288**	0.9281	0.9273
Oval_Face	**0.7938**	0.7796	0.7783
Chubby	**0.9247**	0.922	0.9153
Inner face region			
5_o_Clock_Shadow	0.8579	0.8604	**0.8711**
Arched_Eyebrows	0.6057	0.658	**0.6684**
Bags_Under_Eyes	0.6946	0.7156	**0.7158**
Big_Lips	0.6167	0.5901	**0.6194**
Big_Nose	0.6814	**0.7228**	0.7182
Bushy_Eyebrows	0.774	**0.8288**	0.8267
Eyeglasses	0.9483	**0.9622**	0.9564
Goatee	0.9284	0.9289	**0.9303**
Heavy_Makeup	0.6197	**0.7492**	0.6859
High_Cheekbones	0.5356	0.668	**0.6729**
Male	0.6891	0.7917	**0.8381**
Mouth_Slightly_Open	0.5722	0.603	**0.6305**
Mustache	0.9398	**0.9401**	0.9323
Narrow_Eyes	**0.8925**	0.8839	0.883
No_Beard	0.5359	0.5571	**0.7615**
Pale_Skin	0.9418	0.9446	**0.9451**
Pointy_Nose	0.6239	0.6291	**0.6689**
Rosy_Cheeks	**0.8826**	0.8553	0.8825
Smiling	0.5607	0.6505	**0.6666**
Wearing_Lipstick	0.6234	**0.7721**	0.7579
Young	0.7583	0.7706	**0.7848**

Table 5. Accuracy of attributes (inner and outer face regions).

4.3. Ablation study

In this section, we perform an ablation study on the value of the margin m that controls the similarity between the identities. Concretely, we perform the same anonymization procedure for 50 epochs for the whole CelebA-HQ [25] dataset by changing only the value of m. In particular, we consider two extremes of $m = 0.0$ and $m = 0.9$. The larger the value of m, the more the resulting identity ought to be close to the original, i.e., re-identification results should be worse-off, while facial attributes should be better preserved.

We see from the results in Tab. 4 that m indeed offers this trade-off, with the higher value of m offering better attribute preservation at the cost to slightly worse identity re-identification performance. As one expects, we also see a lower value of FID using the higher margin, given the image is encouraged to be more close to the original.

	CelebA-HQ (labels from [22])	LFW (labels from [22])	LFW (labels from [17])
CIAGAN [27]	0.7721	0.9143	0.7045
DeepPrivacy [16]	0.7902	0.9133	0.7019
Ours	**0.8215**	**0.9157**	**0.7209**

Table 6. Accuracy on CelebA-HQ [25] and LFW [15] of anonymized faces using the pseudo-labels generated by the classifiers of [17, 22].

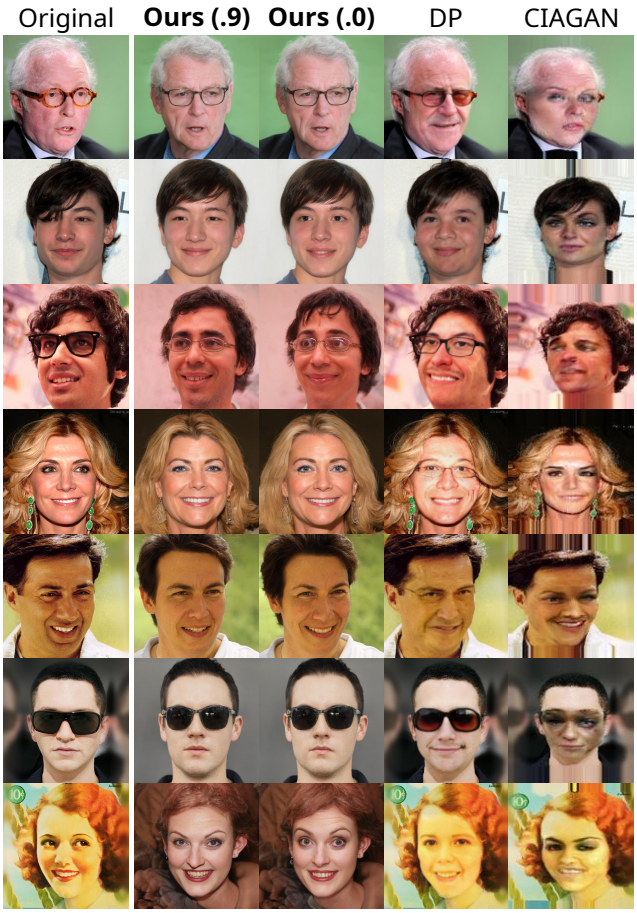

Figure 3. Anonymization results on CelebA-HQ [25] in comparison to DeepPrivacy (DP) [16] and CIAGAN [27].

5. Conclusions

In this paper, we presented a novel anonymization framework that directly optimizes the images' latent representation in the latent space of a pre-trained GAN, using a novel margin-based identity loss and an attribute preservation loss. Our method acts directly in the latent space of pre-trained GANs, avoiding the burden of the need to train complex networks. We showed that our method is capable of anonymizing the identity of the images whilst better-preserving the facial attributes, leading to better de-identification and facial attribute preservation than SOTA.

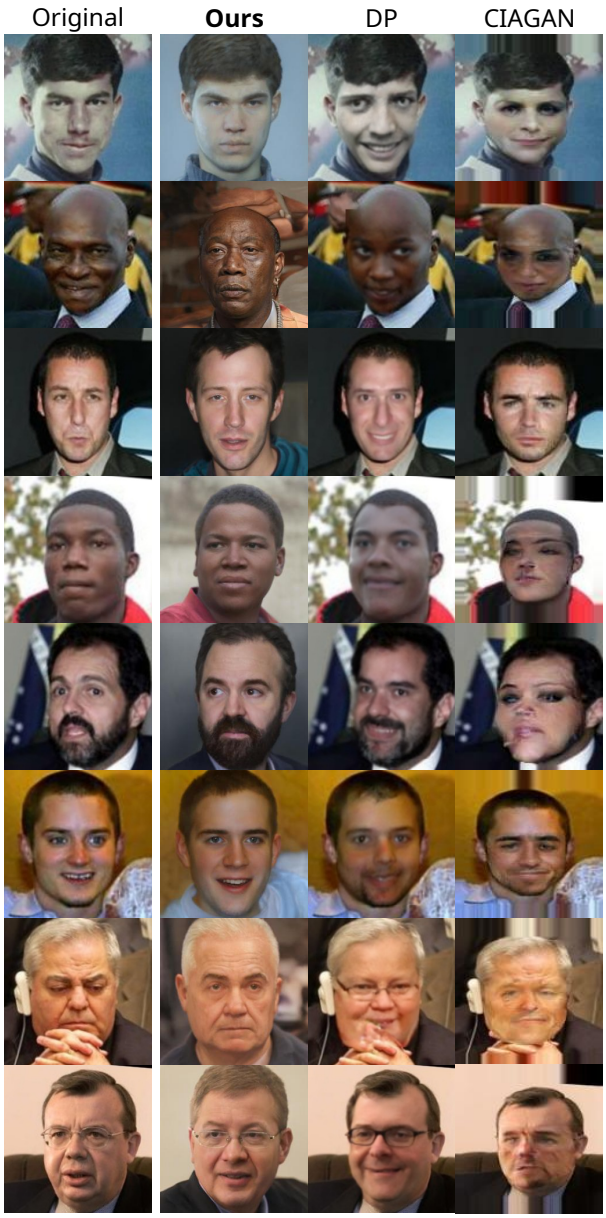

Figure 4. Anonymization results on LFW [15] in comparison to DeepPrivacy (DP) [16] and CIAGAN [27].

Acknowledgments: This work was supported by the EU H2020 AI4Media No. 951911 project.

References

[1] Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*, 12(4):949–961, 2021. 1

[2] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment. *arXiv preprint arXiv:2209.13375*, 2022. 2

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

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2017. 5

[5] Datong Chen, Yi Chang, Rong Yan, and Jie Yang. Tools for protecting the privacy of specific individuals in video. *EURASIP Journal on Advances in Signal Processing*, 2007:1–9, 2007. 2

[6] Durkhyun Cho, Jin Han Lee, and Il Hong Suh. Cleanir: Controllable attribute-preserving natural identity remover. *Applied Sciences*, 10(3), 2020. 3

[7] Bart Custers, Alan M Sears, Francien Dechesne, Ilina Georgieva, Tommaso Tani, and Simone Van der Hof. *EU personal data protection in policy and practice*. Springer, 2019. 1

[8] Nicola Dall'Asen, Yiming Wang, Hao Tang, Luca Zanella, and Elisa Ricci. Graph-based generative face anonymisation with pose preservation, 2021. 3

[9] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Cotsia, and Stefanos P Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2, 4, 5

[10] Ling Du, Wei Zhang, Huazhu Fu, Wenqi Ren, and Xinpeng Zhang. An efficient privacy protection scheme for data security in video surveillance. *Journal of Visual Communication and Image Representation*, 59:347–362, 2019. 2

[11] Niki Maria Foteinopoulou and Ioannis Patras. Learning from label relationships in human affect. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 80–89, New York, NY, USA, 2022. Association for Computing Machinery. 2

[12] Timothy Gerstner, Doug DeCarlo, Marc Alexa, Adam Finkelstein, Yotam Gingold, and Andrew Nealen. Pixelated image abstraction with integrated user constraints. *Computers & Graphics*, 37(5):333–347, 2013. 2

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2, 3

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2, 5, 6, 7, 8

[16] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[17] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog, 2021. 7, 8

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CVPR*, Jun 2019. 1, 4

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020. 3, 4

[20] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019. 2

[21] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy, 2019. 2, 3

[22] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing, 2021. 7, 8

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017. 6

[24] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 1

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 5, 6, 7, 8

[26] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. Cfa-net: Controllable face anonymization network with identity representation manipulation, 2021. 3

[27] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[28] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning, 2016. 2, 3

[29] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 3

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

[31] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. 3

[32] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele. Faceless person recognition; privacy implications in social media, 2016. 2, 3

[33] James Oldfield, Markos Georgopoulos, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Tensor component analysis for interpreting the latent space of gans. *arXiv preprint arXiv:2111.11736*, 2021. 2

[34] James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis A Nicolaou, and Ioannis Patras. Panda: Unsupervised learning of parts and appearances in the feature maps of gans. *arXiv preprint arXiv:2206.00048*, 2022. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015. 5

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 1

[39] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Security Symposium*, 2020. 2, 3

[40] Suriyon Tansuriyavong and Shin-ichi Hanaki. Privacy protection by concealing persons in circumstantial video image. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, PUI '01, page 1–4, New York, NY, USA, 2001. Association for Computing Machinery. 2

[41] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021. 3, 4

[42] Christos Tzelepis, James Oldfield, Georgios Tzimiropoulos, and Ioannis Patras. Contraclip: Interpretable gan generation driven by pairs of contrasting sentences. *arXiv preprint arXiv:2206.02104*, 2022. 2

[43] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. Warpedganspace: Finding non-linear rbf paths in gan latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6393–6402, 2021. 2

[44] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. Identitydp: Differential private identification protection for face images. *Neurocomputing*, 501:197–211, 2022. 2, 3

[45] Yifan Wu, Fan Yang, and Haibin Ling. Privacy-protective-gan for face de-identification, 2018. 2, 3

[46] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch, 2014. 5

[47] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 6

[48] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner, 2021. 1, 2, 3, 4, 5