Survey on Biometric Anonymization Evaluation

Julian Todt julian.todt@student.kit.edu

***Abstract*—With the widespread collection of audio-visual data (video surveillance, social networks, media, etc.) and the advancements in identification of individuals through biometric traits (face, gait, voice, etc.) privacy issues arise. To combat this, anonymization methods have been proposed and used such as blurring faces in video recordings. It is however not clear how to compare different methods in respect to their features and effectiveness. In this paper, we analyze and compare evaluations of anonymization methods for biometric traits. We also point out problems as well as possible solutions for different aspects of current evaluations. We find that there is no standard method to evaluate biometric anonymization methods. While the evaluations that we analyze show similarities in their approaches to evaluate the privacy protection of anonymizations, there are still significant differences. Also, while many anonymization methods make claims about their utility, a significant portion does not evaluate these claims at all.**

***Index Terms*—privacy enhancing technologies, anonymization**

I. Introduction

In the name of public safety, governments and law enforcement around the world have increased video surveillance of the public in recent times. In the United Kingdom, more than 4 million CCTV cameras mean that the average London citizen is caught on cameras 300 times a day [1]. These audio-video recordings of people in public environments are both used for immediate inspection and for storage with subsequent analysis and sharing [2]. This means there is a strong requirement to protect the privacy of the individuals that are being recorded. This is reinforced by a major lack in compliance with data-protection legislation which is insufficient in preventing misuses anyway [3, p. 39] [1]. Social media is another area that contributes to the general increase of audio-video recordings of individuals as well as their quality in recent years.

To protect against the identification of individuals based on their biometric features, a multitude of anonymization methods have been proposed. Their goal is to modify the audio-visual recording in such a way that identifying individuals based on a specific biometric feature becomes significantly less reliable while preserving as much as possible from the original recording. Often however it is not clear, how successful these methods are at achieving this goal and how different methods diverge.

A multitude of different anonymization methods for different biometric traits, each with different privacy and utility goals means there is no standardized evaluation methodology for these biometric anonymization methods. Instead the authors of every method make different assumptions, for example about the attacker model when designing the experiments to evaluate their work. This is problematic as it makes comparisons of different anonymization methods difficult and questions the results of the different evaluations.

In this paper, we analyze and compare these evaluations to determine the current state-of-the-art. We also point out problems as well as possible solutions for different aspects of the current evaluations.

We start with background information in Section 2, especially about how the recognition technologies work which the anonymization methodologies try to disrupt. In Section 3 we analyze the evaluations of a variety of anonymization methodologies for different biometric traits. We compare the analyzed evaluations in Section 4 and 5 both within one biometric trait and across. This is also where we show current issues and propose possible solutions. In Section 4 we consider evaluations of the privacy protection feature of the anonymizations. Meanwhile, we consider the evaluations of utility preservation features in Section 5. We consider related work in Section 6. In Section 7 we make our conclusions and propose future work.

II. Background

In this section, we want to provide some background information.

Biometric traits (or biometric identifiers) are properties of humans that identify a single individual. We differentiate between soft biometric identifiers which are vague physical, behavioral or adhered human characteristics that are not necessarily permanent or distinctive. A single soft biometric trait cannot reliably used for personal identification, but can be used to categorize people or improve identification. Combinations of multiple soft biometric traits can however lead to successful personal identification [4]. Examples are height, weight, eye color, age, gender, race. Other biometric identifiers however are distinctive, measurable, generally unique and permanent. These can be reliably used for personal identification and are split into groups: physiological (face, iris, ear, fingerprint) and behavioral (voice, gait, gesture, lip-motion) [2] [4].

Any single one of the biometric traits can be used to identify the person they belong to using a **recognition method**. To use a recognition method, it first has to be trained with a data set of known biometric feature to identity associations. Afterwards the recognition can output the identity which it perceives to be the closest match when given a new instance of a biometric feature. Often the output also includes a numeric value in $[0, 1]$ which denotes how close the match is. The input data type depends on the biometric trait and the recognition method, but is usually still images for physiological biometric identifiers and audio or video recordings for behavioral biometric identifiers. Recognition methods are evaluated based on their ability to correctly recognize individuals (success rate). These success rates regularly exceed 90% for modern recognition methods in their own evaluations [2] [5]. The specific technology by which recognition methods recognize individuals depends heavily on the specific method, but more recent methods often use machine learning.

For example for **face recognition**, methods can be be categorized as feature-based, holistic and machine learning. Feature-based approaches, which were proposed as early as 1973, work by identifying, extracting and measuring distinctive facial features and then computing geometric relationships between these facial points [5] [6]. These facial parameters are then compared to find matches and identify faces. Holistic recognition approaches include methods using Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), Eigenfaces or similar. Recently, machine learning approaches with convolutional neural networks (CNN) are the dominant face recognition methods with widely available easy-to-use implementations available [5] [7] [8] [9].

The goal of an **anonymization method** (de-identification method) is to disrupt the recognition's ability to correctly identify individuals based on their biometric identifier. Some researchers make a distinction between anonymization and de-identification where de-identification refers to a process of removing or obscuring biometric features that is reversible while an anonymization is irreversible. Since we do not consider the planned reversal of any de-identification in this paper, we use the two terms interchangeable [2]. As both recognition and anonymization are active research topics and have opposing goals, the problem field has been modeled from a game theory perspective in [10] with two opposing players: the anonymizing user and the recognizer. This is because recognition methods evolve to become better at identifying individuals even under suboptimal circumstances, such as when information was purposely removed from a recording by an anonymization method. At the same time, anonymization methods evolve to become better at privacy protection and prevent the identification of individuals even as recognitions improve.

Anonymization methods achieve their goal through a variety of different ways. To illustrate this, Fig. 1 shows for example a variety of different facial anonymization methods. Some anonymization methods (such as (b) and (e) in this example) de-identify by blanking/blacking out parts (or all) of the biometric feature. This leaves less (or none) information for the recognition method to work with which results in worse identification performance. Other anonymizations (such as (c) and (d) here) remove information from the data by blurring it. This means that the data for different individuals is more similar which also results in worse identification performance. Finally, some anonymizations (such as (f) and (g) in Fig. 1) purposely modify the biometric feature. This is done in such a way that based on only the de-identified data the anonymization is not apparent at first glance, while at the same time the modified biometric features make a recognition significantly harder.

The reason for this approach is the secondary goal of an anonymization: **utility**. If it were not for this secondary goal, there would be no need for anonymization methods other than those that completely remove any information on the biometric features. This is because without any biometric information, any recognition cannot do better at identifying an individual than guessing which makes the anonymization method perfectly privacy protecting. However, the reason we use anonymization methods at all is because the anonymized data is supposed to be further used or shared. For both, completely black images and completely muted audio recordings are not useful. This is why most anonymization methods make a trade-off between privacy protection and utility preservation. The specific utility that the anonymization methods preserve from the original data in the anonymized data depends heavily on the anonymization method and the user requirements depend on the specific use case and the considered biometric trait [2] [11] [12].

III. Analysis

In this section, we analyze the evaluations of biometric anonymization methods. For the most part, these are from the evaluation sections of the papers in which the specific anonymization method is introduced, but there are also a few standalone evaluations out there.

In this paper, we want to focus on anonymizations for the biometric traits face, gait and voice. This is because there is the most research on these three and they have the highest need for privacy because the biometric features can be recorded potentially non-consensually and from a distance. On the opposite, the biometric identifiers fingerprint and iris for example are usually used with consent from the individual in authentication systems because they require a short distance for recording [2]. The privacy issue in this case is called biometric template protection and is not subject of this paper [13]. However, we compare our work to this related field in Section 6. For the biometric identifier ear, a number of issues persist in the field of recognition and there has not been extensive research on anonymizations yet [2]. Similarly, compared to other biometric traits, there is less research into the biometric identifiers gesture and lip-motion which is why will not consider them in this paper.

A. Face

The question of how successful facial anonymization methods are has been researched as early as 1999 by the authors of [14]. **Blurring** and **pixelation** (see Fig. 1 (c) and (d) respectively) were then and still are the most common facial anonymization methods in practice. They are for example used in television or more recently in Google Streetview [15]. In their study, the authors show 32 participants images and videos of famous people with their faces blurred or pixelated. The participants are then asked to identify the people. Since the main goal of face anonymization then was to protect people from being recognized by people who know them, the authors argue that a successful anonymizations means that the study participants are not able to identify the famous people. The percentage of faces that the participants are able to correctly identify (recognition rate) is measured for both non-anonymized and anonymized images and then compared. The results are that participants are still able to recognize some of the viewed faces despite the anonymization. While

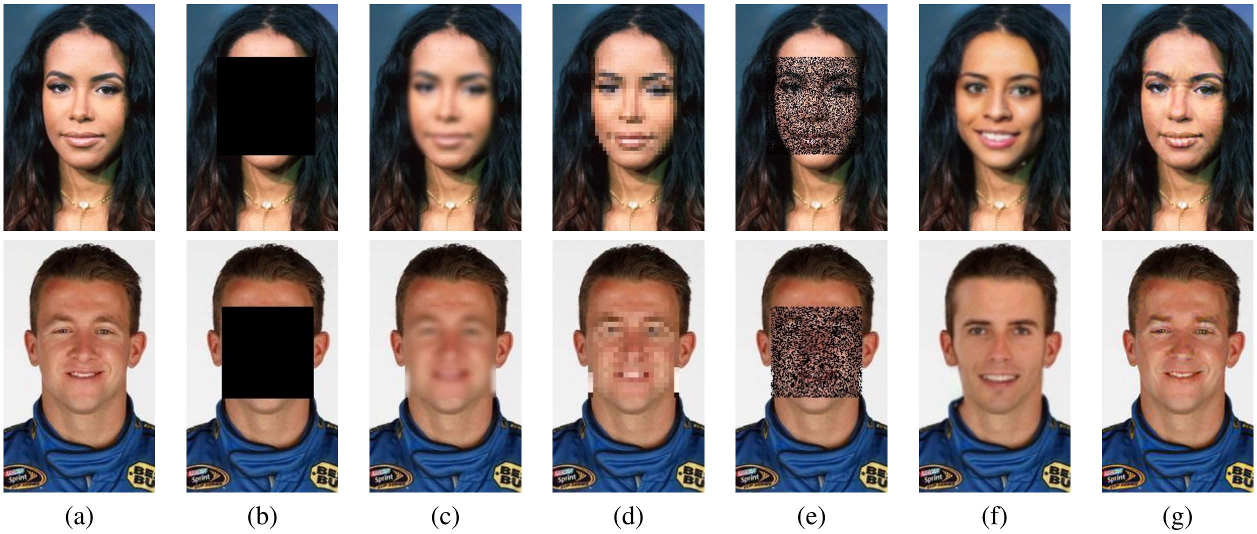

Fig. 1. Different facial anonymization methods. a) original image b) blackbox c) gaussian blur (kernel size 9x9) d) pixelated face (5x) e) random 50% of pixels black f) DeepPrivacy [11] g) Fawkes (middle mode, v1.0) [12]

the recognition rate is lower than without anonymization, it is not completely obscured. Recognition is also higher for videos than for still images.

Similar anonymization methods are evaluated in [16]. Additionally, a **masking** anonymization method is evaluated which removes random pixels from the face, similar to Fig. 1 (e). This time however, computers with deep learning are used instead of humans to evaluate the anonymizations' performance. After facial images from the CelebA dataset [17] are anonymized, specific (the anonymization method matching) deep learning reconstruction algorithms are used with the goal of reversing the anonymization. Finally, to evaluate the anonymization method, both the anonymized and reconstructed image are compared with the original image in three different ways. Firstly, a structural similarity index (SSIM) quantifies the image quality modifications. Secondly, OpenFace [18] computes the identity distance between the images. Thirdly, a pretrained recognition method is used to determine the difference in confidence in matching the images to the correct identity. In this evaluation a successful anonymization method would result in large differences for all three metrics between the original and both the anonymized and recognized image. This is because a larger difference between the original and the anonymized image means that privacy of the depicted is protected because they are less likely to be identified in the anonymized image. A larger difference between the original and the reconstructed image means that the restoration method is not very successful and the anonymization method is robust against reconstruction attacks. The authors find that the reconstruction algorithms are generally very successful in reducing the difference between the original and reconstructed image for the first two metrics, not however for the third metric. Finally, the authors find that the best performing anonymization methods are ”motion blur”, ”gaussian blur” and ”masking” for the first, second and third metric respectively.

In [19] the ***k*-Same**{,-Eigen,-Pixel} class of anonymization methods is introduced. They were the first methods that modified facial images instead of just masking/blurring/pixelating faces. They work by replacing the face with the average of a number of different but close faces. The authors evaluate their approach by testing the recognition performance of the recognition method Eigenfaces on anonymized images. They test the ability to match original images to altered images (naive recognition), altered to original images (reverse recognition) and altered to altered images (parrot recognition). The number of identities in the used dataset is varied in $[2, 100]$ and are randomly chosen from the FERET dataset [20]. For an optimal anonymization method, the recognition cannot do better than guess the correct match which means its correct best match rate is expected to be $1/k$ with k being the number of identities to choose from. The authors find that the k-Same-Eigen and k-Same-Pixel methods both perform better than $1/k$ (on average -1.6%) for naive recognition with similar results for reverse and parrot recognition. Common masking and pixelation anonymization methods are also tested on the same setup with ”bar masks” for example achieving a 2% recognition rate in naive recognition. While the authors point out the improved utility of k-Same compared to masking and pixelation approaches, it is not evaluated.

Extensive evaluation of utility aspects is however done by the authors of **AnonFACES** in [21]. Their anonymization method improves on previous work through improved clustering of the faces used for synthesizing the anonymized face as well as through incorporation of a StyleGAN, a state-of-the-art generative neural network which aims to improve the naturalness of created faces. The evaluation of utility is split into two aspects. In the first, naturalness is examined. This however is subjective to human observation, so the paper includes a variety of sample images and lets the reader judge for themself. In the second, the information loss at different stages of the anonymization process is examined and quantified. The authors find that at each of their three stages between 65% an 80% of information is lost depending on the chosen parameters. The privacy protection property of

AnonFACES is also evaluated by conducting re-identification experiments on the de-identified images. In the experiment, 50 identities are randomly chosen from the CelebA dataset [17] and both naive and reverse recognition are tested. The percentage of anonymized images correctly identified by the Dlib face recognition (recognition rate) is measured and compared to other anonymization methods including k-Same.

Another face anonymization method called **Face-Off** is introduced in [22]. Its use-case however is that users voluntarily upload anonymized images to social media sites instead of original pictures and are therefore able to avoid identification elsewhere. This is why the utility evaluation of Face-Off focuses on whether it is user-friendly and users would be willing to use the anonymization method on images that they upload to social media. In two user studies, the authors first ask whether users would upload a shown anonymized image of someone else, and second to submit a picture of themselves and when shown an anonymized version of it, whether they would upload it to social media. Participants are also asked multiple questions to determine how privacy conscious they are. The authors also evaluate Face-Off's privacy protection feature using commercial face recognition APIs. Azure Face API, Face++ and Amazon Rekognition all offer an API which computes a value in $[0, 1]$ that represents the confidence that two pictures show the same individual. In the evaluation, the original image as well as the by Face-Off anonymized image are passed to the APIs. The anonymization is considered successful if the APIs do not recognize the same person in the two images which is considered the case when the confidence value is below a threshold $\tau = 0.5$. They use randomly chosen subsets of the dataset VGGFace2 [23] and find that depending on API and amplification level the confidence drops between the threshold quickly.

A similar use-case scenario is also assumed in the face anonymization method **Fawkes** [12]. Its idea is to add "imperceptible pixel-level changes" to your own images before uploading them to social media. An example can be seen in Fig. 1 (g). The authors evaluate Fawkes privacy protection feature through a multitude of tests. In the first, two face image datasets VGGFace2 [23] and WebFace [24] as well as two feature extractors DenseNet-121 [25] and InceptionResNet V2 [26] are used. Then a single user's images are anonymized using Fawkes. It is tested whether the feature extractors match an original image of the user seeking protection to this user's anonymized images. They find that Fawkes is 100% effective in preventing this match. In the second more real world test, a face recognition method's ability to correctly identify individuals is tested when the recognition was trained with anonymized images. The cloud face recognition services Microsoft Azure Face, Amazon Rekognition and Face++ are provided with anonymized images of a co-author of Fawkes. It is then tested whether the recognition services match an original image of the same person to the anonymized images. This is measured by a protection success rate which is the percentage of original images for which the recognitions do not identify the individual. They find that some services are still able to correctly match the images when Fawkes' normal cloaks are used, but none are successful for Fawkes' robust cloaks. In a third test, the impact of uncloaked images in the training set as well as sybil accounts on the first experiment is tested. In the final test, countermeasures that try to disrupt or detect the cloaks are considered.

DeepPrivacy [11] is another face anonymization method. The authors claim that it is the first to provide complete privacy while generating realistic images with highly natural faces. An example can be seen in Fig. 1 (f). The claim of highly natural faces is evaluated by testing the ability of the Dual Shot Face Detector (DSFD) [27] to detect (not recognize) faces in the anonymized images. For this, the images of the WIDER dataset [28] are anonymized using DeepPrivacy as well as other anonymization methods like pixelation and blurring. For each of the anonymized datasets, as well as the original dataset, the average precision of face detection is measured. They find that DeepPrivacy is able to preserve DSFD's ability to detect faces in 99.3% of cases while blurring and pixelation only achieve 90.5% and 96.7% respectively.

B. Gait

In more recent times, gait recognition and anonymization has become an active research topic. In [29], the authors introduce a **gait anonymization** method. It works by decomposing the given silhouettes into their shape and phase components, perturbing them separately and then creating new silhouettes from the perturbed components. The authors claim that their methods significantly degrades recognition performance while the anonymized silhouettes still appear natural. They evaluate their method in two ways. Firstly, they evaluate re-identification performance by matching anonymized videos from the OU-ISIR Gait Database [30] to original videos using a multi-layer perceptron with three layers as a gait recognizer. They find that while original videos are matched in 100% of cases, the anonymized videos could only be matched in 30% or less cases. Secondly, the naturalness of the anonymized gait silhouettes is evaluated using two user studies. In the first, participants are asked to pick the anonymized video from a set of five. In the second, participants are asked to rate the naturalness of videos. The authors find that participants are not able to reliably pick the anonymized video in the first study and rate the naturalness of anonymized videos only slightly lower than that of original videos in the second study.

In [31], the authors introduce a gait anonymization method using **deep learning**. It works by combining the original gait and a noise gait using a convolutional neural network (CNN) into an anonymized gait. This method is also evaluated in two ways, both using the CASIA-B gait dataset [32]. Firstly, the naturalness of the anonymized gait is evaluated through a user study. Participants are shown an original and the corresponding anonymized video and are asked to rate the naturalness of the anonymized video using the mean opinion score (MOS). Secondly, the privacy protection feature of the anonymization is evaluated by testing the ability of the gait recognition system by Zheng et al. [33] to match anonymized videos to original videos. They find that the recognition fails to match the videos to the correct identities and the anonymization is therefore successful in between 48% and 86% of cases depending on view angle.

The authors of [34] analyze gait anonymization for a different use case. Smart phone authentication schemes based on gait recognition from the smart phones sensor data have been proposed, such as [35]. At the same time, this sensor data can also be used by attackers to predict key strokes and identify devices. To prevent these attacks on the user's privacy, anonymization of the sensor data by adding noise has been proposed. In [34], the authors evaluate whether the utility of the anonymized sensor data is still sufficient for the proposed authentication schemes. This is done by comparing the rate of successful authentications using original and anonymized sensor data. The authors use a self collected data set with 21 identities and a self implemented authentication system based on the work in [35].

C. Voice

A person not only has a visual identity but also an audio identity. Voice recognition methods use this to identify individuals based on voice recordings [2]. To protect against this, a variety of **voice conversion** methods that aim to protect the privacy of the speakers have been proposed including VoiceMask [36], VTLN-based voice conversion [37] and disentangled representation based voice conversion [38]. These three are evaluated in [39] in two aspects. The first aspect is the privacy protection feature of the voice conversion methods. To evaluate it, the authors measure the equal error rate (EER) for an i-vector or x-vector based voice recognition method when matching anonymized recordings to original ones. They compare the three voice conversion methods against a baseline from non-anonymized recordings. They use the LibriSpeech voice dataset [40] and vary the parameters of the voice conversion methods. Additionally, they consider different attacker models with varying amounts of knowledge of the anonymization: An *Ignorant* attacker which is unaware of the anonymization, a *Semi-Informed* attacker which knows which anonymization was used but not its parameters and finally an *Informed* attacker which knows both. Depending on the attacker model, the evaluation considers that the attacker converts their training and recognition data to the best of their knowledge before attempting to identify the speaker.

The second aspect of the evaluation in [39] is the utility preservation of the three voice conversion methods. The authors consider an anonymization approach to have high utility when the transcription of the anonymized voice recordings using automatic speech recognition is similar to that of the original recording. To measure this, voice recordings from the LibriSpeech voice dataset [40] are anonymized and then transcribed using the ESPnet [41] automatic speech recognition. Afterwards, the difference between the anonymized and original transcriptions is quantified using the word error rate (WER).

IV. Privacy Evaluation

In this section, we want to compare the different privacy protection evaluations that we analyzed in the previous section in order to find the current state-of-the-art. We also point out problems and offer potential solutions for these.

We find that almost all papers that introduce a new anonymization method also evaluate its privacy protection feature. This makes sense since privacy protection is the main objective of an anonymization method. Only the authors of DeepPrivacy [11] do not evaluate its privacy protection while claiming it ”guarantees the anonymization of faces”. They argue that their generator ”ensures 100% removal of privacy-sensitive information in the original face” [11, p. 1, p. 3].

The privacy protection feature of an anonymization method is evaluated by assuming an attacker with the goal to violate the privacy of one or more individuals. While this general approach applies to all evaluations, we find that the different evaluations vary significantly in different aspects. We will take a closer look at these different aspects in the following subsections.

A. Attacker model

The first major difference in the evaluations is the attacker model that the authors assume. With the exception of [39], the authors only implicitly define it which makes comparisons difficult. Tab. I shows an overview of some of the different attacker models in the analyzed evaluations.

A first aspect in which attacker models vary is whether the attacker is aware that an anonymization has taken place. Many evaluations simply assume what [39] calls an *Ignorant* attacker which treats the anonymized data as if it is unmodified. But there are even differences when attackers are **aware of the anonymization**, namely whether the attacker knows the specific anonymization method and potentially even the specific parameters with which it was run. On the one hand, in one of the experiments on Fawkes [12, p. 12f], the authors consider an attacker that cannot detect which images are anonymized, but that unsuccessfully uses a general transformation (augmentation, blurring, noise) on the images before its re-identification experiment. On the other hand, the attacker in [16] and the *(Semi-)Informed* attacker in [39] use their knowledge of the anonymization method to try to reverse the anonymization to the best of their knowledge.

The second aspect describes what **data** the attacker has access to. Depending on the attacker model, the attacker can use original or anonymized data (or a mix) as his training data as well as original or anonymized data for the recognition tests. This results in whether naive recognition, reverse recognition or parrot recognition is performed. The attacker's knowledge of this fact is defined by the first aspect. While some evaluations test multiple variants, most only perform one of them. The choice is mostly based on the use-case for which the anonymization method is designed. The more general purpose anonymization methods are evaluated using naive recognition (training with original data, recognition with anonymized). At the same time, Fawkes [12] is evaluated using reverse recognition because its idea is to ”poison” models that are trained using pictures collected on social media to avoid identification of not-anonymized images.

The final aspect is the **goal** which the attacker pursues. In the vast majority of evaluations, the attacker tries to identify individuals whose privacy is supposedly protected by the

TABLE I
ATTACKER MODELS IN PRIVACY EVALUATIONS

Evaluation	[14]	[16]	[19]	[21]	[22]	[12]	[29]	[31]	[39]
Knowledge of ...									
... anonymization	✓	✓	—	—	—	–/✓	—	—	–/✓/✓
... anonymization method	✓	✓	—	—	—	—	—	—	–/✓/✓
... anonymization parameters	—	—	—	—	—	—	—	—	–/ –/✓
Training data[a]	–[c]	☺	☺/●/●	☺/●	☺ [d]	●	☺	☺	☺
Recognition data[a]	●	●	●/☺/●	●/☺	● [d]	☺	●	●	●
Goal?[b]	Re-ID	Restore/Re-ID[d]/Re-ID	Re-ID	Re-ID	Re-ID[d]	Re-ID/Detect	Re-ID	Re-ID	Re-ID

[a] – none, ☺ original, ● anonymized

[b] Re-ID (re-identification), Restore (reverse the anonymization), Detect (detect that data was anonymized)

[c] humans identified known familiar faces.

[d] same person identified in corresponding original and anonymized data.

anonymization method. This is usually done by comparing the success rate of a recognition method on unmodified data with that when either training or recognition data is anonymized. Only the second attacker of [16] and the attacker of [22] instead use an identity distance metric to determine if original and anonymized data are recognized as the same individual. We also find other attacker goals in the evaluations like trying to reverse the anonymization in [16] or as straight-forward as trying to detect if data was anonymized in [12].

We find that in most cases, a weak attacker model is assumed where the attacker has no knowledge of the anonymization. This however can be problematic in cases where the anonymization can be detected because then the attacker can try to reverse the anonymization or adapt its recognition process to circumvent the anonymization. This seems to have already happened to Fawkes [12], when Microsoft updated its Azure Face API to ”lower the efficacy of the specific version of Fawkes that has been released in the wild” [42]. We recommend that evaluations consider whether the anonymization can be detected and/or whether knowledge of it can make for a more successful attacker. We also find that the choice of naive versus reverse recognition is largely based on the use-case that the anonymizations are designed for and together with different utility expectations for these anonymizations make for a difficult comparison in any case.

B. Parameters

In this subsection, we want to take a look at the values which authors vary within an evaluation. This is done to show how the evaluated anonymization behaves under different circumstances and how different assumptions may have an impact on the anonymization method's performance.

A first category of evaluation parameters are **parameters of the anonymization**. Most anonymization methods can be configured using on or more parameters to regulate their behavior. A gaussian blur anonymization for example can be configured through its kernel size while a pixelation anonymization can be configured through the number of remaining pixels. The evaluation in [14] does exactly this and tests two different pixelation levels and three different blur levels. Anonymization methods that are based on k-anonymity [43] have a parameter k that denotes the number of identities that are taken into account when creating the anonymized data. This parameter k is also often varied in the evaluations of such anonymization methods. But since these anonymization parameters depend on the specific anonymization method, comparisons are not possible.

An evaluation parameter which we find in multiple evaluations is the **evaluation group size (*egs*)**, the number of identities in the set that the anonymized identity is hiding in. This means that a recognition system trying to re-identify anonymized data has the choice between *egs*-many individuals. Please note that evaluations of anonymizations using k often use k as the value for *egs* and do not vary them independently. They will therefore use k to refer to the evaluation group size in the evaluation which can be misleading as the two do not necessarily have to be the same. Also, using k as the evaluation group size can infer that is a parameter of the anonymization method even if it is not which is why we use the distinct *egs* to refer to the evaluation group size. In both [21] and [19] *egs* is varied in $[2, 100]$. In other evaluations, *egs* is fixed for all experiments for different reasons: [29]: $egs = 20$ (”empirically”), [34]: $egs = 21$ (full dataset), [29]: $egs = 121$ (full dataset), [39]: $egs = 29$ (no rationale given). Some evaluations however do not use an *egs* parameter because of their experiment design. This is because they use a pretrained or cloud model or an Open-World-Assumption, such as [16], [22] and [12].

Some evaluations also vary the recognition method. This is often the case when ”blackbox” cloud recognition methods are used, such as in [22] and [12].

The evaluation of Fawkes [12, p. 10f] is the only one that varies the composition of training data. In one of their experiments, the impact of not-anonymized images in the mostly anonymized training data set is tested. The ratio of leaked uncloaked images is varied in [0,0.6].

We find that the only evaluation parameter that can be compared across multiple evaluations is the evaluation group size *egs*. Evaluations that vary *egs* show in their results that its choice is important because the results depend on *egs* a lot. We therefore think it is problematic when evaluations do not vary *egs* and their fixed choice is not thoroughly explained. This is especially the case when the evaluation experiment is based on the evaluations of recognition methods. There, a good recognition method performs well for large *egs*, because that means it can still identify individuals correctly in large groups.

For anonymization methods however we are often interested in worst-case performance to evaluate whether the privacy of individuals is still protected under sub-optimal circumstances. Therefore a good anonymization method performs well for small *egs*, because that means the recognition still fails to identify individuals even in small groups. At the same time, more realistic use-cases for example CCTV and social media, obviously have to deal with huge groups of people making large *egs* more realistic. We therefore recommend that authors either vary *egs* in their evaluations or use a *egs*-independent experiment design.

C. Closed-World-Assumption

Another aspect in which the analyzed evaluations vary is whether they use an open or closed world assumption. This is closely related to the discussion on the choice of *egs* in the previous subsection. When a closed world assumption is used in a re-identification experiment, the recognition method's closest match will be used regardless of its confidence. This means that an anonymization method could for example decrease the recognition confidence from 0.9 to 0.6 and it will be irrelevant in the evaluation as long as there is no closer match. This is especially relevant when there are not many identities to match (small egs). In an extreme example, for $egs = 2$ the anonymization method has to result in the recognition identifying data exactly the wrong way around in order to be successful in $> 50\%$ of cases. One might argue that with a closed world assumption, a successful anonymization has to modify data to appear as someone else and not just not the correct person.

This problem can be solved with an open world assumption and a ***Classify-Verify*** method similar to what is proposed in [44] for stylometry. This means that the closest match of a recognition method might be discarded if the confidence is below a predefined threshold. In our previous example, a threshold of $\tau = 0.7$ would mean that a decrease in confidence from 0.9 to 0.6 through the anonymization method is considered a success even if there is no match with higher confidence.

In the evaluations that we analyzed in Section 3, only one [22] explicitly used an open world assumption. There, the anonymization is considered successful when the confidence of a recognition that original and corresponding anonymized image show the same person is below a threshold of $\tau = 0.5$. Implicitly, an open world assumption is used by [12] since the majority of cloud face recognition services including the ones used only return a match when the confidence is above a (sometimes unknown) threshold. Similarly, the modern face recognition framework DeepFace [9] only returns matches with a difference below 0.4 in its `find`-method.

We find that a closed world assumption can be problematic especially for small identity sets. We therefore recommend using a *Classify-Verify* method with variable threshold.

D. Datasets

Tab. II shows the different datasets that are used by the analyzed evaluations. We find that for each of the three surveyed biometric traits, all recent datasets have similar sizes. The face datasets CelebA [17], VGGFace2 [23] and WebFace [24] all have around 10,000 identities while the gait datasets OU-ISIR [30] and CASIA-B [32] both have around 100 identities. The smaller face datasets from [14] and FERET [20] are both significantly older. It is therefore not surprising that most evaluations do not give a rationale for their choice of dataset since they could have used an alternative just as well.

TABLE II
DATASETS IN PRIVACY EVALUATIONS

Dataset	Identities	Data points
[14]	40	40
CelebA [17]	10,177	202,599
FERET [20]	1,199	14,126
VGGFace2 [23]	9,131	3,310,000
WebFace [24]	10,575	494,414
WIDER [28]	—	393,703
OU-ISIR [30]	102	2,482
CASIA-B [32]	124	1,364
LibriSpeech [40]	1,172	460h

We find that while the analyzed evaluations use a variety of different datasets, they tend to be very similar and we cannot note any major differences.

E. Metrics

All of the analyzed privacy protection evaluations try to quantify the performance of the anonymization method through a metric. While the names and definitions vary slightly across evaluations, the majority of metrics on re-identification experiments (see IV-A) are based on the same base metrics. In a preparatory step, the performance of the recognition method on not-anonymized data is tested. This is measured with the **recognition rate** (or success rate), the percentage of data points which are matched to the correct identity. We refer to it as r_{orig}. Afterwards, the recognition rate is measured again but now the recognition method uses the anonymized data. We refer to this value as r_{anon}. The different metrics used in the analyzed evaluations based on these two are shown in Tab. III. We would like to point out that the terms success rate, error rate, fail rate, etc. can be used referring to both the recognition and the anonymization and that they mean exactly the opposite for these two. For example the success rate of the anonymization is the fail rate of the recognition.

Two evaluations do not use a variant of these metrics because in their experiments, they compare the original and anonymized data for one person at a time. In [16], the metric is the relative decrease of the recognition method's confidence that the image shows the correct person from original image to corresponding anonymized image. In [22], the metric is the recognition method's confidence that the original and the corresponding anonymized image show the same person.

We find that evaluations that use a similar experiment design already use similar metrics to quantify the anonymization method's performance. However, while the $(1-)r_{anon}$ values are often displayed in tables and figures, the values for r_{orig} can be hard to find in the text of the evaluations. This can

TABLE III
METRICS IN RE-IDENTIFICATION EXPERIMENTS

Ref.	Name of metric	Definition	Notes
[14]	hit rate	r_{anon}	r_{orig} assumed as 1
[19]	recognition rate	r_{anon}	r_{orig} measured as 1
[21]	recognition rate	(r_{orig}, r_{anon})	
[12]	protection success rate	$1 - r_{anon}$	r_{orig} measured as 1
[29]	recognition accuracy	r_{anon}	r_{orig} measured as 1
[31]	success rate	$1 - r_{anon}$	r_{orig} not measured
[39]	equal error rate[a]	$(1 - r_{orig}, 1 - r_{anon})$	

[a] error rate when false reject rate equals false acceptance rate

make an evaluation harder to understand. For example, a value $r_{anon} = 0.3$ on its own might make an anonymization appear decent, but in the context of $r_{orig} = 0.4$ not so much. They rather question the recognition method's performance. We find that this can be problematic in cases where r_{orig} is not close to 1. We therefore suggest considering the relative anonymized recognition rate $r_{rel} = \frac{r_{anon}}{r_{orig}}$.

V. UTILITY EVALUATION

Anonymizations make a privacy-utility-trade-off and most of them acknowledge this by claiming both privacy and utility features in their abstracts or introductions. However, while almost all of them evaluate the privacy protection features, many do not evaluate their claims on utility. Tab. IV gives an overview over which utility features are claimed in the papers we analyzed and which of them are evaluated. Additionally, we analyzed utility evaluations that do not belong to a specific anonymization method. Specifically, [34] evaluates the usability of anonymized sensor data for gait based authentication and [39] considers the impact of voice anonymization on the utility of voice recordings.

We find that the utility features that the analyzed anonymization methods and evaluations mention vary widely. They depend on the biometric trait and the specific use-case. In the following subsections we want to taker a closer look at the different evaluations of two utility features, namely naturalness and the preservation of attributes.

A. Naturalness

The first utility feature that we find in multiple evaluations is naturalness. Naturalness means that anonymized data cannot be distinguished from not-anonymized data at first glance because it appears realistic. Most evaluations consider this property to be subjective to humans and therefore conduct **user-studies** to get a measurement of the naturalness of anonymized data. In evaluations of both [29] and [31] participants are asked to rate the naturalness of the anonymized videos on a scale from 1 to 5. However in [29], the participants only see the anonymized video while the participants in [31] see both the original and anonymized versions and are able to compare. This could potentially lead to both distinction and confirmation bias in the second study. [29] also includes another study in which participants are tasked to choose the anonymized video from a set of five videos. A highly natural anonymized video would mean that participants choose correctly in below or around 20% of cases.

TABLE IV
CLAIMED UTILITY OF ANONYMIZATION METHODS

Anonymization	Claimed utility	Evaluation?
k-Same [19]	"many facial characteristics remain"	—
AnonFACES [21]	high naturalness	—
	preserve age, gender, skin tone & more	—
Face-Off [22]	"acceptable cost for the user"	✓
Fawkes [12]	no significant distortions	—
DeepPrivacy [11]	high naturalness	✓
	seamless transition	—
Gait 1 [29]	naturalness	✓
Gait 2 [31]	naturalness	✓

A different approach to evaluate naturalness is taken in the evaluation of DeepPrivacy [11]. There, the anonymized facial images are run through a face **detection** system. The reason is that when a system that is designed to detect real natural faces in images can detect faces in the anonymized images, the faces can be considered natural and realistic. The authors therefore consider the percentage of faces that are detected in the anonymized images of those that are detected in the original images to be a metric of naturalness. This however can be problematic when the definition of what humans consider as a natural/realistic image and what the face detection method detects as a face diverge. It especially only considers the actual anonymized face to be natural not its integration into the complete image.

We find that to evaluate the naturalness of anonymized data time-intensive and expensive user-studies have to be conducted. However more straightforward experiments using a detection method can give a good idea about naturalness of the modified data, but not necessarily its integration into the surrounding data. There has been research into general modification detection for images, e.g. [45]. Perhaps these could be adapted to be used to quantify the naturalness of entire anonymized images.

B. Attribute preservation

Another utility feature is attribute preservation. It means that specific attributes of the original data are still present in the anonymized data. This means that some information (containing the attributes) in the original data has to be left unmodified by the anonymization method while other information (everything identifying the individual) has to be removed. The specific attributes that anonymization methods try to preserve depend on the biometric trait and the use-case of the anonymization.

In [39] the attribute to be preserved are the words that are spoken in the voice recordings. This means that the goal of the anonymization method is to remove the information on the identity while preserving information on the text from the original recording which includes information on both text and identity of the speaker. Whether the anonymization is successful in removing information on the identity is the privacy protection evaluation which we discussed in Sec. IV. The evaluation if the anonymization methods are successful at preserving the spoken text however is a utility evaluation. It is done by using an automatic speech recognition system to transcribe both the original and anonymized voice recordings and measuring the word error rate (WER). The lower the error rate, the less information on the text was lost during the anonymization which means the anonymization method is better at preserving this attribute.

A similar evaluation strategy is used in [34]. There, it is evaluated whether the anonymization method to add noise to sensor data impacts authentication methods using this data to verify the identity of the user. Here, the attribute to be preserved is the identity of the user while the attributes to be removed are the device identity and information from which key presses can be inferred. The evaluation is done by first measuring the utility on the original and then the anonymized data. This means here that the authentication performance which is the percentage of cases where the user's identity is successfully verified is measured. The performance for original and anonymized data is then compared. A performance of anonymized data close to that of the original data means that the anonymization method is successful at preserving these attributes while a significantly lower performance means that the anonymization failed to preserve the attributes.

We find that in general, attribute preservation is evaluated by using a system that can extract the specific attributes from the data and then comparing the extractions from original and anonymized data. This means for example that AnonFACES' claim to preserve age, gender, skin tone and emotional expression in their anonymized faces could be evaluated by using a facial attribute extractor like DeepFace [9]. They could extract the desired attributes from the original face and the corresponding anonymized face and compare whether the extracted attributes are in fact preserved.

VI. Related Work

An area of related work that we want to take a look at is biometric template protection. While we focused on biometric traits where identification of individuals can be done from a distance and potentially without consent, there are also biometric traits where this is more difficult. For example, identification based on fingerprint and iris is often used in authentication systems. Never the less, these systems still have to protect the privacy of the individuals when working with recordings of their biometric features which is called template protection. Contrary to the evaluations that we analyzed in this paper, evaluation of template protection methods has been standardized in ISO/IEC 24745 [13]. The standard defines the privacy protection goal of biometric template protection using three aspects. The first, *Irreversibility*, means that it is difficult to recover the biometric features of an individual from the saved biometric template. *Unlinkability* means that the biometric template cannot be linked to the individual from whom they were derived. Lastly, *Confidentiality* means that the saved templates are protected against unauthorized access. While the last aspect is not relevant in the evaluations that we analyzed in this paper, we can find equivalents of the other aspects. Irreversibility is closely related to the Restoration attacker goal and Unlinkability is related to the Re-identification attacker goal, both of which we dicussed in Sec. IV-A.

Another area of related work are attackers on specific biometric anonymization approaches. In Sec. IV-A, we considered attackers that are aware that an anonymization has taken place and try to revert it. Here, we would like to introduce some papers that evaluate biometric anonymizations by creating attackers on the specific methods and measuring their success.

In [46], the authors construct a CNN with the goal of reversing an anonymization that blurs faces. The authors exploit that faces are highly structured and share key facial landmarks such as eyes and mouths. They find that their deblurring approach results in de-anonymized images that have a lower identity distance (as measured by FaceNet [47]) to the original image than other deblurring approaches and the blurred image.

Similarly, the authors of [48] construct a neural network to attack the facial anonymization methods pixelation, blurring and P3 (a jpeg-specific approach) [49]. However, instead of creating de-anonymized images that are supposed to be close to the original images, they create a recognition system that directly works on the anonymized images (for a specific anonymization method). They find that their recognition method is often able to achieve recognition performance above 50% percent depending on anonymization method and intensity and for example about 70% for 4x4 pixelation on the FaceScrub dataset [50].

The authors of [51] introduce an attacker that reverses facial anonymizations, specifically pixelation, blurring and noise addition. It works by first detecting an obscured face, then determining the anonymization method, estimating the parameters of the anonymization and then applying an anonymization and parameter specific restoration method. The authors find that their anonymization method classification can determine the correct anonymization method in over 90% of cases and that the de-anonymized images can be correctly re-identified significantly more often then the anonymized images.

VII. Conclusion & Future Work

We find that currently there is no standard method to evaluate biometric anonymization methods. While the evaluations that we analyzed show similarities in their approaches to evaluate the privacy protection of anonymizations, there are still significant differences. We also find that there are some problems with the current state-of-the-art, especially a weak attacker model that does not consider more advanced attacks on the anonymization methods. Also, while many anonymization methods make claims about their utility, a significant portion does not evaluate these claims at all. Considering that more recent work focuses on improving utility while not sacrificing privacy, this lack of evaluation is very problematic.

In conclusion, the lack of standard evaluation methodology means that is very difficult to compare different anonymization methods. The weak attacker model means that anonymization methods might not be as effective as they claim to be. This means that anonymizations that are actively in use at the moment such as blurring faces in television, might not actually protect the privacy of the shown individual.

In the future, we would like to see an evaluation framework based on current evaluations that considers the problems that we pointed out. Testing all of the anonymization methods on a single framework would allow them to be properly compared and decisions on which anonymization to use could be more informed.

REFERENCES

[1] A. Cavailaro, "Privacy in video surveillance [in the spotlight]," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 168–166, 2007.

[2] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, 2016.

[3] A. Senior and A. W. Senior, *Protecting privacy in video surveillance*, vol. 1. Springer, 2009.

[4] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication* (D. Zhang and A. K. Jain, eds.), (Berlin, Heidelberg), pp. 731–738, Springer Berlin Heidelberg, 2004.

[5] R. Jafri and H. Arabnia, "A survey of face recognition techniques," *JIPS*, vol. 5, pp. 41–68, 06 2009.

[6] T. Sakai, T. Kanade, M. Nagao, and Y. ichi Ohta, "Picture processing system using a computer complex," *Computer Graphics and Image Processing*, vol. 2, no. 3, pp. 207–215, 1973.

[7] R. Shyam and Y. N. Singh, "A taxonomy of 2d and 3d face recognition methods," in *2014 International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 749–754, IEEE, 2014.

[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[9] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, 2020.

[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)*, pp. 1491–1500, IEEE, 2017.

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

[12] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," 2020.

[13] S. Rane, "Standardization of biometric template protection," *IEEE MultiMedia*, vol. 21, no. 4, pp. 94–99, 2014.

[14] K. Lander, V. Bruce, and H. Hill, "Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 15, no. 1, pp. 101–116, 2001.

[15] 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*, pp. 2373–2380, IEEE, 2009.

[16] J. Tekli, B. al Bouna, R. Couturier, G. Tekli, Z. al Zein, and M. Kamradt, "A framework for evaluating image obfuscation under deep learning-assisted privacy attacks," in *2019 17th International Conference on Privacy, Security and Trust (PST)*, pp. 1–10, 2019.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[18] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," 2016.

[19] 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.

[20] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The feret evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[21] M.-H. Le, M. S. N. Khan, G. Tsaloli, N. Carlsson, and S. Buchegger, "Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security," pp. 87–100, 11 2020.

[22] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, "Face-off: Adversarial face obfuscation," *arXiv preprint arXiv:2003.08861*, 2020.

[23] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018.

[24] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[27] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, "Dsfd: Dual shot face detector," 2019.

[28] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, 2016.

[29] Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi, "Anonymization of gait silhouette video by perturbing its phase and shape components," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1679–1685, 2019.

[30] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi, "Gait recognition: Databases, representations, and applications," *Wiley Encyclopedia of Electrical and Electronics Engineering*, pp. 1–15, 1999.

[31] N.-D. T. Tieu, H. H. Nguyen, H.-Q. Nguyen-Son, J. Yamagishi, and I. Echizen, "An approach for gait anonymization using deep learning," in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, IEEE, 2017.

[32] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4, pp. 441–444, 2006.

[33] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *2011 18th IEEE International Conference on Image Processing*, pp. 2073–2076, IEEE, 2011.

[34] R. Matovu, A. Serwadda, D. Irakiza, and I. Griswold-Steiner, "Jekyll and hyde: On the double-faced nature of smart-phone sensor noise injection," 08 2018.

[35] A. Primo, V. V. Phoha, R. Kumar, and A. Serwadda, "Context-aware active authentication using smartphone accelerometer measurements," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 98–105, 2014.

[36] J. Qian, H. Du, J. Hou, L. Chen, T. Jung, and X.-Y. Li, "Hidebehind: Enjoy voice input with voiceprint unclonability and anonymity," in *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, pp. 82–94, 2018.

[37] D. Sundermann and H. Ney, "Vtln-based voice conversion," in *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*, pp. 556–559, IEEE, 2003.

[38] J. chieh Chou, C. chieh Yeh, and H. yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," 2019.

[39] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2802–2806, IEEE, 2020.

[40] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[41] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[42] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Website of fawkes: Image cloaking for personal privacy," 2021.

[43] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[44] A. Stolerman, R. Overdorf, S. Afroz, and R. Greenstadt, "Breaking the closed-world assumption in stylometric authorship attribution," in *Advances in Digital Forensics X* (G. Peterson and S. Shenoi, eds.), (Berlin, Heidelberg), pp. 185–205, Springer Berlin Heidelberg, 2014.

[45] S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, "Detecting photoshopped faces by scripting photoshop," 2019.

[46] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, "Deep semantic face deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8260–8269, 2018.

[47] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

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

[49] M.-R. Ra, R. Govindan, and A. Ortega, “P3: Toward privacy-preserving photo sharing,” 02 2013.

[50] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 343–347, 2014.

[51] N. Ruchaud and J.-L. Dugelay, “Automatic face anonymization in visual data: Are we really well protected?,” *Electronic Imaging*, vol. 2016, pp. 1–7, 02 2016.