PrivacyProber: Assessment and Detection of Soft–Biometric Privacy–Enhancing Techniques

Peter Rot ©, Member, IEEE, Klemen Grm ©, Member, IEEE, Peter Peer ©, Senior Member, IEEE, and Vitomir Štruc ©, Senior Member, IEEE

Abstract—Soft–biometric privacy-enhancing techniques represent machine learning methods that aim to: i) mitigate privacy concerns associated with face recognition by suppressing soft-biometric attributes in facial images (e.g., gender, age, ethnicity) and ii) make unsolicited extraction of sensitive personal information infeasible. Because such techniques are increasingly used in real-world applications, it is imperative to understand to what extent the privacy enhancement can be inverted and how much attribute information can be recovered from the privacy-enhanced images. While these aspects are critical, they have not been investigated in the literature so far. In this paper, we therefore study the robustness of state-of-the-art soft-biometric privacy-enhancing techniques to attribute recovery attempts. We propose PrivacyProber, a high-level framework for restoring soft-biometrics from privacy-enhanced images and apply it for attribute recovery in comprehensive experiments on three public face datasets (LFW, MUCT, and Adience). Our experiments show the proposed framework is able to restore a considerable amount of suppressed information, regardless of the privacy-enhancing technique used (e.g., adversarial perturbations, synthetic data, etc.), and that there are significant differences between the considered privacy models. These results point to the need for novel mechanisms to improve the robustness of existing techniques and secure them against adversaries trying to restore suppressed information. We also demonstrate that PrivacyProber can be used to detect privacy enhancement (under black-box assumptions) with high accuracy.

Index Terms—Computer vision, deep learning, face recognition, privacy, robustness analysis.

I. INTRODUCTION

FacIAL images represent a rich source of information, from which a multitude of attributes can be extracted automatically using contemporary machine learning models, including gender, age, ethnicity, affective state, or even body mass indices [1], [2], [3], [4], [5], [6]. However, with the rapid proliferation of automatic recognition techniques in facial analytics, privacy–related concerns have also emerged and now represent a key challenge with regard to the trustworthiness of the technology [7], [8].

To address these concerns, researchers are increasingly looking into privacy–enhancing mechanisms capable of ensuring a trade–off between the utility of the data for facial analytics, on the one hand, and privacy protection, on the other [9], [10], [11], [12], [13], [14]. Examples of such mechanisms include: i) deidentification techniques that try to conceal the identity of the subjects in the data, while retaining other useful attributes and information [11], [15], [16], but also ii) so-called soft–biometric privacy–enhancing techniques, which aim to perturb (transform, modify) facial images in a way that makes it difficult (or impossible) for automatic machine learning techniques to infer sensitive attributes (e.g., gender, age, ethnicity), while making only minimal changes to the visual appearance of the images. While a considerable amount of research has been done on deidentification technology, soft–biometric privacy–enhancing techniques represents a relatively new research topic that has so far been underexplored.

The importance of soft–biometric privacy–enhancing techniques is highlighted through various application scenarios where concealing soft-biometrics attributes, such as demographicals, is critical. For example, when sharing personal images online (e.g., on social media), soft–biometric privacy–enhancement can help to avoid automatic user profiling, attribute extraction and other unsolicited forms of processing. Similarly, in face verification systems, where only the identity information is needed, such techniques can conceal information on other attributes that are not required for verification purposes. Such privacy–enhancing mechanisms give people the possibility/freedom to obscure selected (soft-biometric) attribute information from facial images before making them publicly available, instead of making all attributes public by default [11]. These mechanisms represent pivotal tools for addressing societal expectations about the appropriate use of personal data, but also for meeting safeguards and standards defined in data–protection legislation and privacy acts, such as, GDPR [17], CCPA [18], BIPA [19] and others [11]. As noted in relevant recent surveys [11], [20], other types of techniques exist that deal with different aspects of privacy (e.g., encryption, training, synthetic data, etc.). However, the focus of this paper is exclusively on soft–biometric privacy–enhancement.

Several powerful solutions have been proposed in the literature for ensuring soft–biometric privacy and preventing automatic extraction of facial attributes, including synthesis–based...
techniques [21], auto–encoder based models [9], [22] or adversarial perturbations [10], [23]. While these solutions have been shown to successfully obscure soft–biometric information for selected attribute classifiers, experimental performance evaluations are typically limited to vanilla (or zero–effort) evaluation scenarios, where no attempt is made to reconstruct the obscured attributes. Due to this practice, it is not clear if the privacy levels reported in the literature generalize to real–world applications, where a potential adversary may exploit additional knowledge and invest considerable effort and resources to recover the concealed information. Such zero–effort evaluations, hence, raise questions about the reliability of existing privacy models1 and their sensitivity to attribute recovery attempts. Comprehensive reliability studies are, therefore, critical for better understanding the capabilities of contemporary privacy–enhancing techniques and have implications for their deployment in practice. However, to the best of our knowledge, such studies are largely missing from the literature.

In this paper, we address this gap and explore possibilities for recovering obscured (concealed, suppressed) information from privacy–enhanced facial images with the goal of assessing the reliability and robustness of existing soft–biometric privacy models. To facilitate the study, we develop an attribute recovery framework, called PrivacyProber, which uses various types of image transformations (learned over clean, non–tampered images) to reconstruct suppressed information from privacy–enhanced facial images. The proposed framework is based on minimal (black box) assumptions and requires no examples of privacy–enhanced images to restore soft–biometric information, which makes it applicable to a wide variety (of conceptually different) soft-biometric privacy models. To demonstrate the feasibility of our framework, we conduct extensive experiments with multiple state–of–the–art privacy models and over three publicly available face datasets. The results of our experiments show that PrivacyProber is able to recover a significant amount of suppressed attribute information (and invert the privacy–enhancement to a considerable extent), suggesting that sensitivity to reconstruction attacks is still a considerable issue with (most) existing privacy–enhancing techniques. As an additional contribution, we also show that the proposed framework can be used to detect privacy–enhancement in facial images, pointing to another threat vector with respect to existing privacy models that allows for flagging tampered images and treating them differently from non–tampered data, e.g., using manual screening.

As part of our reliability study (illustrated in Fig. 1), we make the following main contributions:

- We conduct, to the best of our knowledge, the first comprehensive investigation into the robustness of soft–biometric privacy–enhancing techniques with respect to attribute recovery attempts. We show that despite the significant progress in this area, a considerable amount of concealed information can still be recovered from (most) privacy–enhanced images, but also that there are significant differences between the tested models.

- We propose PrivacyProber, a framework for attribute recovery from privacy–enhanced facial images that relies on dedicated reconstruction schemes build around inpainting, denoising, face parsing models and self-supervised adversarial defenses, and requires no access to the evaluated privacy–enhancing techniques or prior knowledge about their inner workings.

- We present novel methodology to evaluate the robustness of soft–biometric privacy models. Specifically, we introduce an attribute–recovery robustness (ARR) score that reflects the robustness of privacy–enhancing techniques by comparing attribute–classification accuracy over reference and attribute–recovered images. We demonstrate the value of ARR scores through extensive experimentation with multiple privacy models and datasets.

- We show that PrivacyProber can be used to detect privacy–enhancement in facial images and propose a novel detection approach, called APEND (Evidence Aggregation for Privacy–Enhancement Detection), that consolidates differences between the predictions of an attribute classifier applied before and after a series of attribute-recovery attempts. We show that APEND facilitates the detection of privacy–enhanced images, while ensuring highly competitive performance.

**II. BACKGROUND AND RELATED WORK**

In this section, we provide background information on soft–biometric privacy-enhancing techniques and survey the most relevant prior work. For an in-depth coverage of the broader topic of visual privacy and privacy protection of facial images, we refer the reader to some of the excellent and comprehensive recent surveys in this area, e.g., [11], [20], [24], [25], [26], [27].
A. Problem Definition and Model Taxonomy

Assume an original bona fide face image, $I_{or} \in \mathbb{R}^{w \times h}$, and an attribute classifier $\xi_0 : \mathbb{R}^{w \times h} \rightarrow \{a_1, a_2, \ldots, a_N\}$, with the attribute labels $\{a_i\}_{i=1}^N$ corresponding to classes $\{C_1, C_2, \ldots, C_N\}$. Soft–biometric privacy–enhancement, $\psi$, aims to produce privacy–enhanced images, $I_{pr} = \psi(I_{or})$, from which the class labels $a_i$ cannot be correctly predicted by $\xi_0$ with high confidence. In general, the goal of $\psi$ is to obscure attribute information from machine learning models, but also to ensure that the appearance of the perturbed image is as close to the bona fide one as possible, so that the visual content appears similar for human observers, i.e., $\min\|I_{pr} - I_{or}\|_{L_p}$, where $\|\cdot\|_{L_p}$ is a $L_p$ norm. These characteristics are illustrated on the left part of Fig. 1.

As suggested in [28], existing soft–biometric privacy models can in general be grouped into: i) techniques that try to induce incorrect attribute predictions, and ii) techniques that try to generate approximately equal class probabilities for all attributes. Solutions from the first group aim to enhance privacy by inducing misclassifications, i.e., $\xi_0(I_{pr}) \neq \xi_0(I_{or})$, through various mechanism, e.g., adversarial perturbations and related strategies. With these techniques, an incorrect attribute label is typically predicted from $I_{pr}$ with high probability. Techniques from the second group, on the other hand, commonly rely on (input–conditioned) synthesis methods that enhance privacy by altering image characteristics in a way that makes attribute predictions unreliable, i.e., $p(C_1 | I_{pr}) \approx p(C_2 | I_{pr}) \approx \cdots \approx p(C_N | I_{pr})$, where $p(C_i | I_{pr})$ represents the posteriors of the attribute classes given the privacy–enhanced image $I_{pr}$.

B. Soft–Biometric Privacy Models

A significant amount of research has been presented in the literature recently that addresses different problems related to soft–biometric privacy [9], [12], [14], [23], [29], [30], [31], [32]. Mirjalili and Ross [33], for example, were among the first to explore privacy–enhancing techniques that perturb gender information in facial images. The solution, developed by the authors, first performs Delaunay triangulation (over face feature points) to decompose faces into a set of triangles that can be manipulated with the goal of privacy protection. Next, the texture within the triangles is optimized, such that a selected gender classifier generates unreliable predictions, while the input image is perturbed ever so slightly. The authors showed that their solution results in image manipulations that obscure gender information well, while having only a minimal impact on facial appearance and, consequently, on verification performance.

Another notable technique for soft–biometric gender privacy was introduced by Mirjalili et al. in [22]. This work described so-called Semi–Adversarial Networks (SANs), i.e., machine learning models that rely on conditional image synthesis to conceal gender information in facial images. Conceptually, SANs represent convolutional auto–encoders that are paired with two distinct discriminators that steer the synthesis process – one for enforcing gender privacy and the second for retaining verification accuracy (i.e., appearance similarities). The proposed SAN models were demonstrated to be capable of efficiently suppressing gender information, while retaining the data utility for identity verification tasks. To improve the generalization capabilities of SANs to unseen attribute classification models, FlowSAN models [9] were introduced in the follow-up work of the same authors. The main idea behind these models was to employ multiple SAN transformations one after the other, with the goal of making the privacy enhancement applicable with arbitrary gender classifiers. FlowSANs were demonstrated to generalize better than SANs, while providing a good trade–off between privacy enhancement and utility preservation.

In [12], Marjalili et al. introduced PrivacyNet [12], an extended SAN model build around Generative Adversarial Networks (GANs). Unlike competing solutions, PrivacyNet was shown to be capable of soft–biometric privacy enhancement with respect to different (non-binary, continuous) facial attributes, including race, age and gender. The model, hence, generalized the fundamental idea behind SAN models to arbitrary soft–biometric attributes (as well as their combinations). We note at this point that the SAN–based family of algorithms is not based on adversarial perturbations, but relies on facial synthesis facilitated by auto–encoders driven by a number of competing discriminators.

An adversarial approach for privacy enhancement of $k$ facial attributes ($k$–AAP) was described in [10]. The proposed approach tries to infuse facial images with adversarial noise with the goal of obscuring selected soft-biometric attributes, while preserving others. $k$–AAP relies on the established Carlini Wagner $L_2$ attack [34] and was demonstrated to achieve competitive results with attribute classifiers considered during the construction of the adversarial noise. Similarly, as related adversarial methods, $k$–AAP results in image manipulation that effectively conceal attribute information for machine learning models, while being barely detectable by human observers. Much like the original SAN models, $k$–AAP performs best with a known attribute classifiers, but struggles somewhat with unseen classification models. A conceptually similar idea involving adversarial noise was also investigated in [23] where the Fast Gradient Sign Method (FGSM) method [35] was utilized to explore the robustness of facial feature perturbations.

C. Evaluation of Soft–Biometric Privacy Models

Quantifying the level of privacy enhancement is a challenging task and requires well–defined evaluation methodologies and corresponding performance scores that provide insight into the characteristics of the tested privacy models. To quantify performance, the majority of prior work in this area exploits automatic recognition techniques trained for extracting various facial attributes. Recognition experiments are then performed on the original and privacy–enhanced images, and differences in the observed classification accuracies are used for performance reporting [12], [14], [31].

Additionally, several scalar performance scores have also been proposed in the literature. Othman and Ross [21], for example, introduced gender suppression levels, a performance score for measuring the success of a synthesis–based privacy–enhancing technique with respect to the induced
applied

We assume that privacy–enhancement is achieved by making the image more difficult to interpret. Terhörst et al. proposed the privacy-gain identity–loss coefficient (PIC), which measures the gain of privacy with respect to a chosen visual attribute (e.g., gender) but takes the retained verification utility into account as well. The same authors also proposed [32] a correct classification rate (CCR) to score the success of privacy–enhancing algorithms. More recently, the authors of [30] proposed a set of evaluation protocols with associated performance measures to enable reproducible research on soft–biometric privacy.

While the evaluation methodologies reviewed above provide initial estimates about the performance of biometric privacy–enhancing techniques, they only assume zero–effort evaluation scenarios. In this study, we built on the presented work and propose a more comprehensive evaluation methodology that also considers attribute reconstruction attempts when scoring privacy models. We introduce a novel performance measure that captures the robustness of the models and offers insight into the difficulty of recovering concealed attribute information.

D. Detection of Privacy Enhancement

Soft–biometric privacy-enhancing techniques introduce changes to the visual characteristics of facial images and can, hence, be seen as a form of image tampering. While a significant amount of work has been described in the literature to detect such tampering, e.g., [37], [38], [39], such detection techniques have mainly been studied only within the information-forensics community. The problem of detecting privacy enhancement, on the other hand, is new and, except for the PREM work in [28], where a detection model based on face super-resolution and prediction divergence was presented, has not been studied widely in the open literature. Different from [28], we demonstrate in this paper that the process of attribute recovery, facilitated by the proposed PrivacyProber, can also be exploited to develop efficient privacy-enhancement detectors and that the information aggregated through different instantiations of PrivacyProber lead to highly robust performance.

Finally, we note that because some privacy models are based on adversarial perturbations, the problem of detecting soft–biometric privacy enhancements is also partially related to adversarial attack detection methods [40], [41], [42], [43], [44]. However, because soft–biometric privacy model also include synthesis–based methods (among others), the problem of detecting such modifications is considerably broader.

III. PRIVACYPROBER

In this section, we describe the proposed PrivacyProber, a framework for the recovery of suppressed soft–biometric attribute information from privacy enhanced facial images. We discuss multiple contributions, including novel schemes for recovering attribute information under minimal assumptions, and a novel detector for privacy-enhanced images.

A. Overview of PrivacyProber

Existing evaluation schemes for soft–biometric privacy models typically compare the performance of an attribute2 classifier $\xi_a(\cdot)$, when applied to the original $\xi_o(I_{or})$ and privacy enhanced images $\xi_a(I_{pr})$. While the observed performance difference provides a first estimate of the level of privacy ensured by the privacy models, it also assumes that no attempt is made to recover the information initially contained in the input image $I_{or}$. As a result, the privacy levels empirically determined with such vanilla evaluation methodology may be overestimated and not representative of the actual capabilities of existing privacy models.

The proposed PrivacyProber tries to address this problem by facilitating performance assessments beyond zero–effort evaluation experiments. Specifically, PrivacyProber seeks to recover concealed image information by transforming privacy–enhanced images $I_{pr}$ in such a way that the privacy enhancement is reversed, or formally, such that the classifier $\xi_a$ generates the same predictions for the original, $I_{or}$, and recovered images, $I_{re}$, i.e.:

$$\xi_a(I_{re}) = \xi_a(\chi(I_{pr})),$$

where $I_{re} = \chi(I_{pr})$ and $\chi(\cdot)$ is the transformation applied by the PrivacyProber. By estimating the level of privacy through performance differences of the attribute classifier $\xi(\cdot)$ applied to the original $\xi(I_{or})$ and attribute–recovered images $\xi_a(I_{re})$, where $I_{re} = \chi(I_{pr})$, a more informative estimate of privacy–enhancement performance (or robustness) can be obtained. The main idea is presented in Fig. 1.

While in general, the transformation $\chi(\cdot)$ could be learned in a data–driven manner by defining a loss penalizing the difference in classification outputs between $\xi(I_{or})$ and $\xi(\chi(I_{pr}))$, such an approach is i) model-specific and not universally applicable, ii) requires access to training data for each considered privacy model, and iii) warrants prior knowledge about the internal mechanism governing the privacy–enhancement procedure. In this paper we, therefore, follow a more general approach and design PrivacyProber under the following (minimal) assumptions:

- **Black–box privacy models**: We assume no information about (or access to) the privacy models is available. We only exploit the fact that the privacy models aim to suppress soft–biometrics information, while making minimal changes to the appearance of the images.
- **Target domain**: We assume that privacy–enhancement is applied within a fixed image domain, i.e., on facial images, and not on images of arbitrary scenes.

Thus, we construct PrivacyProber based on a set of i) generative and ii) domain-specific transformations, which can be used separately or in sequence, as illustrated in Fig. 2. In the next sections we propose several possibilities for recovering suppressed soft–biometric attributes from privacy–enhanced images using such transformations.

2Common attributes studied in the literature are gender, age, and race.
B. Generative Transformations

Soft–biometric privacy–enhancing techniques modify or partially corrupt the visual content in the original images $I_{pr}$ with the goal of reducing the utility of the data for some targeted attribute classification task. The use of generative transformations\(^3\) for mitigating such image tampering is motivated by the fact that prior information about (clean, non–tampered) original face images can be incorporated efficiently into generative models (without the need for examples of privacy–enhanced images) and exploited for attribute recovery. Similar approaches have also proven useful for protection against adversarial attacks [45],[46]. We propose three generative transformations for implementing PrivacyProber using dedicated schemes based on i) inpainting, ii) denoising, and iii) image reconstruction.

1) Attribute Recovery Through Inpainting: Inpainting models typically aim to fill in missing pixels in a damaged image and to restore the corrupted content. Our goal, on the other hand, is to recover a complete image $I_{re}$ from $I_{pr}$ with attribute information restored and not just a small portion of the data. We, therefore, design a novel attribute-recovery scheme based on inpainting for this task that sequentially inpaints small parts of the image at the time and then aggregates the results to recover the complete image. The basic assumption with this attribute–recovery strategy is that inpainting can restore (clean) non–tampered content by inferring pixels from contextual information, even if this information was tampered with a soft–biometric privacy–enhancing model.

The proposed recovery procedure, illustrated in Fig. 3, starts with the privacy–enhanced image $I_{pr} \in \mathbb{R}^{w \times h}$ and a set of $N$ binary masks $B_i \in \mathbb{R}^{w \times h}$, where $w$ and $h$ again denote the width and height of the image, respectively, and $i \in \{1, 2, \ldots, N\}$. The binary masks are initialized as matrices of all ones. Next, a chess–like pattern composed of multiple square regions (of size $d \times d$, where $d \ll \{w, h\}$) is constructed and placed into the initialized masks. Pixels in $B_i$ corresponding to the chess–like pattern are set to zero. As illustrated in the top row of Fig. 3, the pattern is positioned in the top right corner for the first mask and then shifted $N$–times in a sliding–window manner (with overlap), such that all positions are traversed in both the horizontal and vertical direction. To facilitate inpainting, the constructed binary masks $B_i$ are then used to remove pixels from the privacy–enhanced image, i.e.:

$$I_{pr}^{(i)} = I_{pr} \odot B_i, \quad \text{for } i \in \{1, 2, \ldots, N\},$$

where $\odot$ represents the Hadamard product and $I_{pr}^{(i)}$ denotes the input image masked by $B_i$. In the next step, all pixels set to zero are reconstructed using a predefined inpainting model, which given a set of masked images $\{I_{pr}^{(i)}\}_{i=1}^{N}$ generates a set of corresponding (partially) recovered images $\{I_{re}^{(i)}\}_{i=1}^{N}$. Finally, the complete attribute–recovered image $I_{re}$ is reconstructed from the inpainted regions only:

$$I_{re} = \left[ I_{re}^{(i)} \odot \bar{B}_i \right]_{i=1}^{N},$$

where $\bar{B}_i$ is an inverted version of $B_i$ used to exclude the original (non–inpainting) areas from $I_{re}^{(i)}$ and $\left[ \cdot \right]$ represents an averaging operation over non–zero pixels. This type of averaging is needed due to the overlap in the masked regions between binary masks. We denote the presented attribute–recovery procedure as $\chi_{in} : I_{pr} \mapsto I_{re}$ hereafter, and implement it using an off–the–shelf inpainting model.

2) Attribute Recovery Through Denoising: The second proposed approach to attribute recovery is based on denoising. Soft–biometric privacy–enhancing techniques typically aim to make only minute changes to the input images, $I_{pr}$, and alter their visual characteristics as little as possible. The changes introduced can, therefore, be well accounted for by the high–frequency part of the privacy–enhanced images $I_{pr}$. We model these high–frequency alterations as noise and try to remove them using a denoising procedure. Such denoising strategies have proven useful for removal of adversarial noise [47], and are, therefore, also expected to be useful for reversing soft–biometric privacy enhancement that shares characteristics with techniques based on adversarial examples. Similarly as in the previous section, we denote the denoising–based attribute–recovery procedure as $\chi_d : I_{pr} \mapsto I_{re}$ hereafter and again implement it using an off–the–shelf inpainting model.

\(^3\)Note that we use the term *generative transformation* loosely here to account for the transformation considered in this group, which include inpainting, auto–encoder, super–resolution and denoising.
3) Attribute Recovery Through Adversarial Defenses: As soft–biometric privacy–enhancing techniques often rely on adversarial noise, the robustness of privacy–enhancement can be also evaluated against existing adversarial defense methods (e.g., [48]). This means that the attribute-recovery component in the proposed PrivacyProber framework (whose main goal is to evaluate the robustness of privacy–enhancing methods) can also be implemented as an arbitrary adversarial defense algorithm. Here, we consider a self-supervised approach that relies on an auto-encoder for image reconstruction, as illustrated in Fig. 4. The auto–encoder, optimized for removal of adversarial noise from an input image, consists of an encoder \( E \) and a decoder \( D \). The privacy–enhanced image \( I_{pr} \) is provided as an input to \( E \), which compresses the image into a latent representation \( z = E(I_{pr}) \). The latent representation \( z \) is then decoded using \( D \) to produce the reconstructed image \( I_{re} \), i.e., \( I_{re} = D(z) \). The result of this process is, therefore, an image \( I_{re} \) that is typically free of high-frequency (adversarial) noise that often serves as means for the privacy enhancement. We denote the reconstruction-based attribute recovery procedure as: \( \chi_{a}: I_{pr} \mapsto I_{re} \) hereafter.

C. Domain-Specific Transformation

Instead of building on generative models, trained on clean non–tampered data, to restore facial attributes, another possibility is to base attribute recovery on specifics of the targeted image domain. With facial images, for example, automatic attribute inference should be based solely on the facial region and ignore other contextual information, e.g., background. With domain–specific transformations, we hence aim to incorporate information on facial semantics into the attribute–recovery procedure and exclude image regions irrelevant for attribute classification from the data.

We propose one domain–specific transform in this work that relies on face parsing. Specifically, given an arbitrary face parser \( f_p \), we extract facial–part information from the privacy–enhanced image \( I_{pr} \) and aggregate all part labels that correspond to the facial region into a binary mask \( B \). The labeled facial parts produced by the face parser and the corresponding binary mask are illustrated in Fig. 5. Once the binary mask is constructed, it is used to exclude background pixels with the goal of making attribute inference less susceptible to artifacts generated by the privacy enhancement:

\[
I_{re} = I_{pr} \odot B,
\]  

(4)
Thus, for a given privacy-enhanced input image, $I_{pr}$, a selected attribute classifier $\xi$ is expected to generate different posterior probabilities $p(C_k|I_{pr})$ than for the image processed through the proposed PrivacyProber $p(C_k|I_{re})$, where $C_k$ denotes the $k$th attribute class (as defined in Section II-A). By comparing the posteriors, it is therefore possible to determine whether an image has been tampered with or not. Based on this insight, we develop in this section a novel (proof-of-concept) approach for detecting privacy enhancements with the proposed PrivacyProber framework.

As illustrated in Fig. 6, the proposed approach aggregates evidence from multiple PrivacyProbers for Privacy-Enhancement Detection (APEND) and consists of the following steps:

- **Step 1. Attribute Recovery**: In the first step, the input image $I_{pr}$ is subjected to a series of $n$ PrivacyProbers $\{\chi_i\}_{i=1}^n$ that generate a corresponding set of $n$ attribute-recovered images $\{I_{re_i}\}_{i=1}^n$. The PrivacyProbers $\{\chi_i\}_{i=1}^n$ used in this procedure are assumed to be distinct and based on different transformations, e.g., functions or function compositions derived from $\{\chi_{in}, \chi_d, \chi_e, \chi_{br}\}$.

- **Step 2. Differential Distribution Scoring (DDS)**: Next, an attribute classifier $\xi$ is applied to the original input image $I_{pr}$ and the privacy-enhanced counterparts $\{I_{re_i}\}_{i=1}^n$, and posterior distributions are calculated, i.e., $p(C_k|I_{pr})$ and $\{p(C_k|I_{re_i})\}_{i=1}^n$. To quantify the differences in the computed distributions, we use the Chi-square distance, which is an established measure for comparing histograms [49]:

$$d_i(I_{pr}, I_{re_i}) = \sum_{k=1}^{N} \left( \frac{p(C_k|I_{pr}) - p(C_k|I_{re_i})}{\sqrt{p(C_k|I_{pr}) + p(C_k|I_{re_i})}} \right)^2,$$

where $i \in \{1, 2, \ldots, n\}$. See the right part of Fig. 6 for an illustration.

- **Step 3. Evidence Aggregation (EA)**: Finally, we aggregate the evidence generated through the $n$ attribute recovery steps into the final detection score $d_{fin}$ using a weighted linear combination, i.e.,

$$d_{fin} = \sum_{i=1}^{N} w_i d_i(I_{pr}, I_{re_i}),$$

where $w_i$ are balancing weights. The motivation for the aggregation step is to consider complementary recovery evidence for a more reliable detection score.

We note that APEND does not require training to facilitate tampering detection. Unlike the majority of modern detection schemes, which are trained in a discriminative manner with examples of bona-fide and tampered images, APEND is training free and knowledge-driven, i.e., designed around the characteristics of soft-biometric privacy models.

### IV. Evaluation of Privacy Models

In this section, we evaluate several state-of-the-art privacy enhancing techniques using standard vanilla methodology as well as the proposed PrivacyProber. The goal of these experiments is to provide insight into the performance, but more importantly robustness of biometric privacy models, and to demonstrate the importance of experimental evaluations that go beyond zero-effort recognition experiments. Furthermore, we also demonstrate how the proposed PrivacyProber can be used to detect privacy enhancement (tampering) in facial images – something that (to the best of our knowledge) has not been attempted widely in the open literature so far.

#### A. Considered Privacy Models

Four recent (soft-biometric) privacy–enhancing techniques are implemented for the experiments, i.e., the $k$-AAP method from [10], the FGSM–based technique from [50], the FlowSAN approach from [9], and PrivacyNet from [12]. Because FlowSAN can strike a balance between the level of privacy protection ensured and the preserved utility of the facial images, two different versions of the model are considered: i) one with three SAN models arranged sequentially one after the other (FlowSAN–3 hereafter), and ii) one with five sequential SAN models (FlowSAN–5). All five techniques are trained to obscure gender information, which is also the most frequently considered attribute in research addressing soft–biometric privacy [9], [14], [21], [32]. The techniques are selected for the experiments because of their state-of-the-art performance and the fact that they rely on different mechanisms for ensuring soft–biometric privacy. As such, they serve as a representative cross-section...
of existing techniques and contribute towards demonstrating the importance of evaluating biometric privacy models beyond zero–effort recognition experiments.

B. Datasets, Setup, and Performance Measures

To assess the performance of the considered privacy models, three publicly available face datasets are used, i.e., Labeled Faces in the Wild (LFW) [51], MUCT [53], and Adience [52]. The datasets ship with the needed gender labels and contain challenging facial images captured in a wide variety of imaging conditions. Moreover, they represent standard datasets for assessing the performance of existing privacy–enhancing techniques (e.g., [9], [12], [13], [14]) and are, therefore, also used in this work.

All images are roughly aligned prior to the experiments such that the faces are cropped to exclude background pixels and then rescaled to a standard size of 224 × 224 pixels. The preprocessed images are subjected to privacy enhancement and used in the experimental assessment to evaluate the following aspects of the privacy models:

1) Level of soft–biometric privacy enhancement: The performance of the evaluated privacy models is measured through gender (g) classification experiments with predefined classifiers on the original (o) and privacy–enhanced (p) images. The level of privacy enhancement achieved is determined by comparing ROC curves generated from the two image sets. Additionally, the gender suppression rate (SR) [29] is also reported and defined in this work as:

\[ SR = k \cdot (AUC_{gp} - AUC_{go}) \in [0, 1] \]  

where \( AUC_{go} \) and \( AUC_{gp} \) denote the area under the ROC curve before and after privacy enhancement, respectively, and the coefficient \( k \) corresponds to:

\[ k = \begin{cases} \frac{1}{(AUC_{go})^2}, & \text{if inverted decision} \\ \frac{1}{(AUC_{go} - 0.5)^2}, & \text{otherwise}. \end{cases} \]  

(8)

Note that \( AUC_{go} \in (0, 1] \), whereas the interval for \( AUC_{gp} \) depends on the type of privacy–enhancing technique, i.e., [0, 1] for methods with inverted decision (e.g., FGSM, k–AAP) and [0.5, 1] for methods aiming to induce equal class probabilities (e.g., FlowSAN, PrivacyNet). In certain cases, the \( AUC_{gp} \) score for methods aiming to induce equal class probabilities (i.e., the otherwise case in (8)) can fall slightly below 0.5, which suggests that soft-biometric information is still leaked and can be recovered by inverting decisions. This is addressed explicitly by using \( AUC_{gp} \rightarrow \max(AUC_{gp}, 1 - AUC_{gp}) \) before computing the SR score. The above definition allows us to report performance for (gender) privacy models that aim to induce misclassifications (i.e., invert classifier predictions for binary problems – invert ROC curves) as well as for models trying to induce random gender–classification probabilities (i.e., AUC scores of 0.5) using a single performance measure. A SR value of 1 indicates perfect attribute (gender) suppression, whereas a value of 0 implies that the suppression has no effect. Evaluating privacy models with SR scores and the approach presented above corresponds to zero-effort (vanilla) evaluation strategies commonly seen in the literature [9], [22].

2) Utility preservation: Soft-biometric privacy–enhancing techniques aim to retain as much of the original image information as possible, while altering the visual appearance of the input images only slightly. In accordance with standard evaluation methodology [9], [12], [21], [32] utility preservation is, therefore, assessed through verification experiments on the original and privacy–enhanced images, where minimal differences in performance are expected. Because there exists a trade–off between utility preservation and attribute suppression, a quantitative measure taking both tasks into account is reported for the experiments. Specifically, a modified version of the privacy–gain identity–loss coefficient (PIC) is used in this work [14], [29], [30], i.e.:

\[ PIC = SR - IL, \]  

(9)

where the identity loss \( IL \) is defined with the degradation in verification performance after the privacy enhancement:

\[ IL = \max \left( \frac{AUC_{vo} - AUC_{vp}}{AUC_{vo}}, 0 \right). \]  

(10)

In the above equations \( AUC_{vo} \) and \( AUC_{vp} \) are the AUC scores of the verification experiments (\( v \)) generated with the original (o) and privacy–enhanced (p) images. PIC is bounded to \([-1, 1]\), with a PIC score of 1 implying an ideal trade–off, i.e., no loss in verification performance and perfect attribute suppression.

3) Robustness: While attribute suppression and utility preservation are standard aspects of privacy models typically evaluated in zero–effort evaluation scenarios, robustness of the models to attribute recovery attempts has so far seen limited attention in the literature. Here, we use the proposed PrivacyProber to recover information suppressed by the studied privacy models. Identity verification and gender classification experiments are then conducted after attribute recovery (\( r \)) and the generated ROC curves are analyzed to assess robustness. A scalar robustness measure is derived from the ROC curves in this paper, i.e., the attribute–recovery robustness (ARR), and reported in the experiments, i.e.:

\[ ARR = g(AUC_{gp}) \cdot \frac{|AUC_{go} - AUC_{gp}|}{AUC_{go}} \]  

(11)

where the subscripts suggest that the AUC score was computed from the recovered (\( r \)) privacy–enhanced (\( p \)) or original (o) images and \( g(x) \), i.e.,

\[ g(x) = \begin{cases} 1, & \text{if } x \leq 0.5; \text{ inverted decision} \\ 2, & \text{otherwise} \end{cases} \]  

(12)

ensures that robustness is measured for all privacy models on the scale [0, 1]. Thus, ARR scores serve as a measure of robustness to attribute recovery attempts and take a value of 0 if after the recovery the same performance is achieved as with the original images and a value close to 1 if no
information can be inferred from the privacy-enhanced images. 

To facilitate the evaluation described above, the experimental datasets are split into training and testing parts. The training parts are used to train gender classifiers for each dataset and matchers for the verification experiments. Because the datasets are unbalanced with respect to gender, the number of male and female subjects in the training and testing sets is (approximately) balanced by randomly excluding images of the more represented gender. It is also made sure that at least two images per identity are present in the testing part for the verification experiments. To ensure a consistent evaluation setup across all datasets and be able to report confidence scores for the experiments, the test images are partitioned into 4 experimental splits. Gender recognition experiments are performed for every test image, while a fixed number of mated and non-mated comparison matches are performed for the verification experiments. Details on the experimental setup are provided in Table I.

### C. Baseline Evaluation of Privacy Models

The first series of experiments studies the performance of the considered privacy models using standard vanilla evaluation methodology. The goal of these experiments is to establish the baseline performance of the models and explore their characteristics.

1) Privacy versus Utility: From an operational point of view, a critical aspect of (soft–biometric) privacy enhancement is the trade–off between privacy protection and utility preservation the models ensure. The first part of our evaluation, therefore, looks at this trade–off through a series of recognition experiments.

   a) Baseline performance: To establish the baseline performance of the privacy models in terms of attribute suppression rates, a gender classifier \( \xi_{g} \) (a VGG16–based model with a two–class softmax at the top) is trained for each of the datasets and used to steer the privacy enhancement.\(^4\) In accordance with standard evaluation methodology [10], [23], the same classifier is also used to evaluate gender recognition performance with the enhanced images on each dataset. Similarly, a ResNet–50 face recognition model [55] is learned for the utility preservation experiments on images from the VGGFace2 dataset. Here, the output of the last fully connected layer of the learned model is utilized as the feature representation of the input face images. The computed representations are then matched with the cosine similarity measure in verification experiments.

   \(^4\)For the FlowSAN models several such classifiers are trained using different training data configurations similarly to [54].

Fig. 7 shows the ROC curves of the experiments on the three experimental datasets before and after privacy enhancement. Note that confidence intervals are not included to keep the plots uncluttered. Instead, standard errors are reported for the (scalar) performance scores in Fig. 8. As can be seen, the two privacy models that aim to induce misclassifications, \( k–AAP \) and FGSM, result in close to ideal gender suppression rates (SR) of 1 on all datasets, except for LFW, where the Carlini–Wagner attack used with \( k–AAP \) is not successful on several test images using our optimization parameters. The suppression rates for the FlowSAN models (which aim to produce random gender classification probabilities for each input image), on the other hand, depend on the number of SAN models used in the sequence. FlowSAN–3, for example, generates lower SR scores than FlowSAN–5, but is, therefore, retaining more identity information, as evidenced by the lower identity losses (ILs) in Fig. 8. PrivacyNet improves on both FlowSAN models both in terms of gender suppression and identity loss, and overall ensures the best overall privacy-utility trade-off among the evaluated synthesis-based privacy models. When comparing \( k–AAP \) and FGSM to the FlowSAN and PrivacyNet models, we observe that the former two models ensure higher overall PIC scores on most dataset. However, this is a consequence of the fact that the FlowSAN and PrivacyNet models do not simply invert classifier probabilities and, therefore, target a more challenging problem, which makes these models applicable to a wider range of application domains. Especially successful in terms of the privacy-utility trade-off is FGSM, which achieves PIC scores of close to 1 on all three datasets.

While most of the findings discussed above hold for all considered datasets, there are slight differences with the Adience dataset. Here, the verification as well as gender recognition performance with the original images is lower compared to the other two datasets. This is due to the characteristics of the dataset, which features real–world images with extreme appearance variations that differ significantly from those present in LFW and MUCT. As a result, even minor (additional) image degradations lead to significant performance drops, which is reflected in the relatively larger IL scores for all methods on this dataset.

b) Generalization to unseen classifiers: The results discussed above were generated with the same gender classifier, \( \xi_{g} \), that was also used for privacy enhancement on each dataset. To evaluate the generalization ability of the privacy models to unseen classification models, a ResNet–50 gender classifier, \( \xi_{g}^{\prime} \), is trained on LFW in the next series of experiments and deployed on the remaining two datasets, i.e., MUCT and Adience. Thus,
Fig. 7. ROC curves displaying the privacy versus utility trade–off ensured by the evaluated privacy models. Results for different datasets are shown in rows and for different privacy models in columns. All evaluated models are in general able to retain a significant portion of the verification performance (red curves), while resulting in different gender suppression rates (green curves). Note again that $k$–AAP and FGSM aim to induce misclassifications (i.e., invert classifier predictions for binary problems), whereas the FlowSAN models aim to produce random gender classification results. Best viewed in color.

Fig. 8. Scalar performance indicators for the privacy versus utility trade–off of the privacy models. Results are reported for three datasets and in the form of mean PIC (higher is better), SR (higher is better) and IL (lower is better) scores with corresponding standard errors computed over the 4 test splits - see Table I for the experimental setup.

Table II provides a summary of the PIC, SR and IL scores generated for this experiment. Here, the colored numbers show the relative change (marked as $\Delta$) in the computed scores when compared to the baseline performance from Fig. 8. As expected, most privacy models degrade in performance, both in terms of PIC and SR score compared to the baseline experiments. The relative drop is very severe for the adversarial techniques, where the gender-classification decisions do not get consistently inverted in most of the cases due to the change in the classifier. The PIC scores drop by 98.2% on Adience and by 91.5% on MUCT for $k$–AAP, and by 81.4% on MUCT and 97.4% on Adience for FGSM. The change in PIC score for FlowSAN–3 is less detrimental in comparison to the adversarial attacks, with a drop of 12.2% on MUCT and an increase by 42.1% on Adience. The performance drop increases for FlowSAN–5 to 58.7% on MUCT, but performance improves slightly (by 0.3%) on Adience. PrivacyNet is observed to generalize best among the tested privacy models, with PIC scores increasing by 57.7% on MUCT and dropping only by 6.2% on Adience.

Overall, the results of these experiments suggest that the relative change in performance differs significantly from model to model and is also dependent on the dataset used. The FlowSAN models, which exploit several gender classifiers for (soft–biometric) privacy enhancement, appear to generalize better across changes in the classification model used and may be preferred over $k$–AAP and FGSM if no assumptions can be made regarding the target classifier used with the final application.
TABLE II
GENERALIZATION ABILITY OF THE PRIVACY MODELS TO AN UNSEEN GENDER CLASSIFIER $\xi^u$

| Privacy Model | Dataset   | PIC | $\Delta$PIC (in %) | SR | $\Delta$SR (in %) | IL |
|---------------|-----------|-----|---------------------|----|-------------------|----|
| $k$–AAP       | MUCT      | 0.084 ± 0.037 | 91.5% $\downarrow$ | 0.087 ± 0.041 | 91.3% $\downarrow$ | 0.004 ± 0.006 |
|               | Adience   | 0.016 ± 0.031 | 98.2% $\downarrow$ | 0.149 ± 0.020 | 85.1% $\downarrow$ | 0.133 ± 0.019 |
| FGSM          | MUCT      | 0.186 ± 0.025 | 81.4% $\downarrow$ | 0.186 ± 0.025 | 81.4% $\downarrow$ | 0.000 ± 0.000 |
|               | Adience   | 0.026 ± 0.010 | 97.4% $\downarrow$ | 0.040 ± 0.015 | 96.0% $\downarrow$ | 0.014 ± 0.008 |
| FlowSAN–3     | MUCT      | 0.301 ± 0.149 | 12.2% $\downarrow$ | 0.306 ± 0.151 | 11.8% $\downarrow$ | 0.005 ± 0.003 |
|               | Adience   | 0.824 ± 0.075 | 42.1% $\downarrow$ | 0.897 ± 0.075 | 37.4% $\downarrow$ | 0.073 ± 0.015 |
| FlowSAN–5     | MUCT      | 0.338 ± 0.053 | 58.7% $\downarrow$ | 0.361 ± 0.051 | 57.1% $\downarrow$ | 0.023 ± 0.004 |
|               | Adience   | 0.718 ± 0.047 | 0.3% $\uparrow$   | 0.886 ± 0.050 | 0.2% $\uparrow$   | 0.169 ± 0.031 |
| PrivacyNet    | MUCT      | 0.941 ± 0.056 | 57.0% $\uparrow$   | 0.941 ± 0.056 | 73.0% $\uparrow$   | 0.000 ± 0.000 |
|               | Adience   | 0.854 ± 0.063 | 6.2% $\downarrow$  | 0.864 ± 0.066 | 6.1% $\downarrow$  | 0.010 ± 0.004 |

The reported results are computed with a gender classifier that differs from the one used for privacy enhancement in terms of training data and model topology. Results are reported (in terms of $\mu$ ± $\sigma$ computed over 4 test data splits) for gender suppression (SR), identity loss (IL) and the combined PIC score. The numbers show the relative change $\Delta$ of the mean score compared to the baseline performance, the arrows indicate whether the score increased (up) or decreased (down).

Even more impressive generalization was observed for the PrivacyNet model, which was trained in an adversarial setting and found to generalize best when utilized with unseen classifiers in the vanilla evaluation scenario.

2) Feature Distribution Exploration: Next, we investigate the effect of privacy enhancement on the distribution of the features generated by the last fully connected layers of the i) ResNet–50 face recognition model and the ii) VGG16 gender classifier. Here, the same setup is used as with the baseline experiments discussed above. The goal of this series of experiments is to better understand what is happening at the representation level as a result of the privacy enhancement and to gain additional insight into the characteristics of the privacy models. To study the ResNet–50 (identity) features, images corresponding to the 10 largest classes of LFW are selected and $t$-distributed Stochastic Neighbor Embedding ($t$–SNE) is used to visualize the feature distributions (in 2D) before and after privacy enhancement. For the gender features, 700 images from LFW are randomly sampled from the dataset for each gender.

The $t$–SNE plots in Fig. 9 show that both gender (top row) and subjects/identities (bottom row) are well separated with the original images. After privacy enhancement with $k$–AAP and FGSM most of the separation between subjects is preserved. The gender distributions, on the other hand, are less clustered and now exhibit a multimodal distribution. Nevertheless, the overlap between the male and female data points is still limited, indicating that the two classes can still be distinguished using a suitable classification model. These observations support the results from the previous section where only small values in IL scores were observed with these two techniques, while the ROC curves from the gender–recognition experiments showed good separation, but with inverted labels.

When looking at the FlowSAN models, a different behavior can be observed. Here, the identity features are not well separated, but for many data points exhibit (reasonably) correct pair–wise similarities. The gender distributions, on the other hand, overlap significantly, with a higher overlap for the FlowSAN–5 model, which is expected given the objective of the privacy enhancement. PrivacyNet appears to combine the best of both worlds and ensures reasonable preservation of the identity clusters, while leading to considerable overlap in the feature distributions w.r.t. gender. The observed distributions point to the shortcoming of the adversarial techniques $k$–AAP and FGSM, where the (gender) label swap could easily be identified using manual inspection of a few sample images. This is not the case with the FlowSAN and PrivacyNet models.

3) Qualitative Assessment: The impact of the privacy models on the visual appearance of a couple of sample images from the LFW dataset is shown in Fig. 10. As can be seen, $k$–AAP and FGSM generate privacy–enhanced images that are almost identical to the original images, the FlowSAN models introduce bigger and visually noticeable changes, whereas PrivacyNet processed images are somewhere in between. The bottom row in each of the two examples presents a visualization of the changes introduced by the privacy models. $k$–AAP and FGSM add a relatively uniformly distributed noise pattern to the input images, while the FlowSAN and PrivacyNet models introduce a structured pattern focused predominantly on the facial area and less so on the background. This observation can be attributed to the design of the privacy models. FlowSAN and PrivacyNet are designed specifically for facial images, while the other two models are applicable to arbitrary images and classification problems and, hence, are not explicitly targeting specific visual categories, i.e., faces. From a qualitative perspective, the adversarial models have an edge over the FlowSAN and PrivacyNet models, but as suggested in the previous sections this edge comes at the expense of robustness to unseen classification models and the simpler
privacy mechanism that aim to induce an incorrect prediction with a chosen classifier.

D. Robustness to Recovery Attempts

The main contribution of this study is a rigorous evaluation of existing (soft–biometric) privacy models with respect to their performance beyond zero-effort recognition experiments. The next series of experiments, therefore, explores the robustness of the evaluated models against attribute recovery attempts facilitated by our PrivacyProber.

1) One–Stage and Two–Stage Attribute Recovery: Multiple versions of PrivacyProber are implemented for the robustness experiments using either one–stage or two–stage attribute recovery. The simpler one–stage implementations consist of a single (generative or domain–specific) transformation, whereas the more complex two–stage implementations use sequential combinations of generative and domain–specific operations. For the one–stage implementations inpainting, image reconstruction, denoising and background removal are considered as stand–alone recovery options, whereas multiple different combination of image transformations are incorporated into the two–stage implementations. A high–level overview of the considered PrivacyProber variants is presented in Table III. Note that these combinations are not exhaustive with respect to all possible attribute recovery options discussed in Section III. However, they do provide a representative cross–section of the existing options for the experimental evaluation.

For the implementation of the recovery techniques state–of–the–art backbone models are selected. Specifically, the GMCNN model\textsuperscript{5} [56] trained on CelebA-HQ [57] is utilized for the proposed inpainting scheme, the auto-encoder from [48] trained on a selection of real-world images of objects is selected for the image reconstruction procedure, the WDnCNN\textsuperscript{6} [58] model trained on the Waterloo Exploration Database [59], the Berkeley segmentation dataset [60] and part of ImageNet [61] is chosen for the denoising operation and the DeepLabv3\textsuperscript{7} face parser [62] trained on CelebAMask-HQ [63] is selected for the background removal process. The backbone techniques ensure state–of–the–art performance for each task and come with publicly available implementations and pretrained weights. It is also important to note that the data used to train the backbone techniques does not overlap (in terms of images or subjects) with any of the test datasets used.

The impact of different PrivacyProbers on the visual appearance of a sample face image is presented in Fig. 11. Here, the numbers below the images correspond to probabilities that the subject on the image is male and the color coded frames indicate whether a gender classifier with a decision threshold at 0.5 correctly determines the subject’s gender (green) or not (red).

As can be seen, all tested privacy models significantly reduce the correct-gender-prediction probability compared to the original image, whereas the majority of PrivacyProbers successfully recover a considerable amount of gender information regardless of the privacy model used. The one–stage implementations based on inpainting (PP-I), image reconstruction (PP-A) and denoising (PP-D) also retain all of the visible semantic content. In terms of image quality, the inpainting scheme results in a slight loss of image contrast and the use of the auto-encoder-based leads to somewhat blurrier outputs. The denoising–based implementation, on the other hand, additionally improves the perceived quality of the facial images. Background subtraction is the only one–stage approach that removes part of the information and drastically changes the image content. Such images may be of limited use for certain applications, but can still facilitate

\textsuperscript{5}GMCNN: https://github.com/shepnerd/inpainting_gmcnn
\textsuperscript{6}WDnCNN: https://github.com/lixiaopeng123456/WDnCNN
\textsuperscript{7}DeepLabv3: https://github.com/tensorflow/models/tree/master/research/deeplab

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 11. Examples of PrivacyProber reconstructions, where the goal is to recover gender information from privacy-enhanced images. The number below each image corresponds to the probability output of a gender classifier – probabilities between 0–0.5 correspond to female and 0.5–1 to male subjects. Note that the privacy enhanced images, $I_{pr}$, generate incorrect gender probabilities (indicated by the red frames) or significantly reduce the correct-gender prediction probability. PrivacyProber, on the other hand, recovers the correct gender information in most cases – indicated by the green frames.

**TABLE IV**

| Privacy Model | Dataset | RR - GAUS models | RR - GUM models | ARR - GAUS models | ARR - GUM models |
|---------------|---------|-------------------|-----------------|-------------------|-----------------|
| k-AAP         | LPW     | 0.251 ± 0.076    | 0.169 ± 0.063   | 0.208 ± 0.017    | 0.410 ± 0.039   |
|               | MUCT    | 0.080 ± 0.043    | 0.078 ± 0.028   | 0.024 ± 0.013    | 0.042 ± 0.014   |
|               | Advers. | 0.024 ± 0.022    | 0.397 ± 0.097   | 0.358 ± 0.046    | 0.079 ± 0.018   |
| FGSN          | LPW     | 0.081 ± 0.023    | 0.060 ± 0.026   | 0.011 ± 0.007    | 0.353 ± 0.028   |
|               | MUCT    | 0.016 ± 0.016    | 0.014 ± 0.007   | 0.011 ± 0.009    | 0.162 ± 0.035   |
|               | Advers. | 0.379 ± 0.027    | 0.147 ± 0.003   | 0.192 ± 0.014    | 0.012 ± 0.007   |
| FlowSAN-3     | LPW     | 0.483 ± 0.030    | 0.425 ± 0.026   | 0.542 ± 0.028    | 0.430 ± 0.041   |
|               | MUCT    | 0.390 ± 0.084    | 0.324 ± 0.155   | 0.424 ± 0.130    | 0.270 ± 0.055   |
|               | Advers. | 0.554 ± 0.026    | 0.422 ± 0.036   | 0.490 ± 0.037    | 0.419 ± 0.069   |
| FlowSAN-5     | LPW     | 0.750 ± 0.057    | 0.748 ± 0.048   | 0.759 ± 0.049    | 0.754 ± 0.046   |
|               | MUCT    | 0.797 ± 0.055    | 0.730 ± 0.109   | 0.803 ± 0.087    | 0.723 ± 0.063   |
|               | Advers. | 0.750 ± 0.057    | 0.570 ± 0.054   | 0.665 ± 0.030    | 0.736 ± 0.061   |
| PrivacyNet    | LPW     | 0.652 ± 0.019    | 0.341 ± 0.097   | 0.520 ± 0.117    | 0.476 ± 0.065   |
|               | MUCT    | 0.016 ± 0.009    | 0.036 ± 0.014   | 0.032 ± 0.019    | 0.011 ± 0.009   |
|               | Advers. | 0.062 ± 0.051    | 0.572 ± 0.073   | 0.610 ± 0.075    | 0.579 ± 0.089   |
| Overall average |         | 0.480 ± 0.044    | 0.281 ± 0.065   | 0.428 ± 0.074    | 0.274 ± 0.060   |

Results are presented as mean±SD. ARR scores computed over 4 experimental splits and grouped by privacyprober type. GAUS = Gaussian and one-stage; GUM = Gaussian and two-stage; DGE = Gaussian and domain-specific and two-stage. The lowest robustness against attribute recovery for each privacy model on each dataset (in row) is marked red, the highest is marked blue. The results for the (overall) best-performing privacyprober variant for each group (GAUS, GUM, DGE) are shown in bold.

For the verification experiments, the ResNet–50 face recognition model is utilized.

1) **ARR analysis**: Table IV presents the attribute recovery robustness (ARR) scores (from (11)) generated from the attribute–recovered images. For each privacy model (and for each dataset), ARR scores for the PrivacyProber that resulted in the lowest robustness are colored red and the scores that correspond to the highest robustness are colored blue. Several interesting observations can be made from the presented results: i) First, considerably more information can be recovered from the adversarial techniques, k–AAP and FGSN on average, than from the FlowSAN and PrivacyNet models. For these techniques, at least one of the PrivacyProbers results in an ARR score in the 0.1 range, which suggests that a significant amount of the gender information can be restored. Among
the FlowSAN models, FlowSAN–3 is less robust to attribute recovery attempts than FlowSAN–5, but still significantly more so than k–AAP or FGSM. As expected, the most robust among all models is FlowSAN–5, where the weakest ARR scores across all PrivacyProbers are between 0.4 (on LFW) and 0.55 (on Adience). PrivacyNet is somewhere in between FlowSAN–3 and FlowSAN–5 with the weakest ARR score on LFW being 0.399 and on Adience 0.562. ii) Second, the robustness of the privacy models varies from dataset to dataset. While the results on LFW and MUCT are relatively consistent for most privacy models, the weakest ARR scores on Adience are higher than the weakest scores on LFW and MUCT. This observation implies that attribute recovery robustness is in parts also affected by the initial data characteristics, and not only by the recovery capabilities of the tested PrivacyProbers. Because the gender recognition performance was already weaker on Adience than on the remaining two datasets, even small degradations from the initial images, $I_{gr}$, cause further degradations. This fact is then also reflected in the success of the attribute recovery attempts. An exception to these observations is PrivacyNet, which performs well on LFW and Adience, but exhibits very low robustness on the MUCT datasets with several variants of PrivacyProber. The reason for this setting lies in the characteristics of the privacy model, which requires images aligned in a specific way. Since the faces in the MUCT dataset are aligned differently than those in LFW and Adience, the privacy enhancement with this data is less stable and, consequently, more susceptible to attribute-recovery attempts. iii) Third, all evaluated PrivacyProbers are able to recover some level of gender information from the privacy-enhanced images, suggesting that the reported performance with zero-effort evaluation scenarios typically reported in the literature often overestimates the actual capabilities of the existing privacy models.

In practice, only a subset of the PrivacyProbers, presented in this work, may be used for the robustness analysis (detailed in Algorithm 1) for efficiency reasons. A reasonable strategy in this case, is to select a single PrivacyProber variant from each of the three groups, i.e., GOS, GTS, and GDT, and utilize the selected models for the experiments. Such a selection ensures that different attribute-recovery strategies are represented and a sufficiently wide range of image-perturbation characteristics (that may be induced by the privacy models) is covered. From Table IV we observe that a selection of PP-I, PP-DI and PP-DB, which yield the lowest average ARR scores in each group, results in consistent findings with the ones made on the larger pool of PrivacyProbers and, therefore, represents a suitable subset of attribute recovery models for the robustness assessment.

2) Attribute recovery versus verification performance: The implemented PrivacyProbers aim to reconstruct attribute information and, as a result, alter the characteristics of the facial images. To analyze the impact of attribute recovery on the trade–off between identity preservation and attribute recovery robustness, we plot the difference in AUC scores between the original and the attribute–recovered images for both verification and gender recognition experiments in Fig. 12. Points indicating that all the identity and gender information contained in the original images was recovered are located at the origin of the plots. These points correspond to the least robust privacy models. Different from the scalar ARR scores analyzed above, the presented plots offer more insight into the behavior of the tested models, but also the characteristics of the attribute recovery attempts.

As can be seen from Fig. 12, the information content can be restored to close to the same level as before the application of $k$–AAP and FGSM on LFW and MUCT given the most effective PrivacyProbers. On these two datasets, FlowSAN–3–recovered images perform similarly to the original ones in terms of verification performance, but still offer a certain level of soft–biometric privacy. This result points to the robustness of the privacy model, but also shows that gender information can be recovered without affecting identity cues. FlowSAN–5 exhibits the greatest level of robustness to recovery attempts, but at the expense of the largest loss of identity information among all models. PrivacyNet preserves identity information very well on LFW and MUCT, but (across these two datasets) offers robustness to recovery attempts only on LFW. On MUCT, on the other hand, a significant amount of soft-biometric information can still be recovered. On Adience, similar observations as above can again be made for FGSM, that is, FGSM is again the least robust of all tested privacy models. $k$–AAP, FlowSAN–3,
We note again that the problem compares the k is comparably larger, but for most methods (except V and FGSM) PrivacyProber: IV-C2 training.

The feature representations generated by the recognition models, as illustrated in Fig. 13, the t-SNE based distributions produced by the face and gender-recognition models. Here, the same setup (involving LFW) as in Section IV-C2 is utilized to generate feature vectors for the plots. To keep the analysis concise, only the best performing one- and two-stage PrivacyProbers are considered, i.e., PP-D and PP-DI.

The presented plots support the observations made in the previous section. Attribute recovery (with PP-D and PP-DI) contributes towards well separated gender classes in the feature space for k–AAP and FGSM with minimal impact of the separability of the identity classes. For FlowSAN-3 and FlowSAN–5 we see less gender separation due to higher robustness of the privacy models and again minimal impact on identity information. A similar observation can also be made for PrivacyNet, where the gender overlap is reduced compared to the original privacy-enhanced images, while the identity separation is not degraded.

4) PrivacyProber and Adversarial Attacks: PrivacyProber is designed to recover attribute information, when this information is protected by soft-biometric privacy models. While such models often also incorporate adversarial perturbations, they are typically implemented with mechanisms that go beyond adversarial methods for two reasons. First, they strive towards a dual goal, i.e., i) to conceal information by fooling a classifier and ii) to preserve the utility of the data (e.g., verification accuracy, image quality etc.). Typical adversarial examples only try to achieve the first goal. Second, besides relying on adversarial noise, methods for soft–biometric privacy–enhancement also often include an image synthesis step, whereas classical adversarial methods do not.

Nevertheless, because adversarial methods and soft biometric privacy models share some commonalities, we explore in this section how PrivacyProber fares as a tool for adversarial defense. For the evaluation, we implement three adversarial attack methods, AdvDrop [64], Carlini-Wagner [34] and FGSM [35], and test them again on test images from LFW and with the two best-performing PrivacyProber variants, PP-D and PP-DI. Additionally, we compare PrivacyProber against a recent baseline adversarial defense mechanism, i.e., the auto-encoder (AE) defense from [48]. The adversarial attacks are designed to induce gender misclassifications and performance is measured in terms of AUC of the ROC curves, generated in gender recognition experiments.

We observe from the results in Table V that PrivacyProber: i) performs comparable to the state-of-the-art defense with the FGSM attack, ii) is slightly behind, but still competitive with the Carlini-Wagner attack, and iii) yields somewhat better performance with the most recent AdvDrop attack. The presented results suggest that despite the fact that PrivacyProber was primarily designed for robust evaluation of soft-biometric privacy models, it can also be used to explore the competitiveness of adversarial attacks.

E. Detecting (Soft–Biometric) Privacy Enhancement

In this section, we now show that PrivacyProber can also be used to efficiently detect image tampering with soft-biometric privacy models. To this end, we implement the proposed APEND detector using three diverse PrivacyProber variants, i.e., PP-A, PP-DI and PP-B. We note that other combinations of PrivacyProber could be used for the experiments with different performance. However, our goal here is to provide a proof-of-concept for the detector and illustrate the benefits of attribute recovery, and not to optimize performance indicators.

1) Quantitative Evaluation: We note again that the problem of detecting privacy enhancement in facial images has, to the
Table V

| Attack model | AUC after attack | AUC after defense |
|--------------|-----------------|-------------------|
| FCSM [35]    | 0.983 ± 0.003   | 0.975 ± 0.012     |
| Carlini-Wagner [34] | 0.962 ± 0.017 | 0.848 ± 0.062     |
| AdvDrop [54]  | 0.600 ± 0.017   | 0.659 ± 0.023     |

While privacyprober is designed as an attribute recovery approach for soft-biometric privacy models, it also ensures competitive performance as an adversarial defense.

Table VI

| Privacy Model | Dataset | APEND (ours) | PREM [28] | T-SVM (A) [44] | T-SVM (B) [44] |
|---------------|---------|-------------|-----------|----------------|---------------|
| k–AAP         | LFW     | 0.982 ± 0.002 | 0.957 ± 0.005 | 0.743 ± 0.013 | 0.984 ± 0.007 |
|               | MUCT    | 0.984 ± 0.002 | 0.965 ± 0.005 | 0.727 ± 0.030 | 0.534 ± 0.021 |
|               | Adience | 0.820 ± 0.004 | 0.629 ± 0.007 | 0.604 ± 0.008 | 0.541 ± 0.007 |
| FCSM          | LFW     | 0.993 ± 0.002 | 0.989 ± 0.003 | 0.894 ± 0.013 | 0.921 ± 0.002 |
|               | MUCT    | 0.995 ± 0.002 | 0.989 ± 0.003 | 0.858 ± 0.015 | 0.552 ± 0.002 |
|               | Adience | 0.907 ± 0.001 | 0.858 ± 0.003 | 0.595 ± 0.016 | 0.471 ± 0.004 |
| FlowSAN–3     | LFW     | 0.984 ± 0.002 | 0.987 ± 0.002 | 1.000 ± 0.000 | 0.997 ± 0.002 |
|               | MUCT    | 0.991 ± 0.002 | 0.993 ± 0.001 | 1.000 ± 0.000 | 0.524 ± 0.020 |
|               | Adience | 0.754 ± 0.003 | 0.775 ± 0.002 | 1.000 ± 0.000 | 0.609 ± 0.005 |
| FlowSAN–5     | LFW     | 0.981 ± 0.002 | 0.988 ± 0.001 | 1.000 ± 0.000 | 0.998 ± 0.001 |
|               | MUCT    | 0.991 ± 0.003 | 0.994 ± 0.001 | 1.000 ± 0.000 | 0.531 ± 0.018 |
|               | Adience | 0.770 ± 0.006 | 0.793 ± 0.011 | 1.000 ± 0.000 | 0.563 ± 0.005 |
| PrivacyNet    | LFW     | 0.989 ± 0.002 | 0.978 ± 0.002 | 0.533 ± 0.017 | 0.765 ± 0.004 |
|               | MUCT    | 0.995 ± 0.002 | 0.995 ± 0.002 | 0.501 ± 0.029 | 0.205 ± 0.020 |
|               | Adience | 0.931 ± 0.002 | 0.948 ± 0.002 | 0.439 ± 0.014 | 0.254 ± 0.006 |
| Average       | 0.940 ± 0.084 | 0.926 ± 0.108 | 0.792 ± 0.009 | 0.630 ± 0.008 |

1 Detection model trained on FlowSAN–5 and LFW.
2 Detection model trained on k–AAP and LFW.

The learning-free (black box) APEND approach is compared against the classification–based T-SVM techniques for adversarial attack detection from [44] as well as the super-resolution based PREM detector from [28].

Best of our knowledge, not yet been studied widely in the open literature. To the best of our knowledge, only the PREM approach from [28] aims to address the same task and is therefore also included in the evaluation. However, because soft-biometric privacy enhancement techniques share some characteristics with adversarial attacks, we also select the recent state-of-the-art transformation-based detection technique (denoted T-SVM hereafter) from [44] as a baseline for our experiments and compare it to APEND. T-SVM combines features from the discrete wavelet (DWT) and discrete sine (DST) transforms with a support vector classifier (SVM) for tampering detection and requires training data to be able to learn how to discriminate between original and tampered images. We, therefore, consider the following setting for the experimental evaluation to ensure a fair comparison, i.e.:

- **APEND and PREM**: These methods require no examples of tampered images or knowledge of the mechanism used for privacy enhancement and are, therefore, tested in a purely black-box scenario.
- **T-SVM**: Because T-SVM is supervised and does require examples of tampered images or training, we design two experimental configurations that test for the generalization capabilities of this detector. With the first configuration, T-SVM (A), is trained for the detection of FlowSAN–5 enhanced images only on data from LFW. For the second configuration, T-SVM (B), is trained for detecting k–AAP based enhancement, again only on LFW. The detection models from both configurations are then tested on all datasets and with all privacy models, thus, simulating a black-box evaluation scenario.

Table VI shows that in the supervised approach, T–SVM, is the most competitive on the datasets and privacy model it was trained on, while the performance with unseen models (and in most cases also with unseen datasets) quickly deteriorates. The APEND and PREM techniques, on the other hand, are training-free and, therefore, generalize better to unseen models and across different datasets. While both PREM and APEND are quite competitive in most experiments, the aggregation of reconstruction evidence integrated in APEND leads to the best overall average performance of 0.940 in terms of AUC, compared to 0.926 for PREM. The added robustness through the aggregation process is especially evident in specific cases, such as for example with the k–AAP technique on the Adience dataset, where APEND outperforms PREM by more than 23% in terms of AUC. The presented results clearly show the added value of training-free tampering detection as well as the aggregation-based approach to privacy-enhancement detection, which lead to highly encouraging results in the presented experiments.

2) **Visual Analysis**: Fig. 14 shows a few example face images, where the proposed APEND detector incorrectly flagged the presence of image tampering (i.e., privacy enhancement) at a decision threshold that ensures equal error (EER) rates in the (two-class) gender recognition experiments. In the top row, we show images, where APEND failed to detect privacy-enhanced
images. In most cases, these images contain strong image artifacts that make it challenging to properly recover attribute information using the PrivacyProber variant used for the implementation of APEND. This leads to minute differences in the gender predictions between the privacy-enhanced and recovered images, and eventually to misdetections. In the bottom row, we show example images that have been flagged by APEND as being tampered, but in fact represent original images. As can be seen, such images are often of poor quality (due to blur, noise, etc.) and get improved by the recovery approaches in APEND. Thus, the predictions of the gender classifier change sufficiently to flag the images as tampered. While the performance of APEND is highly competitive (and close to ideal on most datasets), the presented examples suggest that there is room for further improvement by, for example, combining APEND with supervised detection techniques that should be able to also perform well with poor quality images.

V. CONCLUSION

In this study, we investigated the reliability of privacy-enhancing techniques for soft biometrics and explored their resistance to attribute recovery attempts. Our framework, PrivacyProber, was utilized to recover suppressed soft biometric information from facial images, and was evaluated in comprehensive experiments on the LFW, MUCT, and Adience datasets with three soft-biometric privacy models. The experimental results revealed significant variations among the tested privacy models in terms of visual impact on the privacy-enhanced images and the level of privacy ensured. Notably, our findings indicated that a considerable amount of suppressed attribute information could be recovered regardless of the privacy model used, with synthesis-based techniques such as FlowSANs and PrivacyNet demonstrating greater robustness compared to adversarial approaches like k-AAP and FGSM. These observations highlight the need for further research to enhance the robustness of existing models in the field of biometric privacy.

Additionally, we introduced the APEND detector, which showcased high accuracy in detecting privacy enhancements in facial images, even when different privacy models were employed and diverse data characteristics were present. This fact points to another threat vector with respect to biometric privacy models, in that privacy–enhanced images can easily be identified and subjected to alternative means of processing that is less sensitive to artifacts and perturbations infused with the enhancement. As part of our future work, we plan to extend our analysis to other soft-biometric attributes, such as age and ethnicity, within the same framework employed for evaluating the gender attribute in this study. Furthermore, we aim to explore the robustness of privacy-enhancing techniques across multiple attributes, leveraging the reconstructions from PrivacyProber and employing multiple classifiers to assess the privacy models’ resilience along various dimensions.

REFERENCES

[1] A. Dantcheva, P. Elia, and A. Ross, “What else does your biometric data reveal? a survey on soft biometrics,” IEEE Trans. Inf. Forensics Secur., vol. 11, no. 3, pp. 441–467, Mar. 2016.
[2] H. Pan, H. Han, S. Shan, and X. Chen, “Mean-variance loss for deep age estimation from a face,” in Proc. Comput. Vis. Pattern Recognit., 2018, pp. 5285–5294.
[3] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, “Ordinal regression with multiple output CNNs for age estimation,” in Proc. Comput. Vis. Pattern Recognit., 2016, pp. 4920–4928.
[4] N. Berthouze et al., “EMOPAIN challenge 2020: Multimodal pain evaluation from facial and bodily expressions,” in Proc. IEEE 15th Int. Conf. Autom. Face Gesture Recognit., 2020, pp. 849–856.
[5] P. Li, L. Prieto, D. Mery, and P. Flynn, “Face recognition in low quality images: A survey,” 2018, arXiv:1805.11519.
[6] L. Mao, Y. Yan, J.-H. Xue, and H. Wang, “Deep multi-task multi-label CNN for effective facial attribute classification,” IEEE Trans. Affect. Comput., vol. 13, no. 2, pp. 818–828, Apr-Jun. 2022.
[7] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” Nature Mach. Intell., vol. 1, no. 9, pp. 389–399, 2019.
[8] R. Singh, M. Vatsa, and N. Ratha, “Trustworthy AI,” in Proc. 8th Int. Conf. Adv. Comput. Commun. Syst., 2021, pp. 449–453.
[9] V. Mirjalili, S. Raschka, and A. Ross, “FlowSAN: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers,” IEEE Access, vol. 7, pp. 99735–9974, 2019.
[10] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, “Anonymizing k-facial attributes via adversarial perturbations,” in Proc. Int. Joint Conf. Artif. Intell., 2018, pp. 656–662.
[11] B. Meden et al., “Privacy-enhancing face biometrics: A comprehensive survey,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 417–4183, 2021.
[12] V. Mirjalili, S. Raschka, and A. Ross, “PrivacyNet: Semi-adversarial networks for multi-attribute face privacy,” IEEE Trans. Image Process., vol. 29, pp. 9400–9412, 2020.
[13] P. Terhiö et al., “PE-MIU: A training-free privacy-enhancing face recognition approach based on minimum information units,” IEEE Access, vol. 20, pp. 93635–93647, 2020.
[14] B. Bortolato et al., “Learning privacy-enhancing face representations through feature disentanglement,” in Proc. Autom. Face Gesture Recognit., 2020, pp. 495–502.
[15] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, “Fawkes: Protecting personal privacy against unauthorized deep learning models,” in Proc. 29th USENIX Secur. Symp., 2020, pp. 1589–1604.
[16] B. Meden, R. C. Matti, S. Fabijan, H. K. Ekenel, V. Struc, and P. Peer, “Face deidentification with generative deep neural networks,” IET Signal Process., vol. 11, no. 9, pp. 1046–1054, 2017.
[17] P. Voigt and A. V. D. Bussche, The EU General Data Protection Regulation (GDPR): A Practical Guide, 1st ed., Berlin, Germany: Springer, 2017.
[18] P. Bukaty, The California Consumer Privacy Act (CCPA): An Implementation Guide, Cambridgeshire Business Park, U.K.: IT Governance Publishing, 2019.
[19] Assembly, Illinois General, “740 ILCS/14/Biometric information privacy act,” 2008.
[20] P. Melzi, C. Rathgeb, R. Tolosa, R. Vera-Rodriguez, and C. Busch, “An overview of privacy-enhancing technologies in biometric recognition,” 2022, arXiv:2206.10465.
[21] A. Othman and A. Ross, “Privacy of facial soft biometrics: Suppressing gender but retaining identity,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 682–696.
[22] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, “Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images,” in Proc. Int. Conf. Biometrics, 2018, pp. 82–89.

[23] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boult, “Facial attributes: Accuracy and adversarial robustness,” Pattern Recognit. Lett., vol. 124, pp. 100–108, 2019.

[24] T. Winkler and B. Rinner, “Security and privacy protection in visual sensor networks: A survey,” ACM Comput. Surveys, vol. 47, pp. 1–42, 2014.

[25] J. R. Padilla-López, A. A. Charaouri, and F. Flórez-Revuelta, “Visual privacy protection methods: A survey,” Expert Syst. Appl., vol. 42, no. 9, pp. 4177–4195, 2015.

[26] S. Garfinkel, De-Identification of Personal Information. Gaithersburg, MD, USA: NIST Interagency/Intern. Rep. (NISTIR), Nat. Inst. Standards Technol., 2015.

[27] S. Ribarič, A. Ariyaeinia, and N. Pavesi, “De-identification for privacy protection in multimedia content: A survey,” Signal Process. Image Comm., vol. 47, pp. 131–151, 2016.

[28] P. Rot, P. Pez, and V. Štuc, “Detecting soft-biometric privacy enhancement,” in Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch Eds., Berlin, Germany: Springer, 2022, pp. 391–411.

[29] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuiper, “Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations,” Appl. Intell., vol. 49, pp. 3043–3060, 2019.

[30] P. Terhörst et al., “Privacy evaluation protocols for the evaluation of soft-biometric privacy-enhancing technologies,” in Proc. Int. Conf. Biometrics Special Int. Group, 2020, pp. 215–222.

[31] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, “SensitiveNets: Learning agnostic representations with application to face images,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 6, pp. 2158–2164, Jun. 2021.

[32] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuiper, “Suppressing gender and age in face templates using incremental variable elimination,” in Proc. Int. Conf. Biometrics, 2019, pp. 1–8.

[33] V. Mirjalili and A. Ross, “Soft biometric privacy: Retaining biometric utility of face images while perturbing gender,” in Proc. Int. Joint Conf. Biometrics, 2017, pp. 564–573.

[34] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in Proc. IEEE Symp. Secur. Privacy, 2017, pp. 39–57.

[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proc. Int. Conf. Learn. Representations, 2014, arXiv:1412.6572.

[36] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa, “How are attributes expressed in face DCNNs?” in Proc. IEEE 15th Int. Conf. Autom. Face Recognit. Gender Recognit., 2020, pp. 85–92.

[37] L. Zheng, Y. Zhang, and V. L. Thing, “A survey on image tampering and its detection in real-world photos,” J. Vis. Commun. Image Represent., vol. 58, pp. 380–399, 2019.

[38] K. A. da Costa et al., “A critical literature survey and prospects on tampering and anomaly detection in image data,” Appl. Soft Comput., vol. 97, 2020, Art. no. 106727.

[39] K. B. Meena and V. Tyagi, “Image forgery detection: Survey and future directions,” in Data, Engineering and Applications, Berlin, Germany: Springer, 2019.

[40] E. Nowroozi, A. Dehghantanha, R. M. Parizi, and K.-K. R. Choo, “A survey of machine learning techniques in adversarial image forensics,” Comput. Secur., vol. 100, 2020, Art. no. 102092.

[41] S. Bulusu, B. Kallikhurst, B. Li, P. Varshney, and D. Song, “Anomalous example detection in deep learning: A survey,” IEEE Access, vol. 8, pp. 132330–132347, 2020.

[42] X. Wang, J. Li, X. Kuang, Y.-A. Tan, and J. Li, “The security of machine learning in an adversarial setting: A survey,” J. Parallel Distrib. Comput., vol. 130, pp. 12–23, 2019.

[43] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” 2017, arXiv:1710.10196.

[44] J. Vig, K. Zhang, and J. Kautz, “Defending adversarial image classification,” in Proc. IEEE Int. Conf. Comput. Vision, 2019, pp. 5549–5558.

[45] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, “AdvDrop: Adversarial attack to DNNs by dropping information,” in Proc. IEEE/CVF Int. Conf. Comput. Vision, 2021, pp. 7486–7495.
Klemen Grm (Member, IEEE) received his PhD degree from the University of Ljubljana, Slovenia, in 2020. His PhD work was also awarded the 2019 Max Snijder award by the European Association for Biometrics. He is a research assistant with the Laboratory for Machine Intelligence at the Faculty of Electrical Engineering, University of Ljubljana. His research interests include image processing, machine learning, and face biometrics.

Peter Peer (Senior Member, IEEE) received the doctoral degree in computer science from the University of Ljubljana, in 2003. He is a professor of computer science with the University of Ljubljana, Slovenia. He participated in several national and EU funded R&D projects and published more than 100 research papers in leading international peer reviewed journals and conferences. He serves as an associated editor of IEEE Access and IET Biometrics. He is a member of the EAB, IAPR.

Vitomir Štruc (Senior Member, IEEE) is a professor with the University of Ljubljana, Slovenia. His research interests include biometrics, computer vision, image processing, pattern recognition and machine learning. He is a senior area editor for IEEE TIFS, a Subject Editor for Signal Processing and an associate editor for Pattern Recognition, and IET Biometrics. He is a member IAPR, EURASIP, Slovenia’s national contact point for the EAB and the former president of the Slovenian Pattern Recognition Society.