IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers

VAHID MIRJALILI[1], SEBASTIAN RASCHKA[2], ARUN ROSS[3]
[1]Computer Science and Engineering, Michigan State University, East Lansing, MI 48824 (e-mail: contact@vahidmirjalili.com)
[2]Department of Statistics, Univeristy of Wisconsin – Madison, Madison, WI 53706 (e-mail: sraschka@wisc.edu)
[3]Computer Science and Engineering, Michigan State University, East Lansing, MI 48824

Corresponding author: Arun Ross (e-mail: rossarun@cse.msu.edu).

This work was supported in part by the National Science Foundation under Grant 1618518.

ABSTRACT Privacy concerns in the modern digital age have prompted researchers to develop techniques that allow users to selectively suppress certain information in collected data while allowing for other information to be extracted. In this regard, Semi-Adversarial Networks (SAN) have recently emerged as a method for imparting soft-biometric privacy to face images. SAN enables modifications of input face images so that the resulting face images can still be reliably used by arbitrary conventional face matchers for recognition purposes, while attribute classifiers, such as gender classifiers, are confounded. However, the generalizability of SANs across *arbitrary* gender classifiers has remained an open concern. In this work, we propose a new method, FlowSAN, for allowing SANs to generalize to multiple gender classifiers. We propose stacking a diverse set of SAN models to compensate each other's weaknesses, thereby, forming a robust model with improved generalization capability. Extensive experiments using different unseen gender classifiers and face matchers demonstrate the efficacy of the proposed paradigm in imparting gender privacy to face images.

INDEX TERMS Biometrics, Face Image, Semi-Adversarial Networks, SAN, Gender, Privacy, Adversarial, Deep Learning.

I. INTRODUCTION

FACE images of individuals contain valuable information unique to themselves that facilitates biometric face recognition. Face recognition involves comparing features extracted from a pair of face images, using a *face matcher*, to determine their degree of similarity or dissimilarity [1], [2]. In addition, other auxiliary information such as age, gender, and race, which are called soft-biometrics, can also be extracted from face images using machine learning techniques [1], [3], [4]. The increasing use of face recognition in various applications has brought the issue of data privacy to the forefront [5]–[18]. While extracting soft-biometric information can be useful in many applications [19], we should note that such information can be abused in several ways, such as profiling users, targeted advertisement, and increasing the risk of linkage attacks [20]. Furthermore, extracting this information without the users' consent may be viewed as a violation of their privacy. One aspect of privacy involves granting users the right to determine which personal information to reveal and which to conceal [21], [22]. In this regard, *soft-biometric privacy* was introduced as a means for preserving the biometric utility of face images, while confounding soft-biometric information, such as gender characteristics [23], [24].

Recently, European Union's General Data Protection Regulation (GDPR) [25] has come to effect. One of its goals is to protect the data collected from European users and to regulate its usage. To this effect, it enforces any entity (individual or group) collecting data from European users to disclose the type-of-data collected, the intended usage, and the data-processing techniques that will be used. Accordingly, GDPR prohibits any processing of individuals' information beyond the stated purpose at the time of data collection. For example, consider a scenario where users of a biometric application or service can optionally withhold their gender information; however, such information could still be extracted automati-

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

cally from their biometric data [26]–[34].

In the context of GDPR, biometric data of individuals, such as face photos or fingerprints, are collected solely for the purpose of user recognition, without acquiring other demographic information such as age, gender, and ethnicity. In such a scenario, applying data processing techniques that allow extracting such sensitive information automatically from a person's biometric data [1], [3], [32], [35]–[40] without their knowledge and consent can be a violation of the users' privacy. While GDPR prohibits unsolicited data extraction from European users, the possibility of unlawful data collection still remains and can ultimately lead to negative societal, economic, and political consequences [41]–[43].

Previously, we developed Semi-Adversarial Networks (SAN) [44] for imparting soft-biometric privacy to face images, where a face image is modified such that the matching utility of the modified face image is retained while the automatic extraction of gender information is confounded. In our previous work [44], we empirically showed that the ability to predict gender information, using an unseen gender classifier from outputs of the SAN model, is successfully diminished. In [45], we defined the generalizability of the SAN model as its ability to confound arbitrary unseen[1] gender classifiers. Generalizability is an important property for real-world privacy applications since the lack thereof implies that there exists at least one gender classifier that can still reliably estimate the gender attribute from outputs of the SAN model and, therefore, jeopardizes the privacy of users. In order to address the generalizability issue of SAN models, in this paper, we propose the FlowSAN model, that progressively degrades the performance of unseen gender classifiers. Extensive experiments on a variety of independent gender classifiers and face image datasets show that the proposed FlowSAN method (Fig. 1) results in a substantially improved generalization performance compared to the original SAN method with regard to concealing gender information while retaining face matching utility.

II. RELATED WORK

With regard to privacy concerns in recent years, a new line of research has emerged that focuses on methods for imparting soft-biometric privacy to biometric data and face images in particular [8]–[10], [23], [24], [46]. Othman and Ross [23] first proposed an approach for mixing input face images with candidate images of the opposite gender using Active Shape Model [47]. Subsequently, Mirjalili and Ross [24] developed a scheme that modifies an input face image using adversarial perturbations [48] where the performance of a given gender classifier was confounded while the performance of a face matcher was retained. Chhabra et al. [9] later extended this research by including multiple attribute classifiers. They applied additive perturbations to face images to either preserve or suppress certain soft-biometric attributes [9]. While these proposed schemes successfully confound a target attribute classifier, they fail to generalize to unseen attribute classifiers. Thus, soft-biometric attributes remain susceptible to extraction by unseen classifiers.

In order to derive perturbations that are transferable to unseen gender classifiers, Mirjalili et al. [44] designed a convolutional autoencoder that modifies input face images such that an auxiliary face matcher still retains good matching performance on the modified output image while confounding an auxiliary gender classifier. As a result, since the output of their model is adversarial to one classifier and not to the other, the architecture is referred to as Semi-Adversarial Networks (SAN). The SAN model was shown to be able to derive perturbations that are transferable to two unseen gender classifiers. In [45], we investigated the generalizability of SAN models across multiple arbitrary gender classifiers and formulated an ensemble SAN model with a training scheme based on different data augmentation techniques, to enhance diversity in the ensemble of SAN models. Furthermore, we explored the effectiveness of randomly selecting a perturbed image from an ensemble of SAN models, which we refer to as Ens-Gibbs [45].

While these methods directly apply perturbations to face images, recently, new techniques have emerged where perturbations were applied to face *representation* vectors computed by face matchers [8], [13]. In particular, Morales et al. [8] proposed a neural-network-based model, called SensitiveNet, that is able to remove soft-biometric information from face representation vectors. Therefore, any attribute classifier trained on face representation vectors may not be able to extract such sensitive information. However, these methods are based on the assumption that only face *representation* vectors are stored in a biometric database. This scheme is not desirable in many applications since only storing face representations results in 1) the loss of human interpretability, and 2) the potential lack of compatibility with newer face matchers in the future. Hence, it is desirable to modify the gender information in the original face images directly, which is also the goal of the method presented in this paper, as opposed to modifying face representation vectors. An overview of existing techniques and their properties (transferability, generalization to arbitrary attribute classifiers, and retaining matching utility) is shown in Table 1.

In [45], we investigated how well the SAN model generalizes to multiple unseen gender classifiers and unseen face matchers. To improve the generalizability of the SAN model, we proposed an ensemble scheme based on multiple SAN models trained on different training subsets. However, we observed that even though the generalizability could be improved, the SAN model was still not able to generalize well to all unseen face matchers and gender classifiers tested in this study. In this work, we address the generalization issue of the SAN method using a novel stacking paradigm that will successively enhance the perturbations for confounding an arbitrary unseen gender classifier as illustrated in Fig. 1. We

[1] The term "unseen" indicates that a certain classifier (or face matcher) was not used during the training stage. On the contrary, the term "auxiliary" in this paper refers to the classifier (or face matcher) that is either used or developed during the training phase.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

TABLE 1: Overview of existing methods for imparting soft-biometric privacy and their comparison based on three criteria: transferability, generalizability, and retention of matching performance; transferability refers to the ability to generate perturbations that can successfully confound a different gender classifier, whereas generalizability is a stronger criterion for the ability to confound *any* arbitrary unseen gender classifier.

Authors	Domain	Proposed Method	Transferable	Generalizable	Matching Performance
Othman and Ross [23]	Face images	Mixing faces of opposite gender	Yes	Yes	Severely degraded
Sim and Li [10]	Face images	Multimodal Discriminant Analysis	Yes	Yes	Severely degraded
Mirjalili et al. [24]	Face images	Adversarial perturbations	No	No	Mostly retained
Mirjalili et al. [44]	Face images	Semi-Adversarial Networks	Yes	No	Mostly retained
Chhabra et al. [9]	Face images	Adversarial perturbations	No	No	Mostly retained
Mirjalili et al. [45]	Face images	Ensemble of SAN models	Yes	Yes	Mostly retained
Morales et al. [8]	Face representations	SensitiveNet	Yes	Yes	Mostly retained
Terhörst et al. [13]	Face representations	Noise transformation	Yes	Yes	Mostly retained

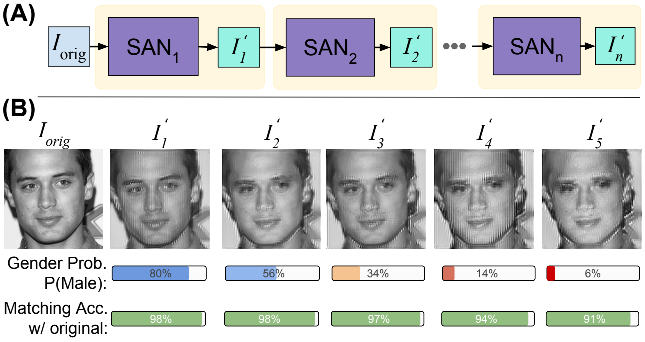

FIGURE 1: Illustration of the FlowSAN model, which sequentially combines individual SAN models in order to sequentially perturb a previously unseen gender classifier, while the performance of an unseen face matcher is preserved. A: An input gray-scale face image I_{orig} is passed to the first SAN model (SAN$_1$) in the ensemble. The output image of SAN$_1$, I'_1, is then passed to the second SAN model in the ensemble, SAN$_2$, and so forth. B: An unmodified face image from the CelebA [49] dataset (I_{orig}) and the perturbed variants I'_i after passing it through the different SAN models sequentially. The gender prediction results measured as probability of being male ($P(\text{Male})$) as well as the face match score between the original (I_{orig}) and the perturbed images (I'_i) are shown.

refer to this method as FlowSAN. The primary contributions of this work are as follows:

- Designing the FlowSAN model that can successively degrade the performance of arbitrary unseen gender classifiers;
- Generalizing the FlowSAN model to multiple arbitrary gender classifiers;
- Demonstrating the practicality and efficacy of the proposed approach in confounding the gender information for real-world privacy applications via extensive experiments involving broad and diverse sets of datasets.

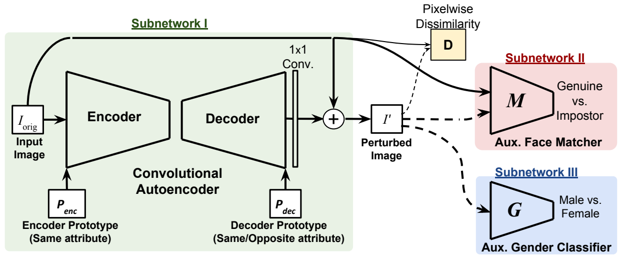

FIGURE 2: Architecture of the original SAN model [44] composed of three subnetworks: I: a convolutional autoencoder [50], II: an auxiliary face matcher (M), and III: an auxiliary gender classifier (G). In addition, the unit D computes the pixelwise dissimilarity between input and perturbed images during model training.

III. PROPOSED METHOD

Original SAN model [44]: The SAN model for imparting gender privacy to face images was first proposed in [44], and the overall architecture is shown in Fig. 2. The SAN model leverages pre-computed **face prototypes**, which are average face images for each gender. SAN consists of three subnetworks: 1) a **convolutional autoencoder** that perturbs an input face image via face prototypes, 2) an **auxiliary face matcher**, which is a convolutional neural network (CNN), and 3) a CNN-based **auxiliary gender classifier**. The input to the convolutional autoencoder is a gray-scale[2] face image I_{orig}, of size $224{\times}224{\times}1$, fused with a face prototype belonging to the same gender (P_{sm}). After the fused input image was passed through the encoder and decoder networks, the face prototypes (P_{sm} prototype face image from the same gender as input image, or P_{op} the prototype face image of the opposite gender) are added as additional channels to the resulting 128-channel feature-map representation. Finally, a $1{\times}1$-convolutional operation is used to reduce the number of channels in the resulting feature-maps to a $224{\times}224{\times}1$-dimensional output image, which is denoted as I'_{sm} or I'_{op}, depending on the type of prototype used by the decoder:

$$\begin{aligned} I'_{\text{sm}} &= \text{SAN}(I_{\text{orig}}; P_{\text{sm}}), \text{ and} \\ I'_{\text{op}} &= \text{SAN}(I_{\text{orig}}; P_{\text{op}}). \end{aligned} \tag{1}$$

[2] Since most face matchers work with gray-scale face images, we used gray-scale images in all experiments to allow for a fair comparison between matchers based on the same input data.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

These output images, I'_{sm} and I'_{op}, are then passed to both the auxiliary face matcher and the auxiliary gender classifier. The auxiliary face matcher predicts whether the original and the perturbed face images belong to the same individual via a face match score. The gender classifier predicts the gender of the input and output images via gender probabilities for male and female.[3] For the auxiliary face matcher, the pre-trained, publicly available VGG-face model [51] is used, which computes the face representation vectors for an input face image, and the similarity between two face representation vectors determines the associated match-score.

Three different loss functions are defined based on the outputs from the autoencoder, the auxiliary gender classifier, and the auxiliary face matcher. The first component of the loss function, $\mathcal{J}_D$, measures the pixelwise dissimilarity between the input and the output from the same-gender prototype I'_{sm}, which is used to ensure that the autoencoder subnetwork is able to construct realistic face images:

$$\mathcal{J}_D(I_{\text{orig}}, I'_{\text{sm}}) = \frac{1}{h \times w} \sum_{i=1}^{h \times w} \mathcal{H}(I^{(i)}_{\text{orig}}, I'^{(i)}_{\text{sm}}), \tag{2}$$

where $\mathcal{H}$ indicates the cross-entropy function for the binary case, defined as

$$\mathcal{H}(p, q) = -\left(p \log(q) + (1-p)\log(1-q)\right). \tag{3}$$

The second loss term, $\mathcal{J}_M$, is the squared L^2 distance between the face representation vectors obtained from the auxiliary face matcher (VGG-face network [51]) for the input image and the perturbed output, making the autoencoder learn how to perturb face images such that the accuracy of the face matcher is retained:

$$\mathcal{J}_M(I_{\text{orig}}, I'_{\text{op}}) = \|\mathcal{R}_M(I_{\text{orig}}) - \mathcal{R}_M(I'_{\text{op}})\|_2^2, \tag{4}$$

where $\mathcal{R}_M(I)$ and $\mathcal{R}_M(I'_{\text{op}})$ indicate the face representation vectors for the input image and the perturbed output based on the opposite-gender prototype.

Finally, the third loss term, $\mathcal{J}_G$, is the cross-entropy loss function applied to the gender probabilities computed by the auxiliary gender classifier, G, on the two perturbed output images. Here, the ground-truth label y of the input image is used for I'_{sm}, but the reverse $(1-y)$ is used for I'_{op}:

$$\mathcal{J}_G(y, I'_{\text{sm}}, I'_{\text{op}}) = \mathcal{H}(y, G(I'^{(k)}_{\text{sm}})) + \mathcal{H}(1-y, G(I'^{(k)}_{\text{op}})). \tag{5}$$

The total loss, $\mathcal{J}_{tot}$, is the weighted sum of the three individual loss functions described in the previous paragraphs,

$$\mathcal{J}_{tot} = \lambda_1 \mathcal{J}_D + \lambda_2 \mathcal{J}_M + \lambda_3 \mathcal{J}_G, \tag{6}$$

where the parameters λ_i are the relative weighting terms that can be chosen uniformly or adjusted via hyperparameter optimization.

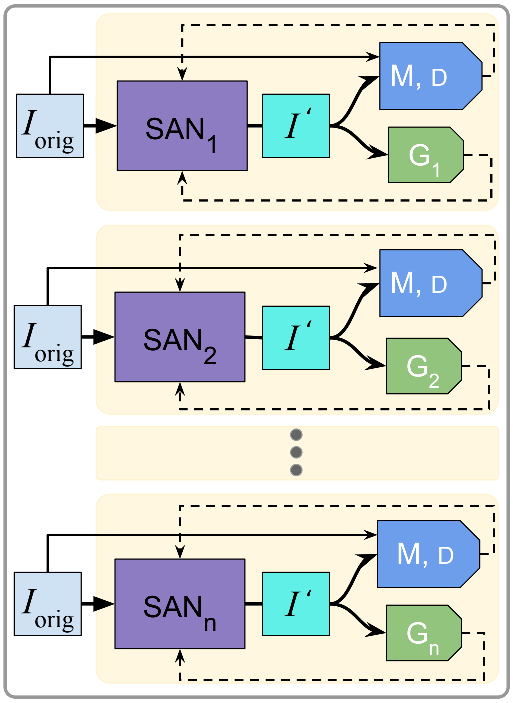

FIGURE 3: Illustration of an ensemble SAN, where individual SAN models are trained *independent* of each other using n diverse, pre-trained, auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, ..., G_n\}$), and a face matcher M that computes face representation vectors for both input face image I_{orig} and the output of the SAN model. D refers to a module that computes pixelwise dissimilarity between an input and output face image.

In the remaining part of the paper, we use notation I' for the output of a SAN model on a face image I_{orig} when using the opposite-gender prototype, i.e., $I' = \text{SAN}(I_{\text{orig}}; P_{\text{op}})$.

Based on our previous study [45], we employed a data augmentation and resampling scheme for training the auxiliary gender classifiers as a means to diversify the SAN models. In particular, by resampling the instances belonging to the underrepresented race in the CelebA [49] dataset, we aimed to balance the racial distribution in the training data. In this regard, we generated five resampled training datasets, where in each one a random disjoint subset of samples from the underrepresented race was replicated 40 times. This is an effort to enhance the diversity among the SAN models in an ensemble. The resampling approaches that are used to mitigate the imbalances in the different training datasets employed in this study are described in [45].

A. TRAINING AND EVALUATION OF AN ENSEMBLE SAN MODEL

In our previous work [45], we proposed an ensemble approach for generalizing SAN models to unseen gender classifiers. The objective of an ensemble SAN was to create n SAN models such that their union can span a larger subset

[3] In this paper, we have assumed binary labels for gender; however, it must be noted that societal and personal interpretation of gender will result in many more classes.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

of the hypothesis space compared to a single SAN model. Therefore, for a new test image and an arbitrary unseen gender classifier, G, it is likely that at least one of these SAN models in the ensemble is able to confound G. For training an ensemble of SANs, we start with n auxiliary gender classifiers, $\mathcal{G} = \{G_1, G_2, ..., G_n\}$, which were trained using different data augmentation schemes (to achieve higher diversity among classifiers), and a pre-trained face matcher M. Then, we train n SAN models, where SAN$_i$ is associated with the auxiliary gender classifier G_i, as shown in Fig. 3. According to the original SAN model proposed in [44], the loss function for training each model is composed of three components: gender loss, matching loss, and pixelwise dissimilarity loss (Eq. 6). Note that the ensemble of SAN models described with this setting can be trained in parallel since each SAN model is independent of others, and each individual SAN model takes unmodified images as input (Fig. 3).

Evaluation of an ensemble of models, that were trained independently, can be performed in two ways:

1) Averaging: Evaluating the ensemble of SANs by computing the average output image from the set of n outputs as shown in Fig. 4-A.
2) Gibbs: Randomly selecting the output of one SAN model (Fig. 4-B).

These two ensemble-based methods serve as a basis for the comparison with the proposed FlowSAN method, which is described in the following section.

B. FLOWSAN: CONNECTING MULTIPLE SAN MODELS

Assume there exists a large set of gender classifiers $\mathcal{G} = \{G_1, G_2, ..., G_g\}$, where each $G_i(I)$ predicts the probability that a face image I belongs to a male individual. Furthermore, suppose there exists a set of m face-matchers denoted by $\mathcal{M} = \{M_1, M_2, ..., M_m\}$, where each $M_i(I_a, I_b)$ computes the match score between a pair of face images, I_a and I_b. Our goal is to design an ensemble of n SAN models, $\mathcal{E} = \langle S_1, S_2, ..., S_n\rangle$, that, once they are sequentially stacked together, can be shown to generalize to confound unseen gender classifiers in $\mathcal{G}$. We hypothesize that stacking diverse SANs sequentially would have a cumulative effect, where each SAN adds perturbations to an input image that confound a particular gender classifier. Therefore, stacking SANs would enhance their generalizability in terms of decreasing the performance of multiple, diverse gender classifiers.

We define a recursive function $\Psi_{\mathcal{E}}(I_{\text{orig}}, t)$ for stacking SAN models in $\mathcal{E} = \{\text{SAN}_1, ..., \text{SAN}_n\}$, as follows:

$$\Psi_{\mathcal{E}}(I_{\text{orig}}, t) = \begin{cases} \text{SAN}_1(I_{\text{orig}}) & \text{if } t = 1, \\ \text{SAN}_t\left(\Psi_{\mathcal{E}}(I_{\text{orig}}, t-1)\right) & \text{otherwise.} \end{cases} \tag{7}$$

By varying t from 1 to n, $\Psi_{\mathcal{E}}(I_{\text{orig}}, t)$ produces a sequence of n output images $\langle I'_1, I'_2, ..., I'_n\rangle$:

- $t = 1 \rightarrow\ I'_1 = \Psi_{\mathcal{E}}(I_{\text{orig}}, 1) = \text{SAN}_1(I_{\text{orig}})$,
- $t = 2 \rightarrow\ I'_2 = \Psi_{\mathcal{E}}(I_{\text{orig}}, 2) = \text{SAN}_2\left(\text{SAN}_1(I_{\text{orig}})\right)$,

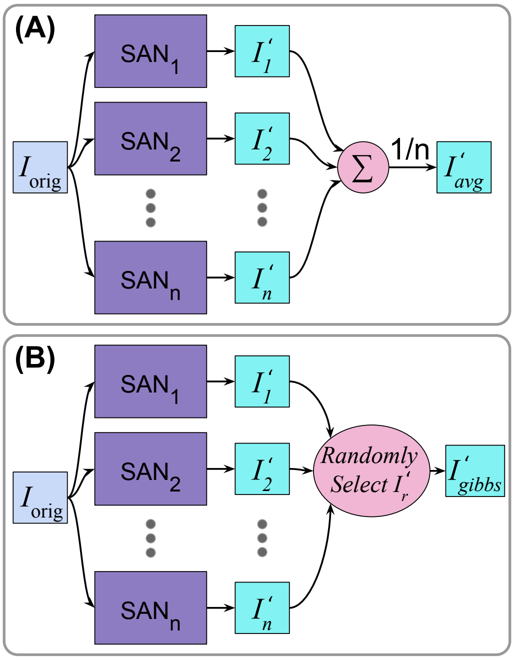

FIGURE 4: Two approaches for evaluating an ensemble of SAN models: Combining a set of n SAN models trained in the ensemble by (A) averaging n output images, and (B) randomly selecting an output (Gibbs).

- ...
- $t = n \rightarrow\ I'_n = \Psi_{\mathcal{E}}(I_{\text{orig}}, n) = \text{SAN}_n\left(... \text{SAN}_1(I_{\text{orig}})\right)$.

In particular, we hypothesize that for each $G_i \in \mathcal{G}$, the stacking of SAN models will progressively confound G_i. Since the individual SAN models were trained to have a minimal impact on face matching performance, we further hypothesize that the perturbations introduced in the output face images $\langle I'_1, ..., I'_n\rangle$ from the stacked SAN models should not substantially affect the face recognition performance of the matchers in $\mathcal{M}$.

Training Procedure for Stacking SAN Models

The goal of this work is to develop a model that leverages the image perturbations induced by individual, diverse SAN models to broaden the spectrum of diverse gender classifiers that can successfully be confounded. To accomplish this goal, we designed and evaluated the FlowSAN model, where multiple individually-trained SAN models were sequentially combined.

This section describes the training procedure for the FlowSAN model, where SAN models $i = 1, ..., n$ are trained in sequential order, each with their corresponding auxiliary gender classifier and an auxiliary face matcher, which is common among all SANs. The first SAN model, SAN$_1$ $\in \mathcal{E} = \{\text{SAN}_1, ..., \text{SAN}_n\}$, takes the original image as input and generates a perturbed output, I'_1, while using the auxiliary gender classifier G_1 during its training. Then, once SAN$_1$ is trained, the entire training dataset is transformed

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

by SAN_1, and the transformed data is then used for training the next SAN model while using its corresponding auxiliary gender classifier. This process is repeated for SAN models $i = 1, ..., n$, to obtain n SAN models that are trained in sequential order. Note that the matching loss is computed between face representation vectors (generated by a face matcher) of the SAN output with that of the corresponding original face image, as opposed to the input to the SAN model (which is already perturbed for $i \geq 2$). This is to ensure that the matching performance does not substantially decline as the sequence is expanded. Furthermore, we considered three different scenarios for the pixelwise dissimilarity loss:

1) Omitting the pixelwise dissimilarity loss term;
2) pixelwise dissimilarity with respect to the input, i.e., I'_{i-1} for SAN_i;
3) pixelwise dissimilarity loss with respect to the original image I_{orig} for each of SAN models $i = 1, ..., n$.

We evaluated all three different pixelwise loss function schemes listed above. However, we were unable to observe any noticeable differences except for some cases where the third scheme slightly outperformed the other two. Therefore, we only report the results of the third case in this paper. The training procedure is illustrated in Fig. 5.

Evaluating the FlowSAN Model

During the model evaluation, the auxiliary networks (the auxiliary gender classifiers and auxiliary face matchers) from the individual SANs are discarded, and the n SAN models are stacked in the same sequence they were trained, in order to enhance their generalizability to arbitrary gender classifiers. In the FlowSAN model, the first SAN model (SAN_1) takes an original image (I_{orig}) as input and generates a perturbed output image I'_1. This output image is then passed into the next SAN model in the sequence to obtain I'_2, and so forth. In general, the ith SAN model (SAN_i for $i = 2, ..., n$) takes the output of the previous SAN model (I'_{i-1}) as input and generates the perturbed output I'_i.

IV. EXPERIMENTS AND RESULTS

We designed two different protocols for training n SAN models:

(a) Training an ensemble of SANs independent of each other as described in [45] (see Section III-A);
(b) Training the FlowSAN model using the sequential procedure described in Section III-B.

Protocol (a) was adapted from [45] and is further described in Section III-A. For evaluating models trained in the ensemble, we applied two techniques: 1) taking the average output from SAN models which we denote as Ens-Avg, and 2) randomly selecting the output which we denote as Ens-Gibbs. In addition, similar to [45], we also define the *oracle best-perturbed* sample for a specific gender classifier, G:

$$\text{best}(I; \mathcal{E}, G) = \begin{cases} \underset{\text{SAN}_i \in \mathcal{E}}{\arg\min}\, G(\text{SAN}_i(I)) & \text{if } y = 1, \\ \underset{\text{SAN}_i \in \mathcal{E}}{\arg\max}\, G(\text{SAN}_i(I)), & \text{otherwise.} \end{cases} \tag{8}$$

The results of best-perturbed samples are denoted as Ens-Best. This analysis indicates which output from the ensemble model $\mathcal{E}$ has resulted in the highest prediction error for a particular gender classifier G if the best output is selected.

TABLE 2: Overview of datasets used in this study. The letters in the "Usage" column indicate the tasks for which the datasets were used. a: training auxiliary gender classifiers, b: SAN training, c: SAN evaluation, d: constructing unseen gender classifiers used for evaluating SAN models.

Dataset	#male	#female	Usage
CelebA-train	73,549	103,772	a, b
CelebA-test	7,929	11,511	c
MORPH-train	41,587	7,567	d
MORPH-test	4,643	863	c
LFW	10,064	2,905	d
MUCT	1,844	1,910	c
RaFD	1,008	600	c

The training of the FlowSAN model was initiated from the pre-trained individual SAN models in [45] and then trained for 10 additional epochs on the CelebA-train subset [49] (see Table 2) using the training procedure described in Section III-B. Then, the models were stacked successively to generate a sequence of perturbed output images, $\langle I'_1, \ldots, I'_n \rangle$.

As the FlowSAN model conceals the gender information in face images incrementally, it naturally produces a sequence of perturbed face images, where the length of this sequence is determined by its ensemble size. By varying the size of the ensemble, we can have a fair comparison between the ensemble approach vs. the FlowSAN model, such that the number of SANs used to obtain an output from the ensemble model is consistent with the number of SANs that are used to generate the output from the FlowSAN model.

For model evaluation and comparison, we used four test datasets: CelebA-test [49], MORPH-test [52], MUCT [53], and RaFD [54]. The number of male and female individuals in each dataset is listed in Table 2.

A. PERFORMANCE IN CONFOUNDING UNSEEN GENDER CLASSIFIERS

In order to evaluate the generalization performance of the three ensemble-based methods discussed in the previous section (Ens-Avg, Ens-Gibbs, Ens-Best) as well as the proposed FlowSAN model, we considered six independent gender classifiers. The experiments designed in this section assess how well the proposed models are able to confound gender classifiers that were unseen during training. These six gender classifiers include three models that were already trained: a commercial-of-the-shelf gender classifier (G-COTS), In-

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

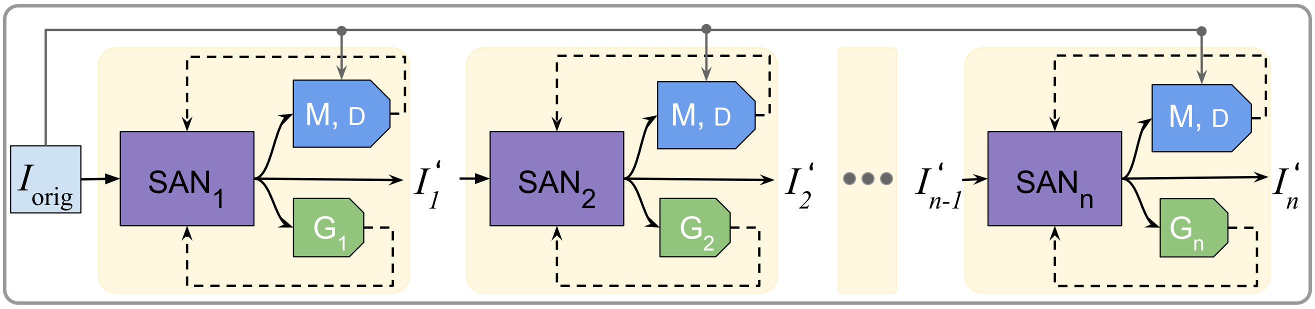

FIGURE 5: An illustration of a FlowSAN model: n SAN models are trained sequentially using n auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, ..., G_n\}$), and a face matcher M that computes face representation vectors for both input image I and the output of SAN model. Both auxiliary face matcher and the dissimilarity unit (D) use the original image along with the output of their corresponding SAN.

traFace [55], AFFACT [56], and three CNN models built in-house, which we refer to as CNN-1, CNN-2 (trained using MORPH-train and LFW, respectively), and CNN-3 (trained on the union of MORPH-train and LFW). Note that these three CNN models have shown a similar level of performance on the original test-sets, compared to the other three pre-trained gender predictors.

Fig. 6 shows the area under the ROC curve as a performance metric for evaluating the generalization performance of each unseen gender classifier on the four independent test datasets. The performance of these gender classifiers on the original images (before perturbations), as well as the outputs from the mixing approach by [23], is also shown for comparison.

In all cases, the FlowSAN approach results in lower AUC values (lower is better) of predictions made by unseen gender classifiers (Fig. 6) compared to the ensemble models Ens-Avg and Ens-Gibbs. In fact, the results of the stacking SAN models are almost on par with the oracle best-perturbed samples (Ens-Best) for each gender classifier. In some cases, the FlowSAN model even outperforms Ens-Best. **It is important to note that selecting the best-perturbed sample (from the individual SAN models) for each gender classifier without *a priori* knowledge of the classifier is infeasible in practice. Yet, we are able to outperform the best result using the FlowSAN model in several cases.**

Note that in a real privacy application, reaching a near random gender prediction performance (AUC ≈ 0.5, and Equal Error Rate (EER) ≈ 0.5) is desired for gender anonymization. As it can be seen in Fig. 6, both Ens-Avg and Ens-Gibbs methods produce samples that are mostly incapable of lowering the AUC of the unseen gender classifiers below 0.75 AUC. Based on the results shown in Fig. 6 (and the EER results shown in Fig. S1), it is evident that, in the majority of cases, a sequential stacking of three SAN models via FlowSAN produces the desired behavior in terms of face gender-anonymization, i.e., AUC ≈ 0.5 (similarly, EER ≈ 0.5). Although, in some cases, the 5th output from Ens-Avg and Ens-Gibbs resulted in a low, desired AUC of ≈ 0.5, it also has a substantially detrimental effect on the face matching performance, as discussed in Section IV-B.

As a result, we conclude that stacking three SAN models in FlowSAN is sufficient to achieve the best gender label anonymization performance across a set of different, unseen gender classifiers and face image datasets. Stacking fewer than three models affects unseen gender classifiers substantially less, and stacking more than three models induces such strong perturbations that flipping the predicted labels could again de-anonymize the perturbed face images with respect to their gender labels.

We shall note that our study was not the first to confound gender classifiers to produce random predictions. In [23], researchers proposed a face mixing approach that also leads to successful gender anonymization (approximately 0.5 AUC gender prediction performance for a specific gender classifier); however, this approach was unable to retain the face matching utility. In different studies, the researchers were able to retain face matching utility but without generalizing to arbitrary gender classifiers [9], [24]. Thus, the FlowSAN model we propose in this paper presents the first successful approach for satisfying both objectives: concealing gender information and retaining matching performance to a satisfactory degree across a variety of independent gender classifiers and face matchers.

B. RETAINING THE PERFORMANCE OF UNSEEN FACE MATCHERS

To assess the effect of the gender perturbations on the matching accuracy, we considered four different unseen face matchers. This includes a commercial-of-the-shelf face matcher (M-COTS), which has shown state-of-the-art performance in face recognition, as well as three publicly available algorithms that provide face representation vectors: DR-GAN [57], FaceNet [58], and OpenFace [59]. For the latter three models, we measured the cosine similarity between face representation vectors obtained from the original images and face representation vectors obtained from the SAN-perturbed output images.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

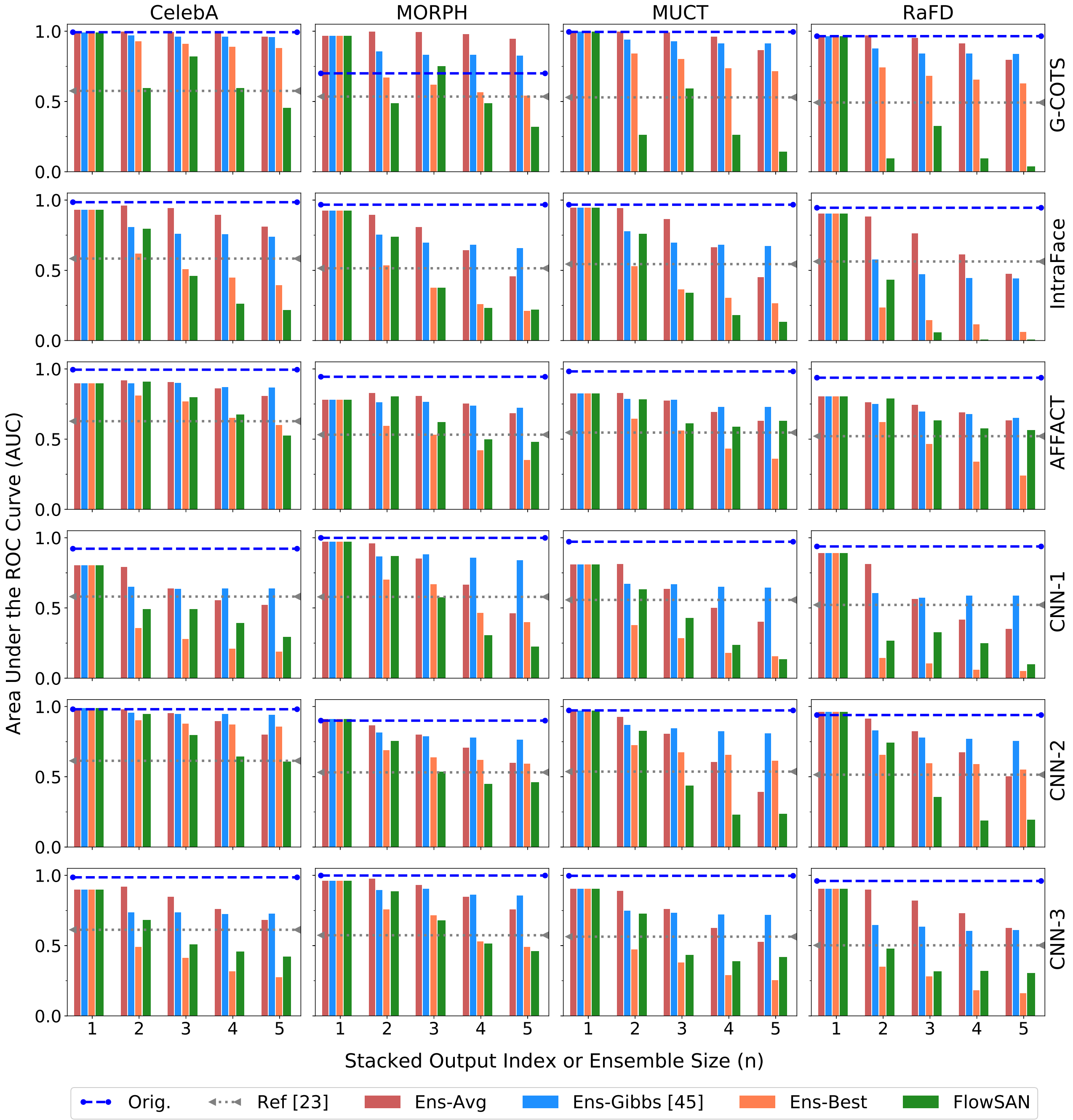

FIGURE 6: Area under the ROC curve (AUC) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images ("Orig.") is shown (blue dashed line) as well as the perturbed samples using the three ensemble-based models (Ens-Avg, Ens-Gibbs, Ens-Best), the proposed FlowSAN model, and the face mixing approach [23] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, ..., I'_5 \rangle$ obtained by varying the ensemble size, n. In almost all cases, stacking three SAN models results in an AUC of approximately 0.5 (perfectly random gender prediction).

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

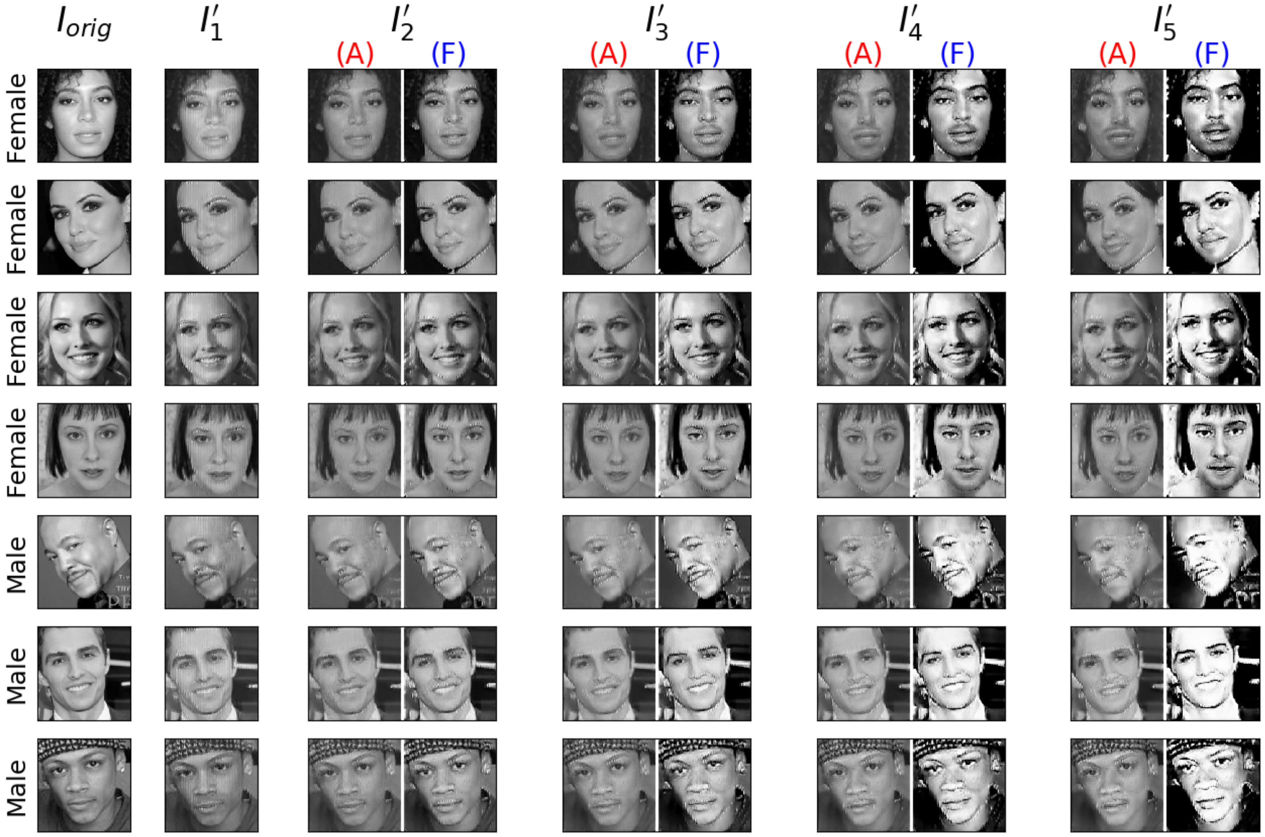

FIGURE 7: A randomly selected set of examples showing input face images and their outputs from I'_1 to I'_5 using (A) the ensemble model, Ens-Avg, and (F) using the FlowSAN model.

Fig. 8 shows the True Match Rate (TMR) values at False Match Rate (FMR) of 0.1% for different ensemble methods. In most cases, the performance of the face matchers regarding the first three outputs (I'_1, I'_2, and I'_3) is similar and relatively close to the matching performance on original images. We note that stacking three SANs in FlowSAN yields the desired performance with regard to confounding unseen gender classifiers. Therefore, the evaluation of the face matching performance for stacking more than three SANs I'_3 (i.e., I'_4 and I'_5) is only included for completeness.

Comparing the performance of face matchers for equal values of n, we observe that the face matchers appear to perform slightly better on outputs produced by the ensemble model compared to the FlowSAN model. However, the extent to which the gender classification performance is reduced by the two models is not the same for equal values of n (Table 3). The ensemble model requires at least $n = 5$ individual SAN models to be able to confound unseen gender classifiers to reach the same level of gender anonymization as the FlowSAN model with $n = 3$. Therefore, if we compare the ensemble models with $n = 5$ to the FlowSAN model with $n = 3$, the face matchers perform substantially better on the face image outputs by the FlowSAN model (Fig. 8). Further, note that the performance of M-COTS on CelebA on the original images is already as low as 85.6%. In fact, all matchers perform poorly on the CelebA dataset, which may be due to different face orientations captured in the wild.

Preserving Privacy

The overall average performance considering the two target objectives of this study, i.e., confounding gender classifiers and retaining the matching utility of face images, is provided in Table 3. In this analysis, the average EER results of all six gender classifiers over all four evaluation datasets were computed for original images, outputs from Ref. [23], as well as outputs from the stacking and the ensemble models using $n = 3$ and $n = 5$. The results clearly show that the FlowSAN model outperforms the ensemble-based methods, including the oracle-best results. On the other hand, the average true matching rate (TMR) values, at a false matching rate (FMR) of 0.1%, are also computed similarly, and the results indicate that the Ens-Gibbs method has the highest performance for both ensemble sizes, while the performance of the FlowSAN model at $n = 3$ is ranked as second, but it is very close to that of Ens-Gibbs. The detailed EER results for each gender classifier is provided in Table S1.

Computational Efficiency

The overall computational cost for training the ensemble-

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

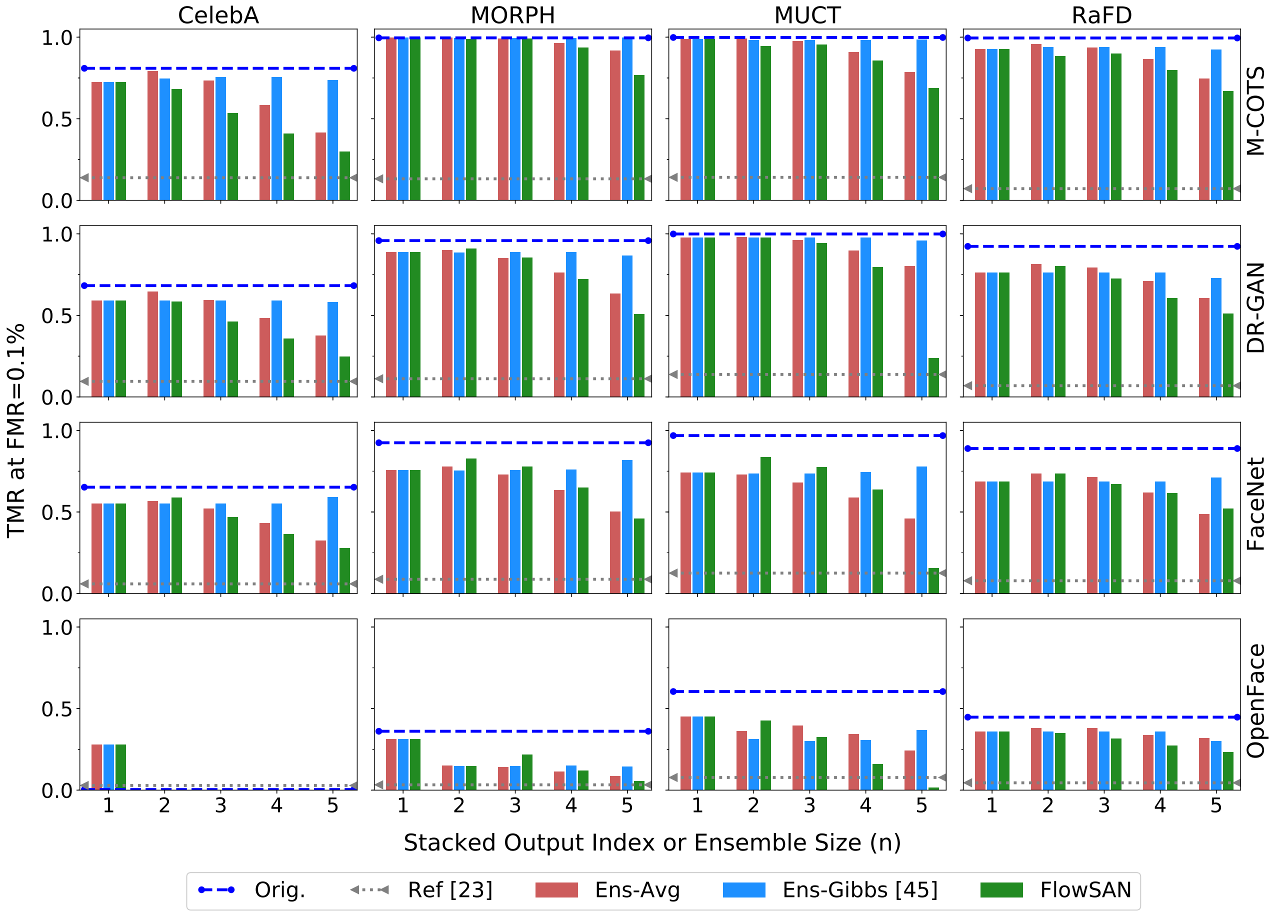

FIGURE 8: True Match Rate (TMR) values at False Match Rate (FMR) of 0.1% obtained using the four unseen face matchers, M-COTS, DR-GAN, FaceNet, and OpenFace, on the original images as well as the perturbed outputs after stacking the SAN models, and using the ensemble models (Ens-Avg and Ens-Gibbs). Note that the matchers' performance obtained after applying the first three SANs in the FlowSAN model is close to the original performance, but it further diminishes when the sequence is extended.

TABLE 3: Comparing the overall average performance of six unseen gender classifiers and four unseen face matchers over the four evaluation datasets using $n = 3$ or $n = 5$ SAN models. This shows that stacking 3 SAN models results in gender anonymization EER $\approx$ 0.5, while the the average matching performance is still comparable to the unmodified images as well as the matching performance on the outputs form other existing methods.

	Gender:		**Matching:**	
	EER		**TMR at FMR=0.1%**	
Orig.	10%		76.3%	
Ref [23]	46%		9.1%	
	$n = 3$	$n = 5$	$n = 3$	$n = 5$
Ens-Avg	23%	40%	64.9%	48.1%
Ens-Gibbs	29%	31%	65.2%	65.6%
Ens-Best	48%	57%	–	–
FlowSAN	**49**%	**64**%	61.9%	35.4%

based approach and the FlowSAN model is similar, except that FlowSAN requires an additional data transformation step between each consecutive SAN training. However, the ensemble approach comes with a bigger advantage that the individual SAN models can be trained in parallel, while the SAN models in the FlowSAN model have to be trained sequentially.

V. CONCLUSION

In this work, we address one of the main limitations of previous gender privacy methods, namely, their inability to generalize across multiple previously unseen gender classifiers. In this regard, we propose the FlowSAN method that sequentially combines diverse perturbations for an input face image to confound the gender information with respect to an arbitrary gender classifier. We compared the performance of the proposed FlowSAN model with two ensemble-based approaches: 1) using the average output of SAN models trained independently of each other (Ens-Avg); 2) randomly

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

selecting the output from the SAN models in the ensemble (Ens-Gibbs).

Our experiments show that the FlowSAN method outperforms the other ensemble-based approaches in terms of confounding a range of gender classifiers. More importantly, while gender classification is successfully confounded, face matching accuracy is mostly retained for all perturbed output face images, thereby preserving the biometric utility of the gender-anonymous face images.

While this work only focused on confounding gender labels to demonstrate this method's efficacy in hiding soft-biometric attributes, our method can be readily extended and generalized to incorporate other soft-biometric attributes (for example, age and ethnicity), which will be the subject of future studies.

VI. ACKNOWLEDGMENT

This study is supported by the National Science Foundation under Grant Number 1618518.

REFERENCES

[1] A. Jain, A. A. Ross, and K. Nandakumar, Introduction to biometrics. Springer Science & Business Media, 2011.

[2] K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2D and 3D facial data," in ACM Workshop on Multimodal User Authentication. Citeseer, 2003, pp. 25–32.

[3] A. Dantcheva, P. Elia, and A. Ross, "What else does your biometric data reveal? A survey on soft biometrics," IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 441–467, 2016.

[4] K. Sundararajan and D. L. Woodard, "Deep learning for biometrics: A survey," ACM Comput. Surv., vol. 51, no. 3, pp. 65:1–65:34, May 2018. [Online]. Available: http://doi.acm.org/10.1145/3190618

[5] A. K. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," IEEE Transactions on Information Forensics and Security, vol. 1, no. 2, pp. 125–143, 2006.

[6] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," IBM Systems Journal, vol. 40, no. 3, pp. 614–634, 2001.

[7] I. Natgunanathan, A. Mehmood, Y. Xiang, G. Beliakov, and J. Yearwood, "Protection of privacy in biometric data," IEEE Access, vol. 4, pp. 880–892, 2016.

[8] A. Morales, J. Fierrez, and R. Vera-Rodriguez, "SensitiveNets: Learning agnostic representations with application to face recognition," arXiv preprint arXiv:1902.00334, 2019.

[9] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing k-facial attributes via adversarial perturbations," arXiv preprint arXiv:1805.09380, 2018.

[10] T. Sim and L. Zhang, "Controllable face privacy," in 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG), vol. 4, 2015, pp. 1–8.

[11] B. Medcn, P. Peer, and V. Štruc, "Selective face deidentification with end-to-end perceptual loss learning," in IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2018.

[12] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer, "k-Same-Net: k-Anonymity with generative deep neural networks for face deidentification," Entropy, vol. 20, no. 1, p. 60, 2018.

[13] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations," Applied Intelligence, 2019.

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

[15] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A survey to deep facial attribute analysis," arXiv preprint arXiv:1812.10265, 2018.

[16] X. Li, J.-M. Wen, A.-L. Chen, and B. Chen, "A method for face fusion based on variational auto-encoder," in 15th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2018, pp. 77–80.

[17] S. Yang, A. Wiliem, S. Chen, and B. C. Lovell, "Using lip to gloss over faces in single-stage face detection networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 640–656.

[18] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Right for the right reason: Training agnostic networks," in International Symposium on Intelligent Data Analysis. Springer, 2018, pp. 164–174.

[19] T. Swearingen and A. Ross, "Label propagation approach for predicting missing biographic labels in face-based biometric records," IET Biometrics, vol. 7, no. 1, pp. 71–80, 2017.

[20] A. Acquisti and R. Gross, "Predicting social security numbers from public data," Proceedings of the National Academy of Sciences, vol. 106, no. 27, pp. 10 975–10 980, 2009.

[21] E. J. Kindt, Privacy and data protection issues of biometric applications. Springer, 2013.

[22] A. Acquisti, L. K. John, and G. Loewenstein, "What is privacy worth?" The Journal of Legal Studies, vol. 42, no. 2, pp. 249–274, 2013.

[23] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in European Conference on Computer Vision Workshop. Springer, 2014, pp. 682–696.

[24] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in Proc. of International Joint Conference on Biometrics (IJCB), 2017.

[25] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016," Official Journal of the European Union, vol. L 119, 2016.

[26] D. Bobeldyk and A. Ross, "Predicting gender and race from near infrared iris and periocular images," arXiv preprint arXiv:1805.01912, 2018.

[27] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," Pattern Recognition Letters, vol. 70, pp. 80–86, 2016.

[28] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Gender classification by deep learning on millions of weakly labelled images," in IEEE 16th International Conference on Data Mining Workshops, 2016, pp. 462–467.

[29] J. R. Lyle, P. E. Miller, S. J. Pundlik, and D. L. Woodard, "Soft biometric classification using periocular region features," in Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Sep. 2010, pp. 1–7.

[30] S. Arora and M. Bhatia, "A robust approach for gender recognition using deep learning," in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018, pp. 1–6.

[31] S. Chhabra, P. Majumdar, M. Vatsa, and R. Singh, "Data fine-tuning," arXiv preprint arXiv:1812.03944, 2018.

[32] D. Bobeldyk and A. Ross, "Predicting soft biometric attributes from 30 pixels: A case study in NIR ocular images," in IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.

[33] A. Sgroi, K. W. Bowyer, and P. J. Flynn, "The prediction of old and young subjects from iris texture," in International Conference on Biometrics (ICB), 2013, pp. 1–5.

[34] M. S. Nixon, P. L. Correia, K. Nasrollahi, T. B. Moeslund, A. Hadid, and M. Tistarelli, "On soft biometrics," Pattern Recognition Letters, vol. 68, pp. 218–230, 2015.

[35] L. Du, M. Yi, E. Blasch, and H. Ling, "GARP-face: Balancing privacy protection and utility preservation in face de-identification," in IEEE International Joint Conference on Biometrics, 2014.

[36] Y. Sun, M. Zhang, Z. Sun, and T. Tan, "Demographic analysis from biometric data: Achievements, challenges, and new frontiers," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.

[37] D. Bobeldyk and A. Ross, "Analyzing covariate influence on gender and race prediction from near-infrared ocular images," IEEE Access, vol. 7, pp. 7905–7919, 2019.

[38] S. Lagree and K. W. Bowyer, "Predicting ethnicity and gender from iris texture," in IEEE International Conference on Technologies for Homeland Security (HST), 2011, pp. 440–445.

[39] A. M. Badawi, M. Mahfouz, R. Tadross, and R. Jantz, "Fingerprint-based gender classification." IPCV, vol. 6, pp. 41–46, 2006.

[40] S. Raschka and V. Mirjalili, Python Machine Learning, 2nd Ed. Birmingham, UK: Packt Publishing, 2017.

[41] P. Lewis and J. Carrie Wong, "Facebook employs psychologist whose firm sold data to Cambridge Analytica," https://www.theguardian.com/news/2018/mar/18/facebook-cambridge-analytica-joseph-chancellor-gsr, 2018, [Online; accessed 26-April-2018].

[42] C. Garvie, A. Bedoya, and J. Frankle, The Perpetual Line-Up: Unregulated Police Face Recognition in America. Georgetown Law, Center on Privacy & Technology, 2016.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

[43] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in European Conference on Computer Vision. Springer, 2018, pp. 556–572.

[44] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-Adversarial Networks: Convolutional autoencoders for imparting privacy to face images," in Proc. of 11th IAPR International Conference on Biometrics (ICB 2018), Gold Coast, Australia, 2018.

[45] V. Mirjalili, S. Raschka, and A. Ross, "Gender Privacy: An ensemble of Semi Adversarial Networks for confounding arbitrary gender classifiers," in Proc. of 9th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Los Angeles, CA, 2018.

[46] P. Chandan Roy and V. Naresh Boddeti, "Mitigating information leakage in image representations: A maximum entropy approach," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[47] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," Computer Vision and Image Understanding, vol. 61, no. 1, pp. 38–59, 1995.

[48] A. Rozsa, M. Günther, E. M. Rudd, and T. E. Boult, "Are facial attributes adversarially robust?" in 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 3121–3127.

[49] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

[50] Y. Bengio, "Learning deep architectures for AI," Foundations and trends® in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.

[51] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in British Machine Vision Conference, vol. 1, 2015, p. 6.

[52] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in 7th International Conference on Automatic Face and Gesture Recognition,. IEEE, 2006, pp. 341–345.

[53] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," Pattern Recognition Association of South Africa, vol. 201, no. 0, 2010.

[54] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg, "Presentation and validation of the radboud faces database," Cognition and Emotion, vol. 24, no. 8, pp. 1377–1388, 2010.

[55] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, and J. Cohn, "Intraface," in 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, pp. 1–8.

[56] M. Günther, A. Rozsa, and T. E. Boult, "AFFACT: Alignment-free facial attribute classification technique," in IEEE International Joint Conference on Biometrics (IJCB), 2017, pp. 90–99.

[57] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[58] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[59] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: a general-purpose face recognition library with mobile applications," CMU School of Computer Science, vol. 6, 2016.

• • •

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

VII. SUPPLEMENTARY MATERIALS

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

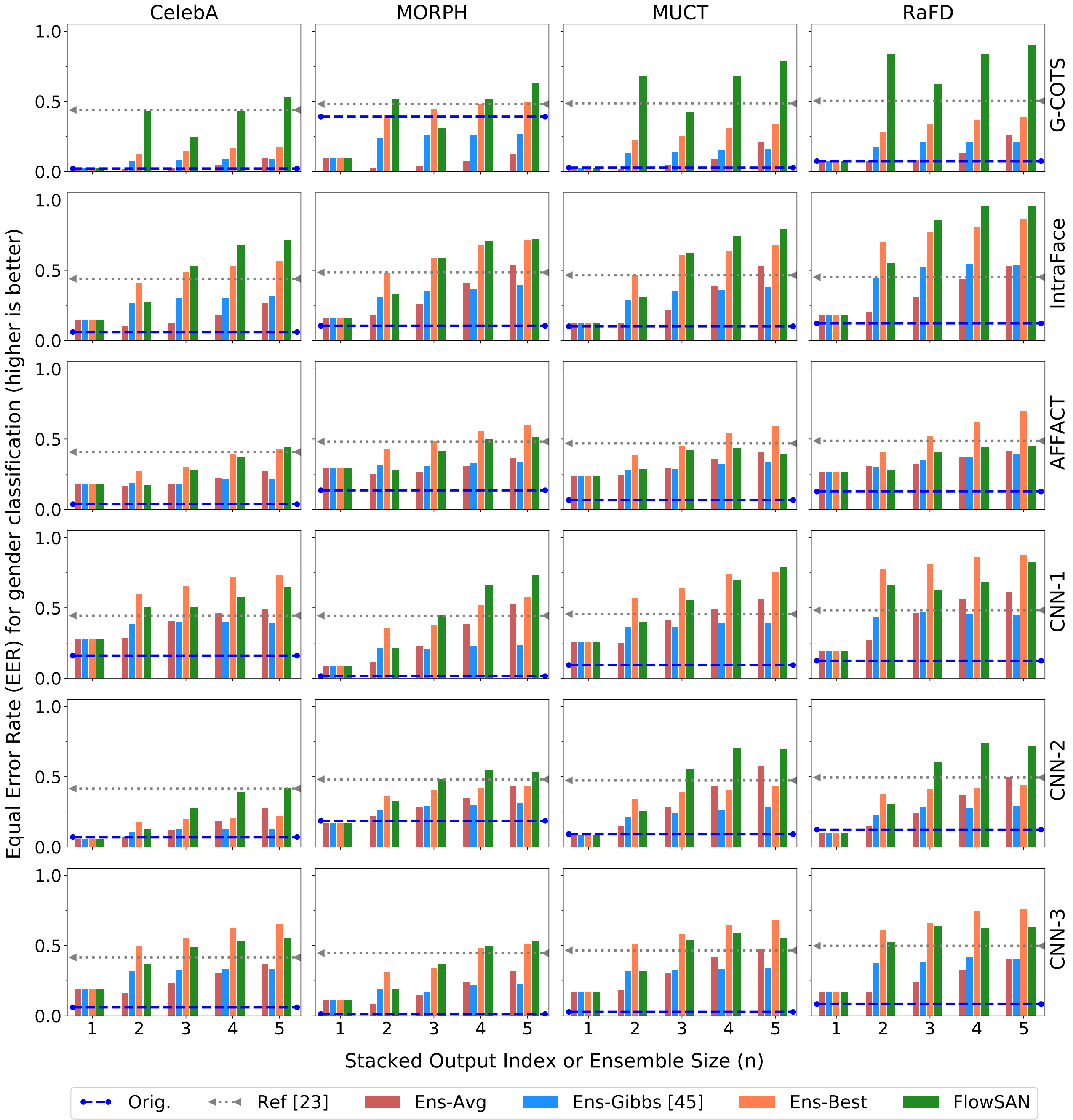

FIGURE S1: Equal Error Rate (EER) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images ("Orig.") is shown (blue dashed line) as well as the perturbed samples using the three ensemble models (Ens-Avg, Ens-Gibbs, Ens-Best), the proposed FlowSAN model, and the face mixing approach [23] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, ..., I'_5 \rangle$ obtained by varying the ensemble size, n.

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2924619, IEEE Access

TABLE S1: Comparing the overall average Equal Error Rate (EER) of six unseen gender classifiers averaged over all four evaluation datasets (CelebA-test, MORPH-test, MUCT, and RaFD), higher is better. Note that the Ens-Best method is the result of “oracle best” selected classifier from an ensemble of multiple SANs, which assumes knowledge of the gender classifier. While this is impractical in a real-world privacy application, we show the results for comparison purposes.

Gender	Orig.	Ref. [23]	$n = 3$				$n = 5$			
Classifier			Ens-Avg	Ens-Gibbs	Ens-Best	FlowSAN	Ens-Avg	Ens-Gibbs	Ens-Best	FlowSAN
G-COTS	0.13	0.48	0.05	0.18	0.30	**0.40**	0.17	0.18	0.35	**0.71**
IntraFace	0.10	0.46	0.23	0.38	0.61	**0.65**	0.47	0.41	0.71	**0.80**
AFFACT	0.09	0.46	0.26	0.28	**0.44**	0.38	0.36	0.32	**0.58**	0.45
CNN-1	0.10	0.46	0.38	0.36	**0.62**	0.53	0.55	0.38	0.74	**0.75**
CNN-2	0.12	0.47	0.23	0.23	0.35	**0.48**	0.45	0.25	0.38	**0.59**
CNN-3	0.05	0.46	0.23	0.30	**0.53**	0.51	0.39	0.32	**0.65**	0.57
Average	0.10	0.46	0.23	0.29	0.48	**0.49**	0.40	0.31	0.57	**0.64**

This work is licensed under a Creative Commons Attribution 3.0 License. For more information, see http://creativecommons.org/licenses/by/3.0/.