IdentityDP: Differential Private Identification Protection for Face Images

Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, Li Song

***Abstract*—Because of the explosive growth of face photos as well as their widespread dissemination and easy accessibility in social media, the security and privacy of personal identity information becomes an unprecedented challenge. Meanwhile, the convenience brought by advanced identity-agnostic computer vision technologies is attractive. Therefore, it is important to use face images while taking careful consideration in protecting people's identities. Given a face image, face de-identification, also known as face anonymization, refers to generating another image with similar appearance and the same background, while the real identity is hidden. Although extensive efforts have been made, existing face de-identification techniques are either insufficient in photo-reality or incapable of well-balancing privacy and utility. In this paper, we focus on tackling these challenges to improve face de-identification. We propose IdentityDP, a face anonymization framework that combines a data-driven deep neural network with a differential privacy (DP) mechanism. This framework encompasses three stages: facial representations disentanglement, ϵ-IdentityDP perturbation and image reconstruction. Our model can effectively obfuscate the identity-related information of faces, preserve significant visual similarity, and generate high-quality images that can be used for identity-agnostic computer vision tasks, such as detection, tracking, etc. Different from the previous methods, we can adjust the balance of privacy and utility through the privacy budget according to pratical demands and provide a diversity of results without pre-annotations. Extensive experiments demonstrate the effectiveness and generalization ability of our proposed anonymization framework.**

***Index Terms*—Face de-identification, face anonymization, differential privacy, generative adversarial networks, privacy protection, utility-privacy tradeoff.**

I. Introduction

TODAY'S popularity of smartphones allows people to take their face photos conveniently. Particularly, the blooming development of media and network techniques makes a vast amount of photos more approachable. At the same time, however, advanced image retrieval and face verification models allow to index and examine privacy relevant information more reliably than ever. Consequently, among those image sources exposed to the public with or without our awareness, the wide range of private information inadvertently leaked is severely under-estimated [1].

Opportunities for misuse of the unprotected face image and advanced computer vision technologies are numerous and potentially disastrous [2]. Restrictive laws and regulations such as the General Data Protection Regulations (GDPR) [3] has taken effect. GDPR requires regular consent from the individual for any use of their personal data to guarantee data privacy, however, it also makes the creation of high-quality datasets that include people becoming extremely challenging. Fortunately, if the data does not allow to identify the corresponding individual, entities are free to use the data without consent. what's more, many computer vision tasks in practice such as detection, tracking, or people counting, do not need to identify the people, but to detect them.

All the troubles and dilemmas mentioned above can be summarized to one issue: given a face image, how can we create another image with similar appearance and the same background, while the real identity is hidden and face detectors are still allowed to work? Traditional anonymization techniques are mainly obfuscation-based and always significantly alter the original face. Other previous work in this field is sparse and limited in both practicality and efficacy: *k*-same algorithm-based methods [4–8] fail to make full use of existing data and deliver fairly poor visual quality; adversarial perturbation-based methods [9–14] usually depend highly on the accessibility of the target system and require special training; recent GAN-based methods[15–25] have trouble generating visually similar de-identified faces as well. Note that there exists a trade-off between privacy protection and dataset utility [26, 27], and previous methods are unable to balance this matter.

To tackle these challenges, we propose IdentityDP, a framework that anonymize face images without significantly distorting the original images, nor destroying the availability of face detectors (see Fig. 1). Especially, individuals are allowed to have control over the anonymization procedure to get the most suitable results in practice. IdentityDP achieves this by helping users adding well-designed obfuscation to photos' high-level identity representations. For example, a user who wants to share photos on social media or the public web can add adjustable perturbations according to his demands through our framework before uploading them. The uploaded photos will look similar to the original ones, but when an adversary employs a general face verificator to compare the user's face images with the altered ones, it will indicate that they are from different people.

The proposed IdentityDP framework consists of three stages. Stage-I aims to perform facial representations disentanglement. We train a specially designed GAN for disentanglement between high-level identity representation and multi-level attribute representations in the feature space. Here

Y. Wen, L. Song and R. Xie are with The Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: wenyunqian@sjtu.edu.cn; song_li@sjtu.edu.cn; xierong@sjtu.edu.cn).

B. Liu is with School of Computer Science, University of Technology Sydney, NSW 2007, Australia (email:bo.liu@uts.edu.au).

M. Ding is with Data61, Sydney, NSW, 1435 Australia (email: ming.ding@data61.csiro.au)

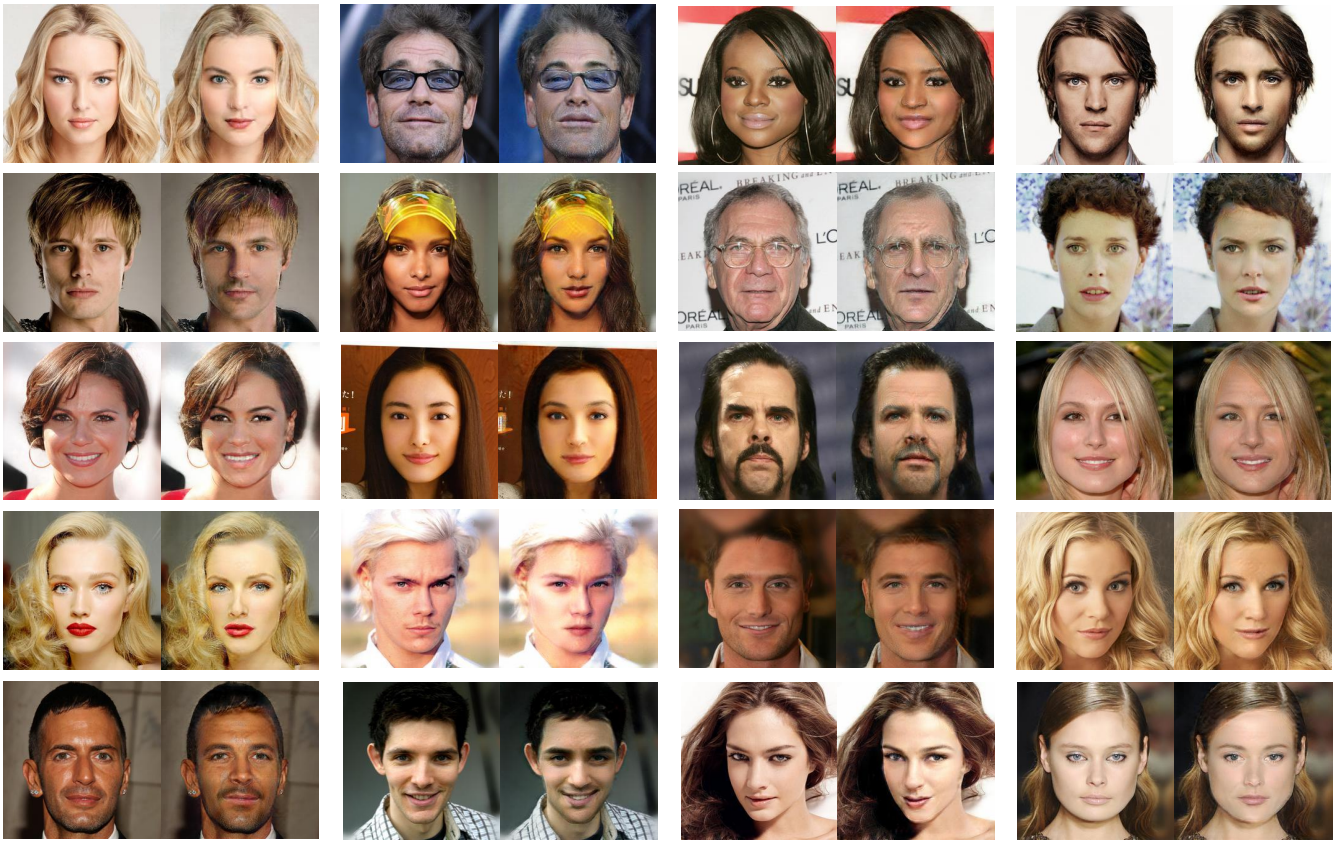

Fig. 1. IdentityDP for face anonymization. In each pair, left is the original image and right is the synthesized result with an altered identity. The results show that face identities are changed in a perceptually natural manner, and in the meantime, each pair of images still shares most of the non-identity related information.

the identity representation affects face verification systems to judge whether it is the same person, and the attribute representation guarantees the visual similarity. Stage-II carries out an ϵ-IdentityDP mechanism, where adjustable differential privacy (DP) [28] perturbations are applied to the identity representation. Stage-III implements the image reconstruction. In more detail, we fix the well-trained GAN network in Stage-I, and generate de-identified face images utilizing the perturbed identity representation as well as the original attribute representations. IdentityDP leverages both the GAN's outstanding ability to disentangle images' representations in the latent space and differential privacy theory, managing to balance the trade-off between image quality and privacy protection according to practical needs. In addition, our framework requires neither pre-annotation nor pre-detection of faces, but can generate numerous anonymous results.

Our contributions in this work are as follows:

- We propose a general framework that is suitable for the de-identification of people in face images.
- As far as we know, we are the first to introduce the rigorously formulated DP theory into the face-anonymous task. The users are able to get not only high-quality anonymous images but also an adjustable privacy protection mechanism.
- We demonstrate that our method does not require special training or targeted adjustments for many unauthorized identity verification systems or face datasets that never seen before.
- We show that images anonymized by our method can be detected by common face detection models, so the processed images are still usable for identity-agnostic computer vision tasks (such as monitoring and tracking).
- We show that our de-identified method is significantly less computationally complex and consumes a small amount of computing resources.

The remainder of this paper is organized as follows. In Section II, we summarize related work. Section III formalizes the face de-identification problem, introduces relevant DP theory and proposes our assumptions. Section IV outlines the three-stages IdentityDP framework. Results of experiments analysing the proposed IdentityDP method and comparisons with existing methods are reported in Section V. we conclude in Section VI with discussions of future research direction.

II. Related Work

In this section, we introduce the related work on face de-identification. We classify face anonymous methods into four categories: traditional obfuscation-based methods, *k*-same algorithm-based methods, adversarial perturbation-based methods and GAN-based methods.

A. Traditional Obfuscation-Based Methods

In traditional computer vision studies, face de-identification technologies are mainly obfuscation-based. To be more specific, individuals can obfuscate privacy sensitive face area in an image by using approaches including blurring, pixelation, masking and so on. These traditional methods are widely used in daily life because of their simplicity and ease of operation. However, researchers have shown these techniques are vulnerable, and the private information in obfuscated images is still in danger of being leaked [29]. McPherson *et al.* [30] showed that deep learning methods especially CNN-based recognition models can successfully identify faces in images encrypted with these techniques with high accuracy. To make matters worse, obfuscation-based approaches towards manipulating images always tend to destroying the usability of images. Vishwamitra *et al.* [31] indicated that both blurring and blocking would impact image perception scores, and even lower scores were observed for images obfuscated by blocking. Moreover, how to conduct a sufficient blur itself is non-trivial [32].

B. k-Same Algorithm-Based Methods

To improve the performance of traditional methods, Newton *et al.* [4] introduced the first privacy-enabling algorithm, *k*-same [33], to the context of image databases. By applying the *k*-same algorithm, a given image is represented by an average face of *k*-closet faces from the gallery. This procedure theoretically limits the performance of recognition to $1/k$, but the resulting images usually suffer from ghosting artifacts due to small alignment errors. Many variants of *k*-same [5–8] were then proposed to improve the data utility and the naturalness of de-identified face images. Although these methods are once a mainstay of anonymous technology, they have notable limitations. Firstly, the *k*-same assumes that each subject is only represented once in the datasets, but this may be violated in practice. The presence of multiple images from the same subject or images sharing similar biometric characteristics can lead to lower levels of privacy protection. Secondly, the k-same operates on a closed set and produces a corresponding de-identified set, which is not applicable in situations that involve processing individual images or sequences of images. Thirdly, their de-identified results always do not look natural enough, let alone resemble the original image. The above limitations indicate that there is still plenty of room for improvement in face de-identification research.

C. Adversarial Perturbation-Based Methods

New techniques and mechanisms are being applied to enhance image obfuscation. A fundamental idea is to generate a small but intentional worst-case disturbance to an original image, which misleads CNN-based recognition models without causing a significant difference perceptible to human eyes. Komkov and Petiushko [9] showed that carefully computed adversarial stickers on a hat could reduce its wearer's likelihood of being recognized. Oh *et al.* [10] introduced a general framework based on game theory to conduct adversarial image perturbations and enforce guarantees on the user's level of privacy. An alternative to evading models is to disrupt their training via a data poisoning attack. Shafahi *et al.* [11] presented an optimization-based method for crafting poison images, in which just one single poison image could control classifier behavior. Liu *et al.* [12] proposed to use adversarial perturbation to protect image privacy from both humans and AI. Zhu *et al.* [13] introduced a new "polytope attack" in which poison images were designed to surround the targeted image in the feature space. Taking both ideas into account, Fawkes [14], the state-of-the-art method, helped users wearing imperceptible "cloaks" to their own photos before releasing them. When used to train facial recognition models, these "cloaked" images produce functional models that consistently cause normal images of the user to be misidentified. Though their obfuscation performances are superb even at imperceptible perturbation level, these methods depend highly upon the accessibility to target systems, so can only be guaranteed for target-specific recognizers. In contrast, we hope to obfuscate identities against general face verification systems, and we are interested in gaining good generalization ability.

D. GAN-Based Methods

GANs represent an inspiring framework for generating sharp and realistic natural face image samples via a minimax game [34]. It has therefore become popular in recent face de-identified techniques, which can be divided into three categories.

Attribute manipulation-based methods. Face attributes are crucial to face identification for human beings, and some methods achieve de-identification by manipulating attributes. Li *et al.* [15] proposed the Privacy-Preserving Attribute Selection (PPAS) algorithm to select and update facial attributes such that the distribution of any attribute was close to its real-life distribution, and provided measurable privacy for face anonymization with privacy guarantees. Wang *et al.* [16] introduced a bi-directional discriminator to alleviate issues of partial inversion of attributes, and executed attribute inversion and obfuscation in a two-stage manner.

Conditional inpainting-based methods. Since face is one of the strongest cues to infer a person's identity, a lot of studies cover up sensitive identity information by conditional inpainting face area. Sun *et al.* [17] generated a realistic head inpainting based on 68 facial keypoints landmarks. Ren *et al.* [18] trained a face modifier to remove privacy-sensitive information, while an action detector was trying to maximize spatial action detection performance. DeepPrivacy [19] directly removed the whole face area and generated new faces based on a sparse pose estimation, which ensured 100% removal of privacy-sensitive information in the original face. Wu *et al.* [20] designed a verificator to help remove biometric information and a regulator to maintain similar image utility. The involvement of these two types of prior knowledge was proved to significantly improve the model performance.

Conditional ID-swapping-based methods. Replacing the identity in a face image with someone else is a direct but effective idea of face anonymization. Meden *et al.* [21] proposed an de-identification pipeline that each generated face

is a combination of k identities. Sun *et al.* [22] explicitly manipulated the identity through identity parameters provided by 3DMM [23]. Gafni *et al.* [24] maximally decorrelated the identity conditioned on the high-level descriptor of a person's facial image, while having the perception (pose, illumination and expression) fixed. CIAGAN [25] leveraged facial landmark and identity one hot-vector to remove the identification characteristics of people, while still keeping necessary features to allow face and body detectors to work.

Although GAN-based methods account for a substantial part of face de-identification study, they suffer from various conditional information requiring either manually annotations or computational resources, not to mention changed expressions, distorted shape, and loss of accessories. In this paper, we introduce a hybrid framework to try to solve the above problems.

III. Preliminaries

A. Problem Formulation

A face de-identification model can be viewed as a transformation function δ that maps a given face image X to a de-identified image $\hat{X}$, aiming to mislead face verification systems. Essentially, we are generating a new fake identity out of the input image. The problem can be formulated as follows:

$$\delta(X) = \hat{X} \tag{1}$$
$$s.t. : \text{Identity}\{X\} \neq \text{Identity}\{\hat{X}\}.$$

Meanwhile, considering image utility, $\hat{X}$ should look similar to X as much as possible and be detectable by general face detectors.

B. Differential Privacy Theory

1) Differential Privacy: Differential Privacy (DP) [28], a cryptography-inspired privacy-preserving model, guarantees that the likelihood of seeing an output on a given original datasets is close to the likelihood of seeing the same output on another datasets that differs from the original one in any single row. Here, the output could be another datasets, a statistical summary table, or a simple answer to a query, etc. Generally speaking, the basic idea of a DP mechanism is to introduce randomness into the original datasets, so that any individuals' information cannot be inferred by an adversary looking at the released output.

A formal definition of DP is shown below:

Definition 1: **(ϵ-DP)** [35]: A randomized mechanism $\mathcal{T}$ gives ϵ-differential privacy if for any neighboring datasets D and D' differing on one element, and all transcripts t:

$$\left|\ln(\frac{Pr[\mathcal{T}(D) = t]}{Pr[\mathcal{T}(D') = t]})\right| \leq \epsilon. \tag{2}$$

This parameter ϵ, which is usually referred to as a privacy budget, is a bound on the ratio of the likelihood probabilities of seeing the same output on neighbouring datasets. The smaller the value of ϵ, the stronger the privacy guarantee.

A random perturbation can be added to achieve the differential privacy. Sensitivity calibrates the amount of noise for a specified query f of dataset D. Δf is the l_1-norm sensitivity defined as

Definition 2: **(l_1-norm sensitivity)** [35]: For any query f: $D \rightarrow \mathbb{R}$, l_1-norm sensitivity is the maximum l_1- norm of $f(D) - f(D')$, i.e.,

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1. \tag{3}$$

The Laplace mechanism is one of the most generic mechanism to guarantee differential privacy [36].

Definition 3: **(Laplace Mechanism)** [35]: Given a function f: $D \rightarrow \mathbb{R}$, the following mechanism $\mathcal{T}$ provides the ϵ-Differential Privacy:

$$\mathcal{T}(D) = f(D) + Lap(\frac{\Delta f}{\epsilon}). \tag{4}$$

2) Local Differential Privacy: In traditional DP setting, there is a trusted curator who applies carefully calibrated random noise to the real values returned for a particular query. However, in many practical scenarios, the curator might not be trustworthy. In the local setting, there is no trusted third party and the data needs to be randomised without the global knowledge. Local differential privacy (LDP) [37–39] is applicable to this case. It is considered to be a strong and rigorous notion of privacy that provides plausible deniability and deemed to be a state-of-the-art approach for privacy-preserving data collection and distribution.

Definition 4: **(ϵ-LDP)** [40]: A randomized mechanism $\mathcal{A}$ satisfies ϵ-LDP, if for any two inputs v, v' and the set of all possible outputs $y \in \mathcal{Y}$, $\mathcal{Y}$ = *Range*($\mathcal{A}$), $\mathcal{A}$ satisfies:

$$Pr[\mathcal{A}(v) = y] \leq e^{\epsilon} \cdot Pr[\mathcal{A}(v') = y]. \tag{5}$$

And the sensitivity in this case equals to

$$\Delta f = \max_{v,v' \in V} ||f(v) - f(v')||_1. \tag{6}$$

3) Two Important Properties: Our approach relies on two key properties of DP. First is the widely used parallel composition property when designing mechanisms:

Property 1: **(Parallel composition)** [41]: Suppose we have a set of privacy mechanisms M = $\{M_1,\ldots,M_m\}$, if each M_i provides ϵ_i privacy guarantee on a disjointed subset of the entire dataset, M will provide (max$\{\epsilon_1,\ldots,\epsilon_m\}$)-differential privacy.

Second is the well-known post-processing property:

Property 2: **(Post-processing property)** [42]: Any computation applied to the output of an (ϵ,δ)-DP algorithm remains (ϵ,δ)-DP.

For example, averaging, rounding or any change to the output will not impact the privacy of the data. This means that an analyst can conduct any data post-processing on a released DP dataset and cannot reduce its privacy guarantee.

C. Face Verification and Our Assumptions

The key idea of face verification is to develop effective representations in feature space for reducing intra-personal variations while enlarging inter-personal differences [43]. The most ideal state is directly learning a mapping from face images to a compact feature space where distances precisely correspond

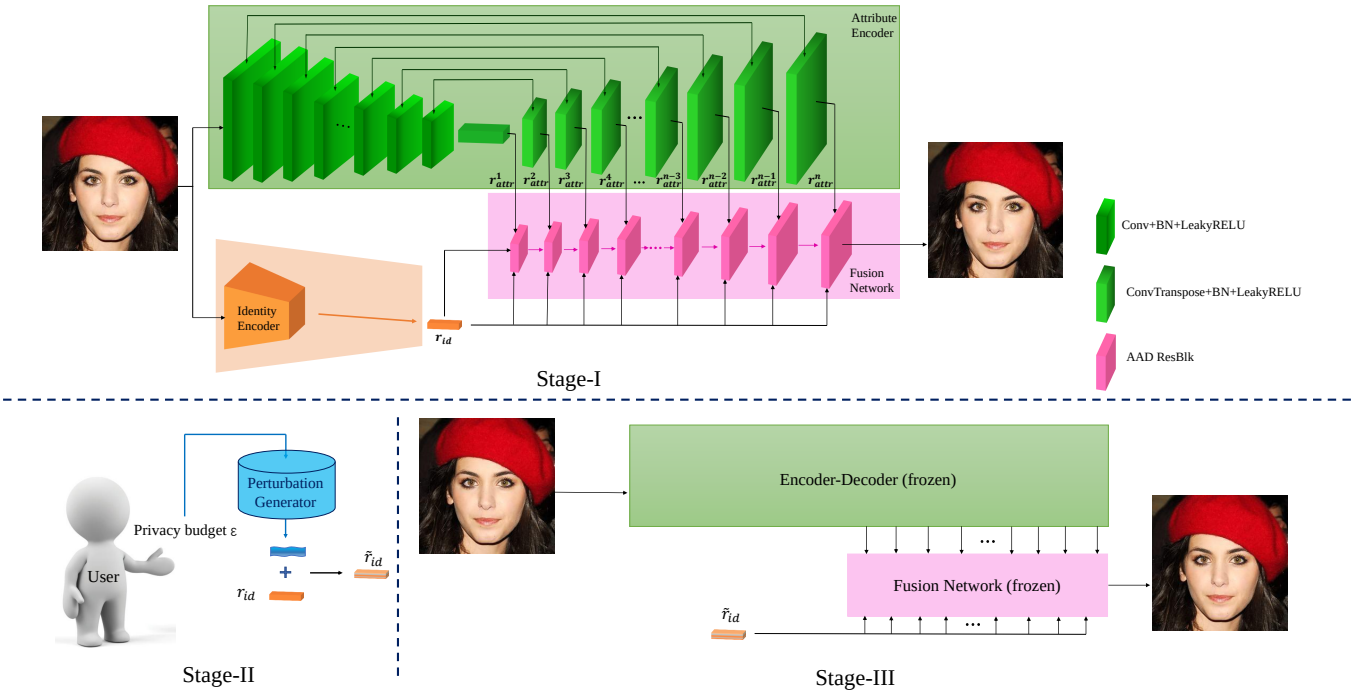

Fig. 2. Architecture of the proposed 3-stages IdentityDP framework, which based on a data-driven deep neural network and a Laplace ϵ-IdentityDP mechanism. Stage-I: training a network to extract the disentangled high-level identity as well as attributes representations and restore the original faces; Stage-II: generating the perturbed identity representation under the Laplace ϵ-IdentityDP mechanism; Stage-III: crafting anonymous faces from perturbed identity representation and original attribute representations through the frozen network.

to a measure of identity similarity. There are currently two main types of solutions: one is metric learning-based, and contrastive loss [44], center loss [45], triplet loss [46] are proposed to enhance the discrimination power of features; the other is angular margin-based, and many efforts [47–50] about angle margin penalty have greatly improved the verification accuracy. To some extent, anonymization can be considered as a task to protect someone's identity representations from being correctly classified.

Here we have an assumption that identity representations of one person in different feature spaces are interrelated. Once a face image's high-level representation in one feature space is disturbed into the wrong identity category, its identity representations in other feature spaces would also be classified incorrectly.

IV. The Proposed IdentityDP Framework

For a given original clean face image X, our proposed IdentityDP framework can be used to generate its anonymous face images $\hat{X}$ in a controllable manner. We factor the face de-identification task into three stages. In the first stage, we use a person's image as input and disentangle the latent space information into two main representations, namely identity and attribute. Among them, identity representation is modeled by embedding features through an encoder, while attribute representations are modeled by multi-level embedding features through a decoder, then the original face image is restored in an adaptively manner. In the second stage, we impose ϵ-IdentityDP perturbations on identity representation according to practical demands. In the third stage, we freeze all the parameters of the network, and reconstruct anonymous face image with the perturbed identity representation. The overall architecture of the IdentityDP framework is shown in Fig. 2.

A. Stage-I: Facial representations disentanglement

In stage I, given an input face image, our goal is to represent the image using two disentangled representations, r_{id} and r_{attr}. r_{id} is expected to contain all the information relevant to the identity, and r_{attr} contains the rest of information carried by the image. We investigate how to generate satisfactory face images with a specific disentanglement intention (i.e. identity and attribute) in mind. The key idea is to explicitly guide the generation process by an appropriate representation of that intention. Therefore, our network consists of 3 components: (1) Identity Encoder; (2) Attribute Encoder; (3) Fusion Generator.

Identity Encoder: As mentioned before, studies on face verification and recognition have made arduous efforts in finding suitable face features that can reduce intra-personal variations while enlarging inter-personal differences, which is in line with our requirement of identity representation. Therefore, we choose a pre-trained state-of-the-art face recognition model [50] as our identity encoder, so as to exploit the existing technology to extract high-level identity representations in latent space. This pre-trained model [50] can provide highly discriminative features for face recognition, and has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. The identity representation $r_{id}(X)$ is defined to be the last feature vector before the final FC layer, which can present off-the-shelf precise facial

identity features to avoid training from scratch and denoted as:

$$r_{id}(X) = f(X). \tag{7}$$

Attribute Encoder: Attribute representation, which determines pose, expression, illumination, background and so on, intuitively carries more spatial information than identity. Johnson *et al.* [51] illustrate that low-level features tend to preserve image content and overall spatial structure, and high-level features tend to preserve color, texture, and exact shape. In order to preserve different level details, we employ multi-level feature maps to represent the attributes. In specific, we feed the input image X into a U-Net-like structure, and then use the feature maps generated from the U-Net decoder as the attributes representations. More formally, we denote

$$r_{att}(X) = g(X) = \left\{r^1_{att}(X), r^2_{att}(X), \cdots, r^n_{att}(X)\right\}, \tag{8}$$

where $r^k_{att}(X)$ represents the *k*-th level feature map from the U-Net decoder, *n* is the number of feature levels.

This attributes encoder does not require any artificial annotations, it extracts the attributes using self-supervised training: we require that the generated de-identified face $\hat{X}$ and the original face X have the same attributes embedding. The loss function will be introduced later in Eq. (16).

Fusion Network: After obtaining the disentangled identity and attribute representations, we would like to learn a way to integrate them to reproduce the original face image, which will be used in our subsequent steps. Through a simple trial, we find that direct feature concatenation can easily lead to blurry results and is not expected to be used. Fortunately, Li *et al.*[52] used *Adaptive Attentional Denormalization* (AAD) ResBlk to achieve remarkable feature integration in multiple feature levels. They argue that the attention mechanism with denormalizations make the effective regions of features more adaptive to adjust; this is an appealing property for fusion network, since identity and attribute representations can participate in synthesizing different parts of the face. We integrate n AAD ResBlks to the body of our fusion network. As illustrated in Fig. 2, in stage-I, after extracting the identity representation r_{id}, and encoding multi-level attribute feature maps r_{att}, the fusion generator integrates them through cascaded AAD ResBlks to restore the original face image X:

$$X = h(r_{id}, r_{att}). \tag{9}$$

The training of h(.) will be discussed in the following sections.

B. Stage-II: ϵ-IdentityDP perturbation

Stage-II generates the perturbed identity representation under a novel Laplace ϵ-IdentityDP mechanism, which is defined as follows:

Definition 5: **(ϵ-IdentityDP Mechanism):** A randomized mechanism $\mathcal{M}$ satisfies ϵ-IdentityDP, i.e. if for any two inputs face images X, X' and the set of all possible outputs $y \in \mathcal{Y}$, $\mathcal{M}$ satisfies: $Pr[\mathcal{M}(X) \in \mathcal{Y}] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(X') \in \mathcal{Y}]$. For a face image X, if:

$$f(X) = r_{id}(X), \tag{10}$$

and

$$\mathcal{M}(X) = r_{id}(X) + Lap(\frac{\Delta f}{\epsilon}) = \tilde{r}_{id}, \tag{11}$$

We say that $\mathcal{M}(X)$ satisfies ϵ-IdentityDP.

And the sensitivity is calculated as follows:

$$\Delta f = \max_{X,X'} ||r_{id}(X) - r_{id}(X')||_1. \tag{12}$$

To achieve ϵ-IdentityDP mechanism, we employ a noise generator to generate suitable Laplace noise whose size equals to the high-level identity representation according to specific privacy budget ϵ. Then we directly add the noise on the identity representation from Stage-I, intending to obfuscate people's identity.

C. Stage-III: Image reconstruction

Stage-III is conditioned on the obfuscated identity representation from Stage-II and the original multi-level attribute features from Stage-I. In order to achieve good de-identified results, we freeze all the parameters of the well-trained fusion network in Stage-I, and generate anonymous face image $\hat{X}$ through the fusion network using obfuscated identity representation and multi-level attribute representations:

$$\hat{X} = h(\mathcal{M}(\mathcal{X}), g(X)) = h(\tilde{r}_{\text{id}}, r_{att}). \tag{13}$$

It can be approved that the generated image $\hat{X}$ follows ϵ-IdeneityDP.

Proof 4.1: First, according to definition in Eq. (11), $\mathcal{M}(\mathcal{X})$ satisfies ϵ-IdentityDP:

$$\begin{aligned}
\frac{Pr(\tilde{r}_{\text{id}}|f(X))}{Pr(\tilde{r}_{\text{id}}|f(X'))} &= \prod_{i=1}^{m} \frac{\exp(-|r_{id(i)} - f(X)_i|/\frac{\Delta f}{\epsilon})}{\exp(-|r_{id(i)} - f(X')_i|/\frac{\Delta f}{\epsilon})} \\
&= \prod_{i=1}^{m} \exp(\frac{\epsilon(|r_{id(i)} - f(X')_i| - |r_{id(i)} - f(X)_i|)}{\Delta f}) \\
&\leq \prod_{i=1}^{m} \exp(\frac{\epsilon|f(X)_i - f(X')_i|}{\Delta f}) \\
&= \exp(\frac{\epsilon \cdot \sum_{i=1}^{m} |f(X)_i - f(X')_i|}{\Delta f}) \\
&= \exp(\frac{\epsilon \cdot \|f(X) - f(X')\|_1}{\Delta f}) \\
&\leq \exp(\epsilon),
\end{aligned}$$

where the first inequality follows from that $|a| - |b| \leq |a - b|$ for any $a, b \in \mathbb{R}$. The rest of proof follows from the post-processing property of DP. Hence, we can conclude that if the identity representation is treated with DP noises, then the reconstructed face image $\hat{X}$ also satisfies the ϵ-IdentityDP defined in Definition 5.

D. Training Process

In Stage-I, we need to build a network which can not only disentangle identity and attribute representations, but also restore the original input face image from these two representations.

We utilize adversarial training for this framework. Let L_{adv} be the adversarial loss for making $\hat{X}$ realistic. It is

implemented as a multi-scale discriminator [53] on the down-sampled output images:

$$L_{adv}(\hat{X}, X) = logD_{img}(X) + log(1 - D_{img}(\hat{X})). \quad (14)$$

An identity preservation loss is used to preserve the identity of the source. It is formulated as:

$$L_{id} = 1 - \cos(r_{id}(\hat{X}), r_{id}(X)), \quad (15)$$

where $\cos(\cdot, \cdot)$ represents the cosine similarity of two vectors. We also use the attributes preservation loss, which is defined as half of the sum of the squared Euclidean distances between the multi-level attributes representations from X and $\hat{X}$. More formally, we denote

$$L_{att} = \frac{1}{2} \sum_{k=1}^{n} \left\| r_{att}^{k}(\hat{X}) - r_{att}^{k}(X) \right\|_2^2. \quad (16)$$

The reconstruction loss as pixel level L-2 distances between the target image $\hat{X}$ and X:

$$L_{rec} = \frac{1}{2} \left\| \hat{X} - X \right\|_2^2. \quad (17)$$

The full objective to train our network in the first stage is a weighted sum of above losses as:

$$L_{total} = L_{adv} + \lambda_{att} L_{att} + \lambda_{id} L_{id} + \lambda_{rec} L_{rec}, \quad (18)$$

where λ_{att}, λ_{id} and λ_{rec} are the weight parameters for balancing different terms.

In practice, GAN is hard to train, so adjusting the training strategy according to real-time generation effect is necessary. In order to use visualization tools to judge our training effect and make appropriate adjustments in time, we extract identity and attribute representations from two faces randomly sampled from the training dataset and then fuse them together in stage-I. It is worth noting the reconstruction loss should be set to $L_{rec} = 0$ when the two faces are different.

E. Some discussions about our research topic

1) The motivation of using differential privacy (DP) for face de-identification.

The reason we need to perform de-identification is that face image is a personal identifier which should be properly protected from the privacy perspective. In more detail, we want to prevent the information leakage of personal identities from releasing face images, and we hope that the privacy protection level can be measured by a formal criterion. Meanwhile, although DP is the most widely used notion for privacy protection, there is no effective and formal DP definition or mechanism in the context of image. This motivates us to use DP to prevent identity information leakage from face images, and we propose the IdentityDP method which makes an initial contribution to this meaningful research topic.

2) Are we just doing adversarial attack-based privacy protection?

Initially, an adversarial attack is perceived as an "attack" method to mislead AI models, i.e., adding small (often human invisible) perturbation to the input data sample so as to corrupt the prediction of a deep learning model. Although there have been a few recent studies [14, 54] that explored the idea of adversarial attack for privacy protection. These methods are distinctively different from our proposed method from the following two aspects:

- Adversarial attack-based privacy protection methods usually assume a machinery adversary, e.g., a deep learning model from previous work. As the adversarial perturbation is often small, the provided protection is not necessarily effective against human eyes. In contrast, our proposed method consider both human and machine as adversaries, and provide effective privacy protection against both types of adversaries.
- There is no formal and strict privacy guarantee provided by the adversarial attack-based privacy protection methods, while the privacy level of our proposed IdentityDP is clearly defined and rigorously guaranteed by the DP criterion.

3) Are we just doing differentially private machine learning?

Researchers in the field of differentially private deep learning [55, 56] are work on preventing model itself from releasing private information of its training datasets, and maintaining a manageable cost in software complexity, training efficiency, and model quality at the same time. However, it is different from our research topic of face de-identification. De-identification is a process which aims to remove all identification information of the person from an image or video, while maintaining as much information on the action and its context with a similar looking appearance [24, 57]. Our concentration is to protect the private identity information of face images, but not to prevent our model from releasing private information of our training face datasets. In more detail, the role of machine learning in this two tasks is different: their topic is to make machine learning system private, i.e., machine learning system is the target of privacy protection; however, our topic is to use machine learning techniques to enhance privacy protection (i.e., prevent the information leakage of personal identifiers from releasing face images). Therefore, these are two different research topics.

4) Recent researches on DP-based face de-identification.

Applying DP in images is a promising research topic because of the increasing concerns on image privacy especially face privacy, and there are a few recent work [58, 59] to study this problem. They all try to implement DP into images, but in different ways. The main idea of these methods is to inject DP noise in the whole feature (latent) space. The disadvantage is that the photo's quality is very sensitive to the amount of noise, and even a small noise perturbation (large epsilon value) will make the photo distorted. Our work solves this problem by only adding noise to the disentangled identity representation. The essential point of our proposed method is that the noise needed for de-identification is much smaller than the existing methods, as the disentangled identity vector has a much smaller norm than the whole latent space vector. In addition, Laplace mechanism is the most often used mechanism to

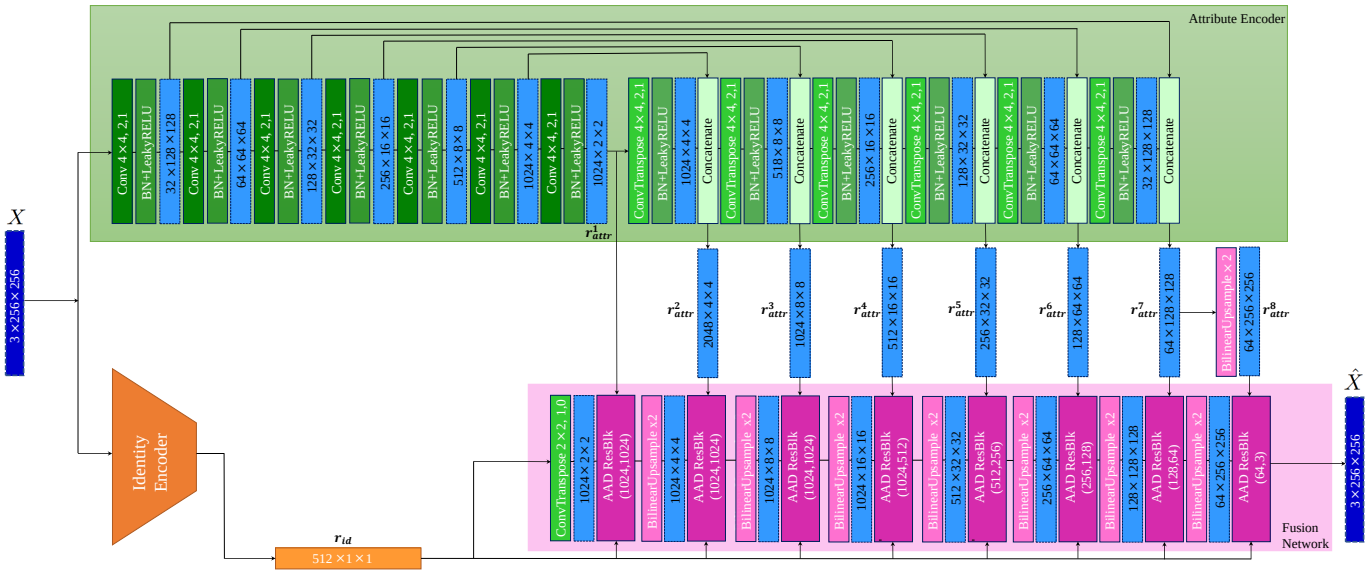

Fig. 3. Network structure of the proposed neural network in stage-I. *Conv k,s,p* represents a Convolutional Layer with kernel size k, stride s and padding p. *ConvTranspose k,s,p* represents a Transposed Convolutional Layer with kernel size k, stride s and padding p. All *LeakyReLUs* have $\alpha = 0.1$. *AAD ResBlk* (c_{in}, c_{out}) represents an AAD ResBlk with input and output channels of c_{in} and c_{out}.

achieve a strict DP privacy guarantee. While other mechanisms such as Gaussian mechanism and Exponential mechanism may also be used, they are not as popular as Laplace mechaism. Hence, the existing methods [58, 59] that implement DP for images all adopt the Laplace mechanism, and we select the Laplace DP mechanism in our method at the second stage too.

V. Experiments

A. Experimental Setup

1) *Datasets*: We choose the CelebA-HQ datasets, which contains 30K high-resolution celebrity images with diverse demographic information like age, gender, and race [60], to train our network in stage-I. We randomly select 27K images for training and 3K for testing. Moreover, in order to demonstrate our generalization ability and compare with conditional comparisons conveniently, we also test IdentityDP on the CelebA [61] datasets. All images are aligned and cropped to size 256×256 covering the whole face, as well as some background regions.

2) *Comparison methods*: To validate the effectiveness of the proposed IdentityDP framework, we compare to traditional anonymization methods as well as state-of-the-art methods.

- Traditional Anonymization methods. We use Pixelization, Noise and Blur of faces.
- State-of-the-art methods. We select 4 methods: AnonymousNet [15], DeepPrivacy [19], CIAGAN [25] and Fawkes [14].

B. Evaluation Metrics

We evaluate all methods in privacy metrics as well as utility metrics.

1) Privacy metrics. Two different metrics are used to measure the performance of privacy protection.

- *Identity Distance* $\mathcal{ID\_DIS}$. We employ FaceNet identification model [46] based on Inception-Resnet backbone, pre-trained on two public datasets: CASIA-Webface [62] and VGGFace2 [63], whose LFW accuracy can reach 99.05% and 99.65% individually. The output distance of FaceNet can indicate the pairs of input faces' identity difference.
- *Protection success rate* $\mathcal{PSR}$. Besides publicly available datasets and known model architectures for academic usage, we also wish to understand the performance of IdentityDP on public facial verification services that people may touch in daily life. Therefore Microsoft Azure Face [64] is employed to evaluate real-world effectiveness of a method. It gives judgement of whether the input pairs are of the same people. The protection success rate is the proportion of faces that are judged as different from the original ones.

2) Utility metrics. Two different metrics are used to evaluate the utility of processed images.

- *PSNR* and *SSIM*. We choose peak-signal-to-noise ratio (PSNR) as well as structural similarity index measure (SSIM) as two objective measures of similarity between anonymous results and original faces.
- *Face detection rate* $\mathcal{FDR}$. We evaluate whether the processed images are still usable for identity-agnostic computer vision tasks by performing face detection using HOG [65] Detector, and we calculate the proportion of faces that can be detected in the protected images.

C. Implementation Details

We implement our framework as shown in Fig. 2. The number of attribute representation is set to $n = 8$ (Eq. (8)). The detailed network structure is given in Fig. 3. In the training process, we use the Adam optimizer [66] with

momentum parameters $\beta_1 = 0, \beta_2 = 0.999$. The learning rate is set to 0.0004. The parameters in Eq. (18) are set to $\lambda_{att} = \lambda_{rec} = 10, \lambda_{id} = 5$.

D. ϵ-IdentityDP Mechanism Analysis

To explicitly understand the differential privacy mechanism in our proposed IdentityDP, we design an experiment to explore how the privacy budget ϵ affects the face anonymization performance. First of all, we extract every test image's identity representation and calculate the l_1-norm sensitivity Δf, i.e., $\Delta f = \max_{X,X'} ||r_{id}(X) - r_{id}(X')||_1$, $X, X' \in$ *test datasets*. Then we increase ϵ from 1.1 to 800, and accordingly adjust the IdentityDP framework. Since our ϵ-IdentityDP mechanism $\mathcal{M}(X)$ is $\mathcal{M}(X) = r_{id}(X) + Lap(\frac{\Delta f}{\epsilon})$, we double ϵ for better display effect and 100 anonymous faces are generated for every test face under each ϵ. Finally, various statistical mean metric values are calculated at each ϵ value.

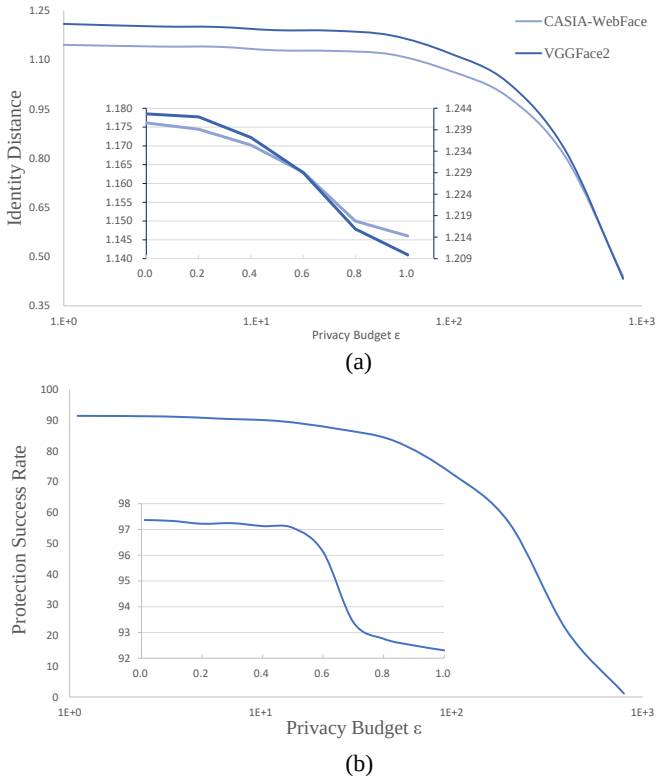

Fig. 4. Identity protection performance: (a) the identity distance calculated by FaceNet model trained on CASIAWebface and VGGFace2 datasets respectively; (b) the Protection success rate calculated through public facial verification service [64].

For privacy protection, when ϵ increase from 0.01 to 800, Fig. 4 (a) shows that the average identity distance decreases gradually and Fig. 4 (b) shows that the protection success rate decrease from 97.367% to 1.125%, illustrating that a smaller privacy budget guarantees better de-identified results. We show anonymous image whose identity distance is closest to the mean distance under every ϵ in Fig. 6 for visual observation, which also implies the diversity of our de-identified results. For data utility, Fig. 5 (a) plots PSNR and SSIM vs. ϵ, indicating that the visual similarity gets better as the privacy budget increases. Fig. 5 (b) shows that our face detection rate always remains at a high level, demonstrating that identity-agnostic computer vision technologies can still work on our processed faces. Specially, when the privacy budget ϵ is small (i.e., a strong privacy protection), the subtle differences between the de-identified faces and the corresponding original ones can be perceived by humans easily (e.g., e.g., dfferent eyebrow shapes, different iris colors, and different lip shapes), while they still share a great visual similarity on the whole.

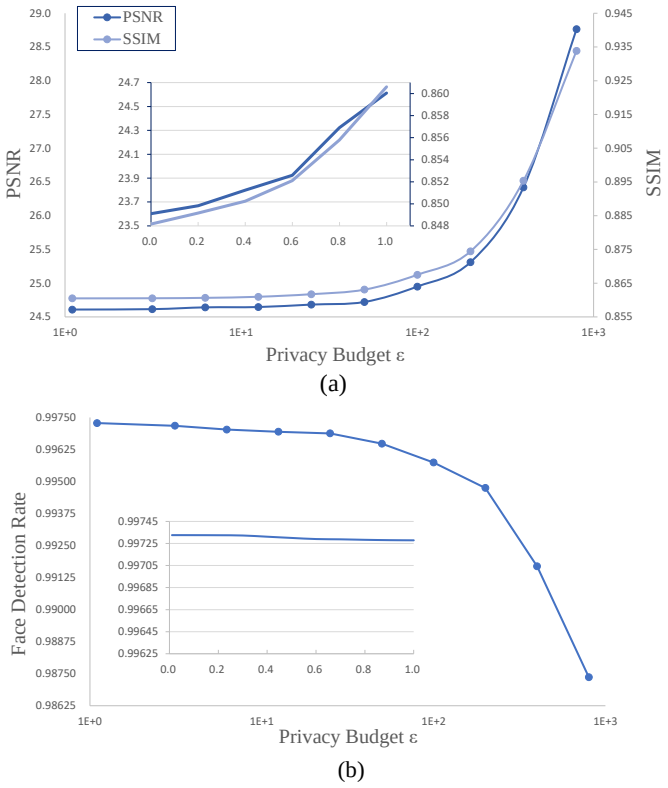

Fig. 5. Image utility performance: (a) PSNR and SSIM; (b) the Face detection rate calculated through HOG detector.

Furthermore, an unexpected issue is that the face detection rate decreases slightly as ϵ increases. After research, we find the reason is that partially severely blocked faces in test datasets can recover some facial features in the blocked area using our framework, resulting in the detection of originally undetectable faces.

Fig. 1 illustrates some de-identified results in pairs, where left is the original image and right is the result generated by our framework. It demonstrates that human identities are obfuscated in a perceptually natural manner, in the meantime, each pair of images still shares similar appearance, as well as the same expression and background. It is worth noticing that our results can well retain the unique attributes of characters, such as rare hairstyles, beards, glasses and other accessories, which is hard to achieve in previous GAN-based methods.

Based on a large number of experiments, we get some experience in choosing a suitable privacy budget value: if the image's hue is light or the people's expression is exaggerated, a smaller privacy budget should be chosen. In fact,

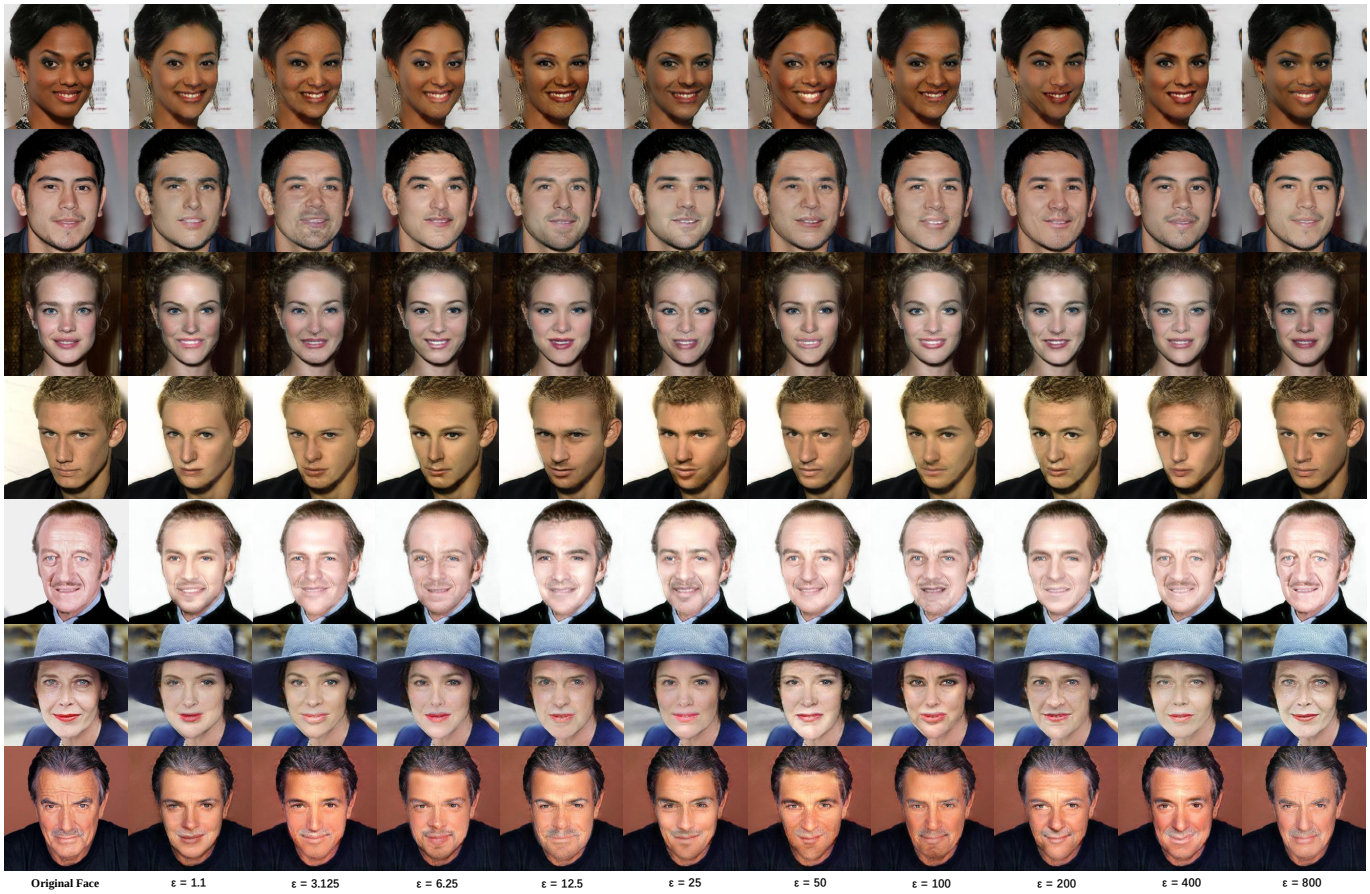

Fig. 6. Qualitative comparison of the influence of parameter ϵ. The first column shows the original face images. The rest columns demonstrate anonymous face whose identity distance is closest to the mean distance under every ϵ.

we believe that setting the privacy budget to any positive number less than 10 will get the advanced anonymization effect successfully, and We recommend the user to set their privacy budget between 0.5 and 7 to obtain anonymous face efficiently with quite well-preserved appearance. Specifically, during the subsequent experiments, we set privacy budget to 6 and 0.57.

E. Comparisons with Traditional Methods

In this subsection, the following traditional methods are implemented: (1) Pixelization: we cluster face region's pixels that are close in 2D space and color space, and then replace each cluster (8×8, 16×16) with its average value. (2) Noise: we add Gaussian noise (σ = 9, 49) on each pixel's RGB value of the face region; (3) Blur: following *Ryoo et al.* [67], we downsample the face region to extreme low-resolution (7×7, 19×19) and then upsample back. We set the privacy budget to 6. It can be seen that for the fairness of comparison, we select two parameters for each traditional method: one aims to make the identity distance close to our approach, at this time, the utility metrics are mainly compared; the other aims to make PSNR or SSIM close to our method, at this time, the privacy metrics are mainly compared.

Fig. 7 shows the qualitative results. It is obvious that our approach achieves a great advantage in visual similarity as well as realism. The detailed quantitative results are shown in Table I, illustrating that the traditional methods fail to improve the privacy-utility trade-off and perform poorly in preventing practical face verification.

TABLE I
QUANTITATIVE EVALUATION ON CELEBA-HQ DATASETS UNDER DIFFERENT METRICS

Method	$\mathcal{ID\_DIS}$		$\mathcal{PSR}$	PSNR	SSIM	$\mathcal{FDR}$
	CASIA	VGGFace2				
Pixelization(8×8)	0.8646	0.8993	0	26.735	0.7671	0.923
Pixelization(16×16)	1.1541	1.2195	0.017	23.926	0.7223	0.058
Noise(σ = 9)	0.3317	0.2723	0.002	23.831	0.8312	0.986
Noise(σ = 49)	1.1267	1.0280	0.012	14.370	0.5533	0.425
Blur(7×7)	0.8491	0.8380	0	27.405	0.806	0.888
Blur(19×19)	1.1102	1.1857	0.669	24.829	0.7719	0.518
DeepPrivacy	1.0860	1.1829	0.961	21.012	0.7808	0.989
Fawkes	0.7267	0.8585	0	**35.898**	**0.9487**	0.985
Ours(ϵ=6)	1.1403	1.2012	0.908	24.640	0.8606	0.997
Ours(ϵ=0.57)	**1.1644**	**1.2307**	**0.967**	23.909	0.8519	**0.997**

F. Comparisons with State-of-the-art Methods

In this subsection, we compare our IdentityDP with state-of-the-art face de-identification methods. Among them, DeepPrivacy and Fawkes are trained and tested on CelebA-HQ datasets. Anonymousnet and CIAGAN require pre-annotations

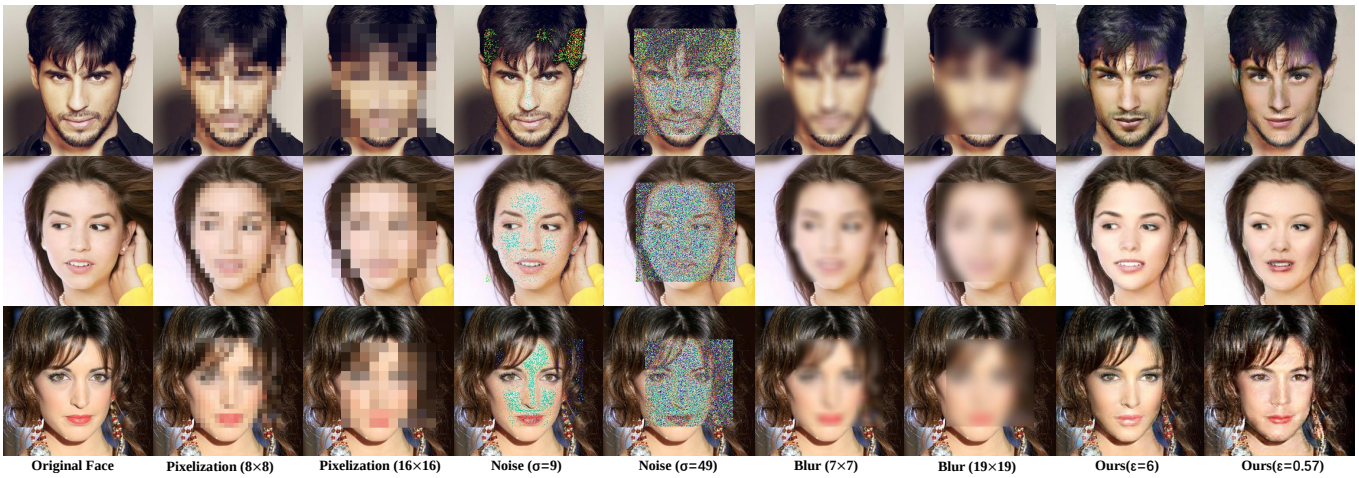

Fig. 7. Qualitative comparison with traditional methods. From left to right: original faces, faces obfuscated by Pixelization(4×4, 8×8), Noise(σ = 9, 18), Blur(8×8, 16×16), faces generated by our method(ϵ=6 and ϵ=0.57).

and are trained on CelebA datasets, so we transfer our framework on CelebA and compare with them for fairness. We evaluate performance with these methods respectively.

1) *Comparisons with Attribute manipulation-based Anonymization*: Facial attributes, including gender, age, haircut and so on, should be an important reference for identifying faces' identities, especially affecting human's subjective judgment. Therefore, manipulating face attributes to make faces anonymous seems reasonable. AnonymousNet, a privacy-preserving attribute selection algorithm for facial image obfuscation, is a typical representative.

Fig. 8 shows the anonymous faces generated from our framework and those from AnonymousNet. Due to the change of several face attributes, the anonymous face generated by AnonymousNet is often visually different from the original face, especially when modifying gender, which is not conducive to the normal use of the images. In contrast, our method achieves significant improvement in visual similarity. As can be seen from Table II, our method performs better under both privacy metrics and utility metrics, not to mention that AnonymousNet requires detailed data annotations. Moreover, it is worth noticing that although anonymous faces generated by AnonymousNet are visually very different from the original one, face verification service API can still judge them correctly, which suggests that general face attributes are not directly related to human identity.

2) *Comparisons with Conditional inpainting-based Anonymization*: Exposure of faces is the source of private information leakage. Therefore, some methods directly feed their networks with face-removing images as well as auxiliary annotations to automatically generate anonymous human faces. In this way, the generator never touches original faces, ensuring the removal of any privacy-sensitive information. DeepPrivacy is such a method which requires two annotations: a bounding box to identify the privacy-sensitive area and a sparse seven keypoints pose estimation of the face. It generates de-identified faces considering the original pose and image background. We compare our method with it.

Fig. 9 reports the difference of methods. We can see that the face generated by DeepPrivacy can maintain the facial pose well, but is not visually similar to the original image. Besides, distortions and artifacts often occur. Our method produces more visual-pleasing anonymous faces which look similar to the original one. Table I shows quantitative results, when ϵ is set to 6, our method is slightly inferior to DeepPrivacy in terms of privacy protection, but has remarkable data utility improvement, moreover, when ϵ is set to 0.57, our method outperforms DeepPrivacy with almost no distortion or artifact.

TABLE II
QUANTITATIVE EVALUATION ON CELEBA DATASETS UNDER DIFFERENT METRICS

Method	$\mathcal{ID\_DIS}$		$\mathcal{PSR}$	PSNR	SSIM	$\mathcal{FDR}$
	CASIA	VGGFace2				
AnonymousNet	0.8896	1.0589	0.295	18.892	0.7192	0.892
CIAGAN	0.8155	1.0271	0.945	21.863	0.7401	0.958
Ours(ϵ=6)	0.9345	1.0918	0.905	**23.353**	**0.8188**	0.986
Ours(ϵ=0.57)	**0.9622**	**1.1176**	**0.961**	22.7639	0.8005	**0.987**

3) *Comparisons with Conditional ID-swapping-based Anonymization*: Since anonymizing a face is intended to hide its original identity, swapping the original ID with others may be a straightforward idea. Conditioned on face landmark and masked background image of the input image, CIAGAN generates a new fake identity out of the input image to achieve anonymization.

We compare images generated from our proposed framework and those from [25]. From Fig. 8 we can see that the two methods produce comparable results, while ours enjoy a better visual similarity and less artifacts. Table II shows quantitative results. When ϵ is set to 6, CIAGAN protects privacy better from the perspective of PSR, however, when setting ϵ to 0.57, we outperform CIAGAN from all metrics and maintain a better visual similarity on the whole.

Moreover, CIAGAN has some notable flaws: 1) It needs to borrow someone else's identity as operation guidance,

Fig. 8. Qualitative comparison of our method with AnonymousNet [15] and CIAGAN[25]. The top row shows original faces, the second row and the third row show corresponding anonymous faces generated by AnonymousNet and CIAGAN. The last two rows shows our results(ϵ=6 and ϵ=0.57).

Fig. 9. Qualitative comparison of our method with DeepPrivacy [19] and Fawkes [14]. The top row shows original faces, the second row and the third row show corresponding anonymous faces generated by DeepPrivacy and Fawkes. The last two rows show our results(ϵ=6 and ϵ=0.57).

which may affect the privacy and security of the identity provider; 2) Faces de-identified by CIAGAN is visually similar to original ones only when the fake ID provider shares the same gender, a similar age as well as similar makeup with the person with the original ID, which makes it not very convenient to use in practice; 3) CIAGAN cannot maintain certain special attributes, such as glasses, heavy makeup, and thick beard, unless the identity provider also has; 4) CIAGAN depends on landmark detection to provide pre-annotations, which tends to miss any face that has not been detected in the anonymization process. In contrast, our approach does not have these problems, as ours does not need the assistance of other identities, can retain the special attributes of original faces, and does not need pre-annotations.

In summary, our method can surpass the privacy preservation ability of CIAGAN while maintaining a similar appearance to their original ones, and our proposed method significantly performs better in terms of the utility metrics, which is preferable for real-life applications. Hence, our proposed method is superior to CIAGAN in terms of the privacy-utility tradeoff and the capacity for providing provable and strict privacy guarantee.

4) *Comparisons with Adversarial Perturbation-Based Anonymization*: De-identified methods based on adversarial examples are continuously popular because of their almost the same anonymous results. However, their performance depends largely on the accessibility of the target system's internal parameters, or special training on the target system. Fawkes, as one of the latest representatives, is selected as our comparison.

Fig. 9 demonstrates that Fawkes can generate faces that look extremely like the original one, except for a few strange spots that sometimes appear. We just provide a comparable result. However, Table I shows that Fawkes performs poorly under privacy metrics, which means that faces processed by Fawkes are unable to obfuscate the previously inaccessible systems. In contrast, although our method suffers less visual similarity, it works significantly better in preserving face privacy.

TABLE III
ADDITIONAL QUANTITATIVE EVALUATION WITH STATE-OF-THE-ART METHODS ON LFW DATASETS

Method	FaceNet model		FID
	CASIA	VGGFace2	
Original	0.965 ± 0.016	0.986 ± 0.010	0
AnonymousNet	0.037 ± 0.015	0.044 ± 0.016	6.8479
DeepPrivacy	0.029 ± 0.012	0.039 ± 0.014	2.7122
CIAGAN	0.019 ± 0.008	0.034 ± 0.015	2.1756
Fawkes	0.898 ± 0.010	0.917 ± 0.012	**1.2681**
Ours(ϵ=6)	0.019 ± 0.010	0.031 ± 0.015	2.0201
Ours(ϵ=0.57)	**0.016** ± **0.011**	**0.024** ± **0.014**	2.0401

5) *Additional Discussion*: To make the comparison more convincing and fairer, we follow the evaluation protocol that has been used in CIAGAN, and add two experiments with the state-of-the-art methods to evaluate the performance of privacy and utility respectively.

Firstly, we use the evaluation method for privacy protection, which is conducted on the LFW benchmark. In this experiment, we employ two FaceNet identification models (pre-trained on CASIA-Webface [62] and VGGFace2 [63]), and the main evaluation metric is the true acceptance rate. Tab. III presents the results on de-identified LFW image pairs for a

Fig. 10. Our de-identification results on examples labeled as challenging or very challenging in the NIST Face Recognition Challenge [68]. The first row shows original faces, and the following row shows our corresponding de-identified results.

given person, while the de-identification method is applied to the second image of each pair. It can be seen that all the state-of-the-art methods can let the true positive rate drop from almost 0.99 to less than 0.05 except Fawkes. In particular, when ϵ is 0.57, our method yields the lowest true positive rate when two FaceNet models pre-trained on CASIA dataset and VGGFace2 dataset are employed.

Then we evaluate the utility of the images by using the FID score on LFW dataset, as it can measure the distance between the generated distribution and the real distribution. The results are shown in Table III. Among the methods that can effectively drop the true acceptance rate and well protect the identity information, our method achieves the best FID score. It demonstrates that our de-identified images exhibit a closer similarity to the original ones in terms of data distribution, which is consistent with our intuitive expectation.

Besides, there is no formal and strict privacy guarantee provided by the state-of-the-art privacy protection methods, while the privacy level of our proposed IdentityDP is clearly defined and rigorously guaranteed by the DP criterion. Therefore, our method has the advantage of providing provable and strict privacy guarantee.

G. Generalization Ability

Our IdentityDP provides great generalization to various face images. In previous experiments, it has been proved by showing remarkable qualitative and quantitative results on CelebA, a datasets that our IdentityDP has never trained on before. To further demonstrate the robustness of our method, we apply our framework to face images from the very difficult inputs of [68]. As can be seen in Fig. 10, our method is robust to very challenging illuminations.

In addition, we apply our framework on artistic portraits. All artworks are taken from Wikiart.org. Fig. 11 shows the interesting results, illustrating that faces in different styles are anonymized successfully without causing significant distortions or artifacts.

H. Computational Overhead

In this subsection, we evaluate our computational overheads for anonymizing faces. IdentityDP adds little overhead for processing, as the only additions are a random noise tensor. On an NVIDIA GTX 1080 Ti, IdentityDP takes on average 0.329s per image. The low computational overhead is beneficial to process a large amount of face images.

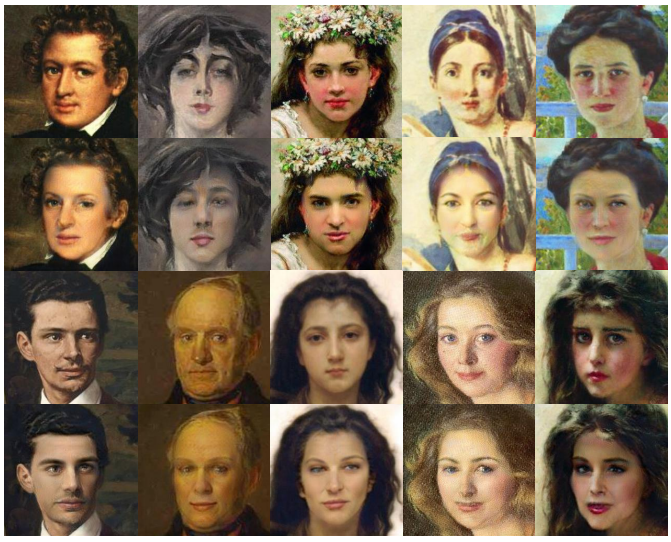

Fig. 11. Our anonymization results on challenging artistic portraits. The first and the third row show the artistic portraits, while the second and the fourth row show our corresponding anonymous results.

VI. Conclusions and Future Work

In this paper, we propose the IdentityDP framework that combines differential privacy mechanisms with deep neural networks to achieve image privacy protection for the first time. Our framework consists of three stages: deep representations disentanglement, ϵ-IdentityDP perturbation and image reconstruction. In our framework, DP perturbation is directly added on to the identity representation to ensure privacy protection, while the attribute representation is unchanged and it preserves visual similarity well. Furthermore, the adjustable privacy budget guarantees the diversity of anonymization results. Experiments demonstrate the effectiveness of our framework in terms of privacy protection and image utility, and produce satisfactory results compared with the traditional as well as state-of-the-art methods. Moreover, our framework has a good generalization ability. In the future, we will further explore the trade-off between user privacy and authorized use of work. In addition, extending this work to videos and achieving temporal consistency would be an interesting direction.

References

[1] V. Mirjalili, S. Raschka, and A. Ross, "Privacynet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[2] I. Barron, H. J. Yeh, K. Dinesh, and G. Sharma, "Dual modulated qr codes for proximal privacy and security," *IEEE Transactions on Image Processing*, vol. 30, pp. 657–669, 2021.

[3] E. Commission, "2018 reform of eu data protection rules," 2018.

[4] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005.

[5] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2005, pp. 227–242.

[6] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," in *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*. IEEE, 2006, pp. 161–161.

[7] L. Du, M. Yi, E. Blasch, and H. Ling, "Garp-face: Balancing privacy protection and utility preservation in face de-identification," in *IEEE International Joint Conference on Biometrics*. IEEE, 2014, pp. 1–8.

[8] A. Jourabloo, X. Yin, and X. Liu, "Attribute preserved face de-identification," in *2015 International conference on biometrics (ICB)*. IEEE, 2015, pp. 278–285.

[9] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," *arXiv preprint arXiv:1908.08705*, 2019.

[10] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial image perturbation for privacy protection a game theory perspective," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1491–1500.

[11] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 6103–6113.

[12] B. Liu, J. Xiong, Y. Wu, M. Ding, and C. M. Wu, "Protecting multimedia privacy from both humans and ai," in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. IEEE, 2019, pp. 1–6.

[13] C. Zhu, W. R. Huang, A. Shafahi, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," *arXiv preprint arXiv:1905.05897*, 2019.

[14] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 1589–1604.

[15] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[16] H.-P. Wang, T. Orekondy, and M. Fritz, "Infoscrub: Towards attribute privacy by targeted obfuscation," *arXiv preprint arXiv:2005.10329*, 2020.

[17] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5050–5059.

[18] Z. Ren, Y. Jae Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 620–636.

[19] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International Symposium on Visual Computing*. Springer, 2019, pp. 565–578.

[20] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-gan for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, 2019.

[21] B. Meden, R. C. Mallı, S. Fabijan, H. K. Ekenel, V. Štruc, and P. Peer, "Face deidentification with generative deep neural networks," *IET Signal Processing*, vol. 11, no. 9, pp. 1046–1054, 2017.

[22] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 553–569.

[23] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[24] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9378–9387.

[25] M. Maximov, I. Elezi, and L. Leal-Taixé, "Ciagan: Conditional identity anonymization generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5447–5456.

[26] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 594–603, 2020.

[27] R. Hasan, E. Hassan, Y. Li, K. Caine, D. J. Crandall, R. Hoyle, and A. Kapadia, "Viewer experience of obscuring scene elements in photos to enhance privacy," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[28] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[29] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," in *European Conference on Computer Vision*. Springer, 2016, pp. 19–35.

[30] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[31] N. Vishwamitra, B. Knijnenburg, H. Hu, Y. P. Kelly Caine *et al.*, "Blur vs. block: Investigating the effectiveness of privacy-enhancing obfuscation for images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 39–47.

[32] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, B. Wu, A. Bissacco, H. Adam, H. Neven, and L. Vincent, "Large-scale privacy protection in google street view," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 2373–2380.

[33] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, and D. Warde-Farley, "Generative adversarial nets in advances in neural information processing systems (nips)," 2014.

[35] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[36] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[37] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[38] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[39] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," *Advances in neural information processing systems*, vol. 27, pp. 2879–2887, 2014.

[40] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 729–745.

[41] T. Zhu, G. Li, W. Zhou, and S. Y. Philip, *Differential privacy and applications*. Springer, 2017.

[42] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.

[43] Y. Zhong, W. Deng, J. Hu, D. Zhao, X. Li, and D. Wen, "Sface: Sigmoid-constrained hypersphere loss for robust face recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2587–2598, 2021.

[44] Y. Sun, X. Wang, and X. Tang, "Sparsifying neural network connections for face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4856–4864.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[46] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[47] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[48] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[49] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[50] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[51] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[52] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[53] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[54] H. Xue, B. Liu, M. Din, L. Song, and T. Zhu, "Hiding private information in images from ai," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland. IEEE, Jul 2020.

[55] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov,

K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[56] M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, "A survey on differentially private machine learning," *IEEE Computational Intelligence Magazine*, vol. 15, no. 2, pp. 49–64, 2020.

[57] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299–310, 2011.

[58] B. Liu, M. Ding, H. Xue, T. Zhu, D. Ye, L. Song, and W. Zhou, "Dp-image: Differential privacy for image data in feature space," *arXiv preprint arXiv:2103.07073*, 2021.

[59] T. Li and C. Clifton, "Differentially private imaging via latent space manipulation," *arXiv preprint arXiv:2103.05472*, 2021.

[60] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[61] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[62] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[63] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[64] "Microsoft azure face api," https://azure.microsoft.com/en-us/services/cognitive-services/face/.

[65] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[67] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," *arXiv preprint arXiv:1604.03196*, 2016.

[68] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, *An introduction to the good, the bad, & the ugly face recognition challenge problem*. IEEE, 2011.