PrivacyNet: Semi-Adversarial Networks for Multi-attribute Face Privacy

Vahid Mirjalili[1]   Sebastian Raschka[2]   Arun Ross[1]

[1] Department of Computer Science and Engineering, Michigan State University

[2] Department of Statistics, University of Wisconsin – Madison

Abstract—Recent research has established the possibility of deducing soft-biometric attributes such as age, gender and race from an individual's face image with high accuracy. However, this raises privacy concerns, especially when face images collected for biometric recognition purposes are used for attribute analysis without the person's consent. To address this problem, we develop a technique for imparting soft biometric privacy to face images via an image perturbation methodology. The image perturbation is undertaken using a GAN-based Semi-Adversarial Network (SAN) - referred to as PrivacyNet - that modifies an input face image such that it can be used by a face matcher for matching purposes but cannot be reliably used by an attribute classifier. Further, PrivacyNet allows a person to choose specific attributes that have to be obfuscated in the input face images (e.g., age and race), while allowing for other types of attributes to be extracted (e.g., gender). Extensive experiments using multiple face matchers, multiple age/gender/race classifiers, and multiple face datasets demonstrate the generalizability of the proposed multi-attribute privacy enhancing method across multiple face and attribute classifiers.

Index Terms—privacy, semi-adversarial, neural networks, autoencoder, face image, perturbation, soft biometrics, deep learning.

◆

1 INTRODUCTION

THE use of automated methods to compare face images in order to determine the identity of an individual or to verify a claimed identity is known as face recognition [1]. Face recognition has been widely used in several applications, including, access control in smartphones, surveillance for public safety, and finding missing children [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Examples of face recognition methods include Elastic Bunch Graph Matching [12], Active Appearance Models [13], Sparse Representation [14], as well as more recent techniques based on Deep Learning [15].

The primary purpose of collecting and storing face images in a biometric system is for the recognition of individuals. Yet, face images stored in a database implicitly contain auxiliary information about each individual [16], [17]. These auxiliary information, sometimes referred to as *soft biometric attributes*, include gender, age, ethnicity, body mass index, and health characteristics [16], [18], [19].

Soft biometrics can facilitate a large variety of applications, such as improving face recognition performance, clustering users, or developing targeted advertisements [20], [21], [22]. Recent advances in machine learning has made it possible to extract such soft biometric attributes from face images automatically and, in many cases, with a high degree of accuracy [16], [18], [23]. However, users of such biometric systems may prefer not to be profiled based on their demographic attributes and may wish to opt-out of such services due to privacy concerns [24], [25], [26]. In this regard, certain privacy laws allow users to choose what information about themselves to reveal and what information to conceal [27], [28], [29], [30]. Moreover, in the near future, biometric applications are expected to implement best-practices with regard to respecting the privacy of users by preventing automatic information extraction from face

images in the absence of users' consent [24], [31], [32]. However, even if information is not extracted intentionally, user images stored in a database are still susceptible to privacy breaches via third party users or applications [33]. Thus, to provide actionable means and guarantees for preventing the automatic mining of personal information from face images, recent research has explored the possibility of imparting soft biometric privacy to face images by modifying the image data directly [34], [35], [36], [37], [38], [39].

To provide a practical approach for imparting gender privacy to face images, Mirjalili et al. previously developed the Semi-Adversarial Network (SAN) model [36]. This SAN model is able to conceal gender information in face images while retaining satisfactory face matching accuracy. In later studies, improvements of the SAN model resulted in state-of-the-art matching performance with arbitrary face matchers under the constraint that arbitrary gender classifiers were not able to extract gender information from the modified face images [40], [41].

While the previously developed SAN model is only capable of hiding gender information, in this paper, we propose a new model, named *PrivacyNet*, for imparting multi-attribute privacy to face images that includes gender,[1] race,[2] and age. The overall objective of this work is to develop a model that can induce selective (which attributes to conceal) and collective (how many attributes to conceal) perturbations to impart soft biometric privacy to face images

---

1. The terms 'gender' and 'sex' have been used interchangeably in the biometric literature. It must be noted that gender is a social or cultural construct, while sex is based on biological characteristics.

2. The terms 'race' and 'ethnicity' have been used interchangeably in the biometric literature. An exact definition of either of these two terms appears to be debatable.

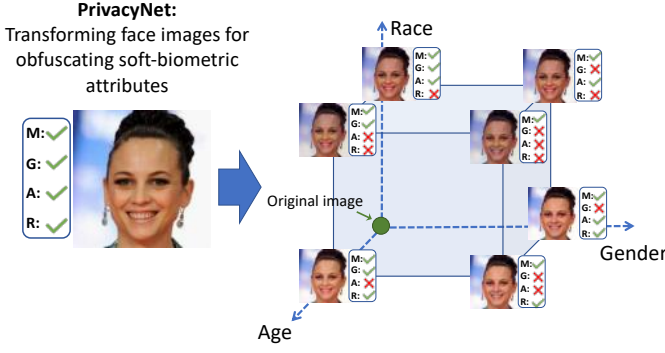**PrivacyNet:** Transforming face images for obfuscating soft-biometric attributes

Fig. 1: Illustration of the overall objective of this work: transforming an input face image across three orthogonal axes for imparting multi-attribute privacy selectively while retaining biometric recognition utility. The abbreviated letters are M: Matching, G: Gender, A: Age, and R: Race.

(Fig. 1) while retaining biometric matching performance. The main differences between PrivacyNet and previous work on perturbing facial attributes while maintaining biometric matching utility are as follows:

- The adversarial examples generated in [42] were only focused on the gender attribute, and the resulting outputs were not generalizable to unseen[3] gender classifiers.
- The generalizability issue of [42] was addressed by developing semi-adversarial networks [36] and their subsequent variants [41] and [40]. However, all of the aforementioned methods are focused on a single attribute, gender, whereas PrivacyNet can perturb race, gender, and age attributes.
- Furthermore, PrivacyNet is a GAN-based model with cycle consistency loss, in contrast to the regular convolutional autoencoders used in [36], [40], [41]

The organization of this paper is as follows. Section 2 discusses existing methods for imparting privacy to face images. In Section 3, we describe our proposed PrivacyNet method, including a detailed description of the neural network architecture and datasets used in this study. Section 4 presents and discusses the experimental results obtained by analyzing the matching utility of face images after perturbing facial attributes.

2 RELATED WORK

With recent advances in machine learning and deep learning for computer vision, the prediction of soft biometric attributes such as age, gender, and ethnicity from facial biometric data has been widely studied [16], [43], [44], [45], [46]. For instance, the use of convolutional neural networks for predicting the gender from face images has resulted in models with almost perfect prediction accuracy [45], [47], [48], [49], [50]. Methods for estimating the apparent age from face images are similarly well studied, and current-state of the art methods can predict the apparent age of a person

3. The term "unseen" refers to matchers, classifiers or data that were not used during training.

with a prediction error below three years on average [43], [44], [51], [52].

While tremendous progress has been made towards the automatic extraction of personal attributes of face images, the development of methods and techniques for imparting soft biometric privacy is still a relatively recent area of research. In 2014, Othman and Ross introduced the concept of soft biometric privacy, where a face image is modified such that the gender information is confounded while the recognition utility of the face image is preserved [34]. The researchers proposed a face mixing approach, where a face image is morphed with a candidate face image from the opposite gender. As a result, the resulting mixed face image contains both male and females features such that the gender information was fully anonymized. Sim and Zhang then developed methods for imparting soft biometric privacy to multiple attributes based on multi-modal discriminant analysis, in which certain attributes can be selectively suppressed while retaining others [35]. They proposed a technique that decomposes a face image representation into orthogonal axes corresponding to gender, age, and ethnicity, and the identity information is left as a residual of this decomposition. This enables transforming a face image along one axis resulting in modifying the corresponding attribute, while other information of the face image remains visibly unchanged to the human eye. They also showed that their proposed method can alter identities of face images, which is useful for face de-identification [53]. However, Sim and Zhang's [35] method cannot explicitly preserve the matching performance of transformed face images and, therefore, the biometric utility of the resulting face images is severely diminished.

In 2013, Szegedy et al. [54] studied the vulnerability of Deep Neural Networks (DNNs) towards adversarial perturbations. Adversarial perturbations are small perturbations added to an input image, typically imperceptible to a human observer, that can cause the DNN to misclassify images with high confidence. In recent years, several methods for generating such adversarial perturbations have been proposed, and the development of methods that make DNN-based models more robust against these so-called adversarial attacks remains an active area of research [55], [56], [57], [58], [59], [60], [61], [62]. The vulnerability to adversarial attacks raises several security concerns for the use of machine learning systems in computer vision applications [32], [55], [63], [64]. Recently, Rozsa et al. [65] investigated the robustness of binary facial attribute classifiers to adversarial attacks. Based on the concept of adding adversarial perturbations to an input image, Mirjalili and Ross [42] investigated the possibility of generating adversarial perturbations for imparting soft-biometric privacy to face images. This scheme was further extended by Chhabera et al. [37] to conceal multiple face attributes simultaneously. While these perturbation-based methods are shown to successfully derive adversarial examples based on a specific attribute classifier, the perturbed output images are not generalizable across unseen attribute classifiers. For a real-world privacy application, generalizability of adversarial examples to unseen attribute classifiers is critical [41].

Recently, methods have been developed that impart privacy through the design and use of specific face representa-

tion vectors, which have been derived from the original face images without including the sensitive information that is to be concealed [66], [67], [68], [69], [70]. For instance, the SensitiveNet [68] model generates agnostic face representations for biometric recognition such that gender and race information are removed from these representations [71]. However, storing face representation vectors may not be desirable in many applications since these vectors are neither interpretable by humans nor compatible with future face recognition software. In this work, we develop a generally applicable method that applies perturbations to the face images directly instead of the derived representations.

In previous work [36], Mirjalili et al. developed a deep learning-based model to generate perturbed examples for obfuscating gender information in face images. The neural network was coined Semi-Adversarial Network (SAN) and is composed of a convolutional autoencoder for synthesizing face images such that the gender information in the synthesized images is obfuscated while their matching utility is preserved. The SAN model is trained using an auxiliary gender classifier and an auxiliary face matcher. After training, the auxiliary subnetworks are discarded and the convolutional autoencoder is used for performance evaluation on unseen data. It was shown that this model is able to suppress gender information as assessed by some *unseen*[4] attribute classifiers while the matching utility, assessed by unseen face matchers, was retained. Moreover, the generalizability of SAN models to fool arbitrary gender classifiers can be further enhanced by diversifying the auxiliary classifiers during training [41] or by combining multiple, diverse SAN models [40].

With the development of Generative Adversarial Networks (GANs) [72], [73], [74], different GAN-based models for editing facial attributes have been proposed in the literature [75], [76], [77], [78], [79]. These methods focus on selected face attributes such as facial hair, eyeglasses, hair color and skin-tone, which can be modified without significantly affecting other attributes and biometric recognition capabilities. The aforementioned facial attributes are only loosely connected to the identity information in the face image. On the other hand, gender and race are intricately tied to the identity of the subjects, and, therefore, cannot be easily modified without affecting the recognition capabilities of face matchers. Consequently, the aforementioned GAN models and the PrivacyNet method we propose in this paper have different objectives. To the best of our knowledge, the only GAN-based model that has a similar objective as PrivacyNet is the Deep Identity-Aware Transfer (DIAT) model [80]. Besides transferring attributes that are loosely connected with identity, DIAT also considers gender transfer. DIAT relies on a perceptual loss via a pre-trained VGG-face [81] network to preserve the recognition performance of the attribute-transferred output images. However, their experimental results still showed significant decrease in face verification performance after transferring the eyeglass attributes of the face images.

The main contribution of this work is the design of a

---

4. In contrary to "auxiliary" classifiers, the term "unseen" indicates that the attribute classifier (or face matcher) was not used during the training stage.

multi-attribute face privacy model to provide **controllable soft-biometric privacy**. This proposed GAN-based privacy model, which we refer to as "PrivacyNet," modifies an input face image to **obfuscate soft-biometric attributes** while **maintaining the recognition capability** on the generated face images. To the best of our knowledge, the "PrivacyNet" model proposed in this paper is the first method for multi-attribute privacy that generalizes to unseen attribute classifiers while preserving the recognition utility of face images across unseen face matchers.

3 PROPOSED METHOD

3.1 Problem Formulation

Given a face image X, let S_{obf} be a set of face attributes to be obfuscated and S_{keep} be a set of attributes to be preserved. The overall objective is to find function ϕ that applies some perturbations to the input image X such that $X' = \phi(X)$ has the following properties:

- For a soft biometric attribute $a \in S_{\text{obf}}$, the performance of an unseen attribute classifier f_a is substantially reduced.
- For the remaining set of attributes $b \in S_{\text{keep}}$, the performance of an arbitrary classifier f_b is not noticeably adversely affected; that is, the performance of an attribute classifier f_b on perturbed image X' is close to its performance on the original face image X.
- The primary biometric utility, which is face recognition, must be retained for the modified face image, X'. In other words, given pairs of image examples before ($\langle X_1, X_2 \rangle$) and after ($\langle X'_1, X'_2 \rangle$) perturbations, the matching performance as assessed by an arbitrary face matcher (f_M) is not substantially affected, i.e.,
$f_M(X_1, X_2) \approx f_M(X'_1, X'_2) \approx f_M(X_1, X'_2) \approx f_M(X'_1, X_2)$.

3.2 PrivacyNet

According to the objectives described in Section 3.1, the PrivacyNet neural network architecture (Fig. 2A) is composed of four sub-networks: A generator (G) that modifies the input image, a source discriminator (D_{src}) which determines if an image is real or modified, an attribute classifier (D_{attr}) for predicting facial attributes, and an auxiliary face matcher (M) for biometric face recognition. Along with the input image, both generator and discriminator receive the attribute labels as conditional variables, that are spanned to the same width and height as the input image, (224×224). Together, these subnetworks form a cycle-consistent GAN [82] as illustrated in Fig. 2B. Given an RGB input face image X, the attribute label vector $\mathcal{V}_0 \in \mathbb{Z}^c$ corresponds to the ground truth attribute labels of the original face image. The target label vector $\mathcal{V}_t \in \mathbb{Z}^c$ (c is the total number of attributes) denotes the desired facial attributes for modifying the face image. Given a target vector $\mathcal{V}_t \neq \mathcal{V}_0$, the objective of the generator G is to synthesize a new image $X' = G(X, \mathcal{V}_t)$ such that X' is mapped to the target label vector $\mathcal{V}_t$ by an attribute classifier D_{attr}. The other component of the GAN model is a source discriminator D_{src}, which is trained
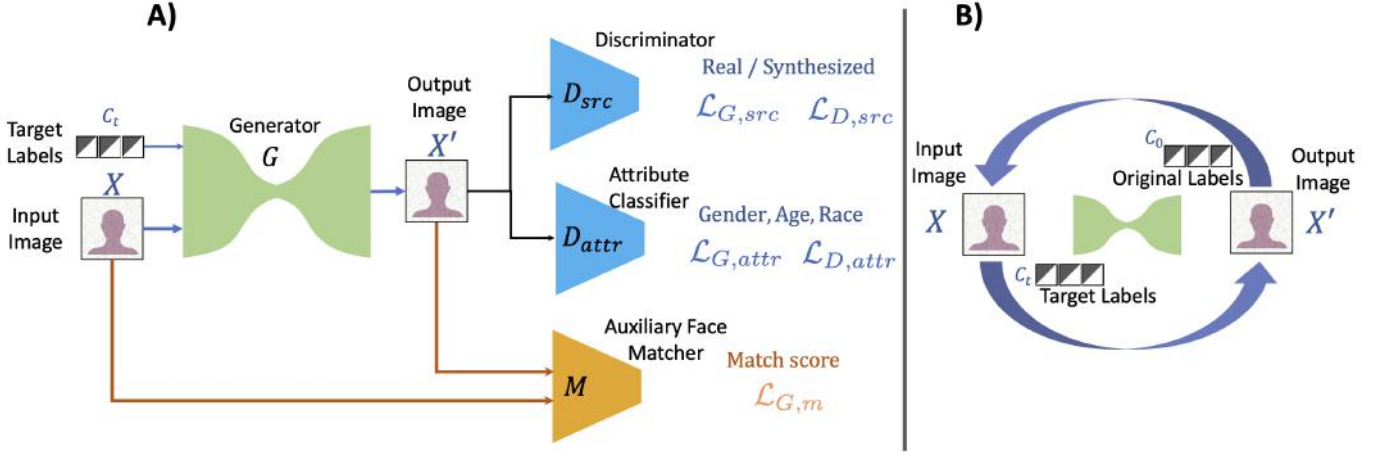
Fig. 2: Schematic representation of the architecture of PrivacyNet for deriving perturbations to obfuscate three attribute classifiers – gender, age and race – while allowing biometric face matchers to perform well. (A) Different components of the PrivacyNet: generator, source discriminator, attribute classifier, and auxiliary face matcher. (B) Cycle-consistency constraint applied to the generator by transforming an input face image to a target label and reconstructing the original version.

to distinguish real images from those synthesized by the generator.

The total loss terms for training the discriminator ($\mathcal{L}_{D,tot}$) and the generator ($\mathcal{L}_{G,tot}$) are as follows:

$$\mathcal{L}_{D,tot} = \mathcal{L}_{D,src} + \lambda_{D,attr}\mathcal{L}_{D,attr}, \qquad (1)$$

and

$$\begin{aligned}\mathcal{L}_{G,tot} = \ & \mathcal{L}_{G,src} + \lambda_{G,attr}\mathcal{L}_{G,attr} + \\ & \lambda_m\mathcal{L}_{G,m} + \lambda_{rec}\mathcal{L}_{G,rec},\end{aligned} \qquad (2)$$

where, λ coefficients are hyperparameters representing the relative weights for the corresponding loss terms. The individual terms of the total loss for the discriminator ($\mathcal{L}_{D,tot}$) and the generator ($\mathcal{L}_{G,tot}$) are described in the following paragraphs.

For the discriminator, the loss term associated with source discrimination (i.e., discriminating between real and synthesized images) is defined as

$$\begin{aligned}\mathcal{L}_{D,src} = \ & \mathbb{E}_{X,\mathcal{V}_0}\Big[-\log\big(D_{src}(X,\mathcal{V}_0)\big)\Big] + \\ & \mathbb{E}_{X,\mathcal{V}_t}\Big[-\log\big(1 - D_{src}\left(G(X,\mathcal{V}_t),\mathcal{V}_t\right)\big)\Big],\end{aligned} \qquad (3)$$

where, $\mathbb{E}_{X,\mathcal{V}}\big[f(X,\mathcal{V})\big]$ represents the expected value of the random variable $f(X,\mathcal{V})$ taken over distribution of X given the conditional variable $\mathcal{V}$. Similarly, the loss associated with the source discrimination for the generator subnetwork is defined as

$$\mathcal{L}_{G,src} = \mathbb{E}_{X,\mathcal{V}_t}\Big[\log(1 - D_{src}\big(G(X,\mathcal{V}_t),\mathcal{V}_t\big))\Big], \qquad (4)$$

where, $D_{src}(X)$ returns the estimate of the probability that the input image X is real or was synthesized by the generator.

Next, the loss terms for attribute classification are defined as

$$\mathcal{L}_{D,attr} = \mathbb{E}_{X,\mathcal{V}_0}\Big[-\log\big(D_{attr}(\mathcal{V}_0|X)\big)\Big] \qquad (5)$$

and

$$\mathcal{L}_{G,attr} = \mathbb{E}_{X,\mathcal{V}_t}\Big[-\log\big(D_{attr}(\mathcal{V}_t|G(X,\mathcal{V}_t))\big)\Big], \qquad (6)$$

where, $D_{attr}(\mathcal{V}|X)$ is the probability that input image X belongs to attribute class $\mathcal{V}$.

The loss term for optimizing the performance of the biometric face matcher $\mathcal{M}$ on the perturbed images is defined as the squared L_2 distance between the normalized features of the original face image X and those of the synthesized image $G(X,\mathcal{V}_t)$:

$$\mathcal{L}_{G,m} = \mathbb{E}_{X,\mathcal{V}_t}\Big[\|R_\mathcal{M}(X) - R_\mathcal{M}(G(X,\mathcal{V}_t))\|_2^2\Big], \qquad (7)$$

where, $R_\mathcal{M}(X)$ is the normalized face descriptor of face image X after applying a face matcher $\mathcal{M}$.

Lastly, a reconstruction loss term is used to form a cycle-consistent GAN that is able to reconstruct the original face image X from its modified face image $X' = G(X,\mathcal{V}_t)$:

$$\mathcal{L}_{G,rec} = \mathbb{E}_{X,\mathcal{V}_0,\mathcal{V}_t}\Big[\|X - G(G(X,\mathcal{V}_t),\mathcal{V}_0)\|_1\Big]. \qquad (8)$$

Note that the distance term in Eq. 8 is computed as the pixel-wise L_1 norm between the original and modified images, which empirically results in less blurry images compared to employing a L_2 norm as the distance measure [83].

3.3 Neural Network Architecture of PrivacyNet

The composition of the different neural networks used in PrivacyNet, generator G, real vs. synthetic classifier D_{src}, attribute classifier D_{attr}, and face matcher $R_\mathcal{M}$ is described in Fig. 3. The generator and the discriminator architectures were adapted from [84] and [82], respectively.

Generator. The generator G receives as input an RGB face image X of size $224 \times 224 \times 3$ along with the target labels $\mathcal{V}_t$ concatenated as extra channels. The first two convolutional layers, with stride 2, reduce the size of the input image to a to 32×32 with 128 channels. The convolutional layers are followed by instance normalization
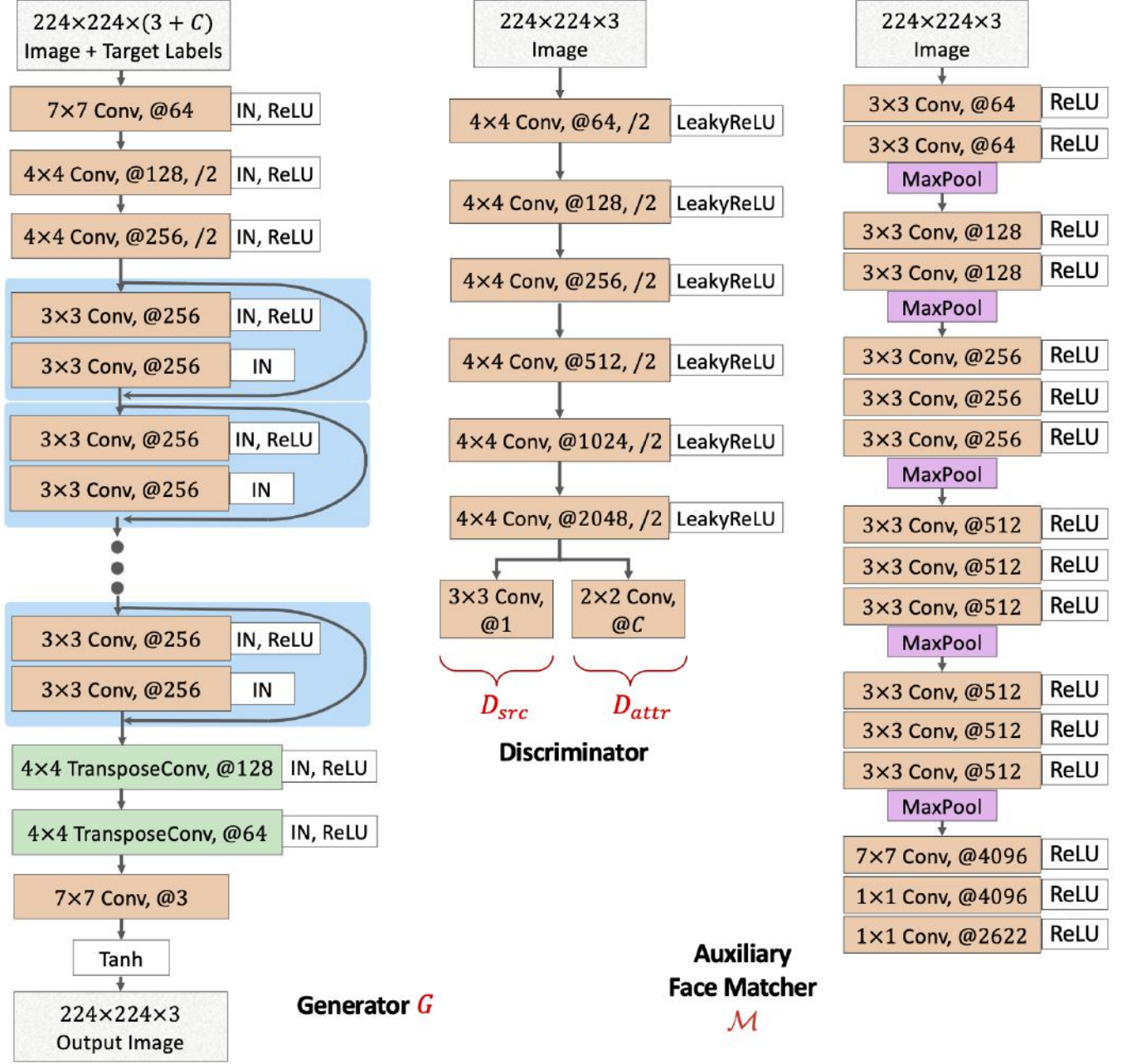
Fig. 3: The detailed neural network architecture of the four sub-networks of PrivacyNet: the generator G, the discriminators D_{src} and D_{attr}, and the pre-trained auxiliary face matcher $\mathcal{M}$. Note that D_{src} and D_{attr} share the same convolutional layers and only differ in their respective output layers.

layers (InstanceNorm) [85]. The layer activations are computed by applying the non-linear ReLU activation function to the InstanceNorm outputs. Then, 6 residual blocks [86] are applied, followed by two transposed convolution for upsampling the image size to 224×224. Finally, the output image X' is constructed by a 1×1 convolution layer and the hyperbolic tangent ($Tanh$) activation function, which returns pixels in the range $(-1, 1)$ (the input image pixels are also scaled to be in range $[-1, 1]$).

Discriminator and Attribute Classifier. The discriminator, as shown in Fig. 3, combines the source discriminator D_{src} and the attribute classifier D_{attr} into one network

where all the layers except the last convolution layer are shared among the two tasks. All the shared convolution layers are followed by a Leaky ReLU non-linear activation with a small negative slope of $\alpha = 0.01$. In the last layer, separate convolutional layers are used for the two tasks, where D_{src} returns a scalar score for computing the loss according to Wasserstein GAN [87], and D_{attr} returns a vector of probabilities for each attribute class.

Face Matcher. Lastly, the auxiliary face matcher is adapted from the publicly available pre-trained VGG-Face CNN model that receives input face images of size $224 \times 224 \times 3$ and computes their face descriptors of size 2622 [88].

3.4 Datasets

We have used five datasets in this study: CelebA [89], MORPH [90], MUCT [91], RaFD [92], and UTK-face [93]. Table 1 shows the number of examples in each dataset, including the number of examples for each face attribute. Since the race label distribution in CelebA is heavily skewed towards Caucasians, and MORPH is heavily skewed towards persons with African ancestry, we combined CelebA and MORPH for training. Both the CelebA and MORPH datasets are split into training and evaluation sets in a subject-disjoint manner. The two training subsets from CelebA and MORPH are merged to train the PrivacyNet model with a relatively balanced race distribution. The other three datasets, MUCT, RaFD, and UTK-face are used only for evaluation. While all five datasets provide provide binary attribute gender labels [5], each dataset lacks the ground-truth labels for at least one of the other attributes, age or race.

Gender Attribute: All the five datasets considered in this study provide ground-truth labels for the gender attribute. Furthermore, since gender is a well-studied topic, there are several face-based gender predictors available for evaluation. In this study, we have considered three gender classifiers for evaluation: a commercial-off-the-shelf software G-COTS, IntraFace [94], and AFFACT [45].

Race Labels: We consider binary labels for race: Caucasians and African descent. Samples that do not belong to these two race groups are omitted from our study since the other race groups are under-represented in our training datasets. We have used the ground-truth labels provided in the MORPH and UTK-face datasets, but for the other three datasets, we labeled the samples in multiple stages. First, an initial estimate of the race attribute is computed using commercial software R-COTS. Next, the predictions made by R-COTS from all samples of the same subject are aggregated, and subjects that show discrepant predictions for different samples are visualized and the discrepant labels are manually corrected. Finally, one random sample from every subject is visually inspected to verify the predicted label. Furthermore, note that since RaFD did not have any sample from the African-descent race group, we did not use this dataset for race prediction analysis.

Age Information: The ground-truth age information is only provided in the MORPH and UTK-face datasets. Therefore, for the remaining datasets (CelebA, MUCT, and RaFD) we used the commercial-off-the-shelf A-COTS software to obtain the class labels of the original images. For the evaluation of our proposed model, we use the Mean Absolute Error (MAE) metric to measure the change in predicted age on the output images of PrivacyNet from the predicted age on the original face images. Therefore, the combination of all five datasets shows both changes in age prediction with respect to the original (for CelebA, MUCT, and RaFD) as well as the ground-truth age values (for MORPH and UTK-face datasets). For training the PrivacyNet model, we create three age groups based on the age values:

$$y_{\text{age}} = \begin{cases} 0 & \text{age} \leq 30; \\ 1 & 30 < \text{age} \leq 45; \\ 2 & 45 < \text{age}. \end{cases} \quad (9)$$

Due to the non-stationary nature of patterns in face aging [43], [52], creating age groups does not fully capture the non-linearity in the textural changes. However, this scheme is consistent with the treatment of the other two attributes, gender and age. Further, it should be emphasized that our objective is *not* to synthesize face images in particular age groups (which is known as age synthesis); instead, the goal of the proposed method is to disturb the performance of arbitrary age predictors.

Identity Information: For matching analysis, we exclude the UTK-face dataset since the subject information is not provided. We used three face matchers, a commercial-off-the-shelf software M-COTS, and two publicly available face matchers DR-GAN [95] and SE-ResNet-50 [96] (SE-Net for short) which were trained on the VGGFace2 dataset [81].

A summary of the datasets and the number of subjects and samples in each dataset is provided in Table 2.

4 EXPERIMENTAL RESULTS

The proposed PrivacyNet model is trained on the joint training subsets of CelebA and MORPH as explained in Section 3.4. Due to the memory-intensive training process, we used a batch-size of 16. The models were trained for $200,000$ iterations. The optimal hyperparameter settings for the weighting coefficients of the attribute loss terms were $\lambda_{attr,d} = 1$ and $\lambda_{attr,g} = 4$. The matching term coefficient was set to $\lambda_m = 4$, and the hyperparameter for the reconstruction term was set to $\lambda_{rec} = 4$. After training the PrivacyNet model, both the discriminator and the auxiliary face matcher subnetworks are discarded and only the generator is used for transforming the unseen face images in the evaluation datasets.

Additionally, we also trained a cycle-GAN model [84], without the auxiliary face matcher, as a baseline to study the effects of the face matcher. The cycle-GAN model is trained using the same protocol that was described for training PrivacyNet. In the remainder of this paper, we will refer to this method as "baseline-GAN". The transformations of five different example images from the CelebA-test dataset are shown in Fig. 4.

The following subsections summarize the results of the experiments and analyze how the performance of the attribute classifiers and face matchers is affected by the face attribute perturbations via PrivacyNet.

4.1 Perturbing Facial Attributes

The performance assessment of the proposed PrivacyNet model involves three objectives:

1) when an attribute is selected to be perturbed, the performance of unseen attribute classifiers must decrease;

2) the attribute classifiers should retain their performance on attributes that are not selected for perturbation;

---

5. In this paper we treat gender as a binary attribute with two labels, male and female; however, it must be noted that societal and personal interpretation of gender can result in many more classes.

Fig. 4: Five example face images from the CelebA dataset along with their transformed versions using PrivacyNet and baseline-GAN models. The rows are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is specified as A0 (young), A1 (middle-aged), or A2 (old).

TABLE 1: Overview of datasets used in this study, with the number of face images corresponding to each attribute. Samples which belong to a race other than the two categories shown below, as well as those whose age-group could not be determined, are omitted.

Dataset	Gender		Race		Age groups		
	Male	Female	African-descent	Caucasian	Young	Middle-aged	Old
CelebA	84,434	118,165	11,119	142,225	79,848	91,373	16,337
MORPH	47,057	8,551	42,897	10,736	25,009	26,614	3,985
MUCT	1,844	1,910	1,030	1,480	1,326	1,807	620
RaFD	1,008	600	0	1,608	1,276	332	0
UTK-face	12,582	11,522	4,558	10,222	12,980	6,068	5,056

TABLE 2: Summary of the datasets used in this study, with the number of subjects and samples in the train-test partitions. The "Excluded Experiments" column indicate datasets that were removed from an experiment for the reasons given in the text.

Datasets	Train		Test		Excluded
	# Subj	# Samples	# Subj	# Samples	Experiments
CelebA	8,604	150,530	167	2,795	–
MORPH	11,176	45,512	1,968	8,038	–
MUCT	–	–	185	2,508	–
RaFD	–	–	67	1,608	Race
UTK-face	–	–	NA	14,182	Matching

3) in all cases, the performance of unseen face matchers must not be drastically affected.

We conducted several experiments to assess whether the proposed PrivacyNet model meets these objectives.

Gender Classification Performance: We considered three gender classifiers: a commercial-off-the-shelf software (G-COTS), AFFACT [45] and IntraFace [94]. For this comparison study, all five evaluation datasets listed in Table 2 were considered. The performances of the different gender classifiers on the original and perturbed images are measured using the Equal Error Rate (EER); the results are shown in Fig. 5. For a given image, PrivacyNet can produce up to 15 distinct outputs, depending on the combination of attributes that are selected for perturbation.

The EER results shown in Fig. 5 indicate that PrivacyNet increases the error rate of the cases where the gender attribute is willfully perturbed, which is desired. At the same time, it can preserve the performance of gender classifiers when gender information is to be retained. The EER of gender classification using G-COTS software on gender-perturbed outputs increases to 20-40%, and the EER of gender classification using AFFACT and IntraFace on these outputs surpasses 60%. Comparisons between the gender prediction results on the outputs of PrivacyNet and the outputs of the face-mixing approach by Othman and Ross [34], as well as the model by Sim and Zhang [35], show that in case of G-COTS, the PrivacyNet results are superior in terms of increasing the EER (Fig. 5).

Note that we did not include the results of the GAN model in Fig. 5 for readability sake. However, we observed that the GAN model shows larger deviations (which is advantageous) in cases where gender was intended to be perturbed. This is expected since the GAN model does not have the constraints from the auxiliary face matcher. Therefore, there is more flexibility for modifying the patterns of the face. However, a disadvantage of the GAN model is that

it also significantly degrades the matching utility as shown in Section 4.2.

Race Prediction Performance: We conducted the race prediction analysis using a commercial-off-the-shelf software, R-COTS. Similar to the gender classification experiments, we show the EER of race classification on original images as well as the different outputs of the PrivacyNet model in Fig. 6. Since the face mixing approach proposed in [34] was only formulated for gender and not race perturbations, we did not include it in this section.

The EER results in Fig. 6 show that PrivacyNet successfully meets the objectives of our study for confounding race predictors. The outputs where race is not intended to be perturbed (shown in blue) exhibit low EER values similar to the EER obtained from the original images (EER $\sim 1\%$). On the other hand, when race is selected to be perturbed, the EER values increase significantly (EER $\sim 20\%$ for CelebA and UTK-face, and EER $\sim 10\%$ for MORPH and MUCT datasets). The results of separately perturbing gender and race using the controllable face privacy method proposed in [35] are also shown for comparison. When the race attribute is perturbed according to [35], the performance is slightly higher than our model. However, the disadvantage of the controllable face privacy method [35] is that when it perturbs the gender attribute, it also affects the race predictions.

Age Prediction Performance: To assess the ability of PrivacyNet for confounding age information, we used a commercial-off-the-shelf age predictor (A-COTS), which has shown remarkable performance across the different datasets tested in this study (Fig. 7). We used the Mean Absolute Error (MAE) values in unit of years to measure the change in age prediction before and after perturbing the images (Fig. 7). As mentioned previously (Section 3.4), the ground-truth age values for three datasets – CelebA, MUCT, and RaFD – are not provided. Therefore, for these three datasets, the MAE values are computed as the difference between the age predictions on the output images and the predictions on the original images, while for the other two datasets, MORPH and UTK-face, the ground-truth values are used for computing the MAE values.

The results of age-prediction show that the MAE obtained from the outputs, where age is not meant to be perturbed, remains at approximately 5 years. However, when we intend to modify the age of face images, using the label A2 results in the highest MAE (around 20 years for RaFD and 15 years for the other four datasets) compared to A0 and A1. A possible explanation for this observation is that, due
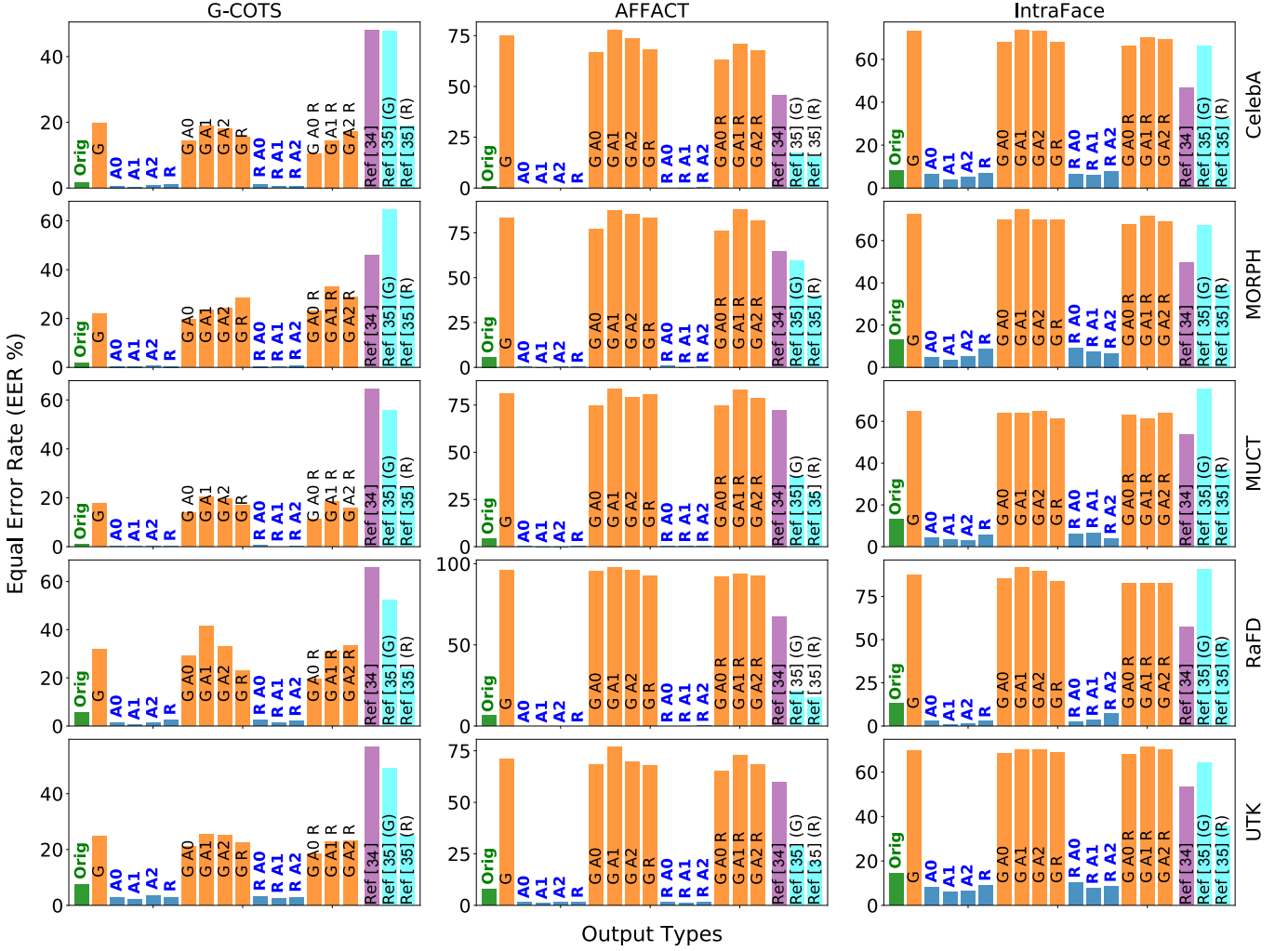
Fig. 5: Performance of three gender classifiers – G-COTS, AFFACT, and IntraFace – on original images as well as different outputs of the proposed model (the larger the difference the better). The results of a face mixing approach, as described in [34], are also shown. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is abbreviated as A0 (young), A1 (middle-aged), and A2 (old). The outputs of PrivacyNet, where the gender attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.

the nature of the aging process, larger textural changes occur in face images belonging to A2. The MAE of the A0 group is also relatively large (except for RaFD), which may be caused by the reversal of the textural changes. However, the results of the middle-age group (A1) is similar to the cases where we did not intend to modify the age. We hypothesize that the small changes in A1 are also due to the non-stationary aspect of aging patterns; the age perturbations via the PrivacyNet model can potentially be improved by using an ordinal regression approach for age prediction.

4.2 Retaining the Matching Utility of Face Images

Besides obfuscating soft-biometric attributes in face images, another objective of this work is to retain the recognition utility of all outputs of PrivacyNet. For this purpose, we conducted matching experiments using three unseen face matchers: commercial-off-the-shelf software (M-COTS) and two publicly available matchers, SE-ResNet-50 trained on the VGGFace2 dataset [81] (SE-Net for short), and DR-

GAN [95]. Fig. 8 shows the ROC curves obtained from these matching experiments for four datasets – CelebA, MORPH, MUCT, and RaFD. The UTK-face dataset is removed from this analysis since it does not contain subject information. Since PrivacyNet generated 15 outputs for each input face image, the minimum and maximum True Match Rate (TMR) values at each False Match Rate (FMR) value are computed and only the range of values for these 15 outputs are shown. Note that it is expected for the matching utility to be retained in all these 15 outputs. Similarly, the range of TMR values at each FMR obtained from the 15 different outputs of the GAN model that did not have the auxiliary face matcher for training, is also shown for comparison. The ROC curves of PrivacyNet are very close to the ones obtained from the original images for each dataset, compared to the baseline results, which both show significantly larger deviations. It is worth noting that the baseline-GAN is equivalent to removing the matching loss term $\mathcal{L}_{G,m}$ from PrivacyNet. As shown in Figures 4,8 and
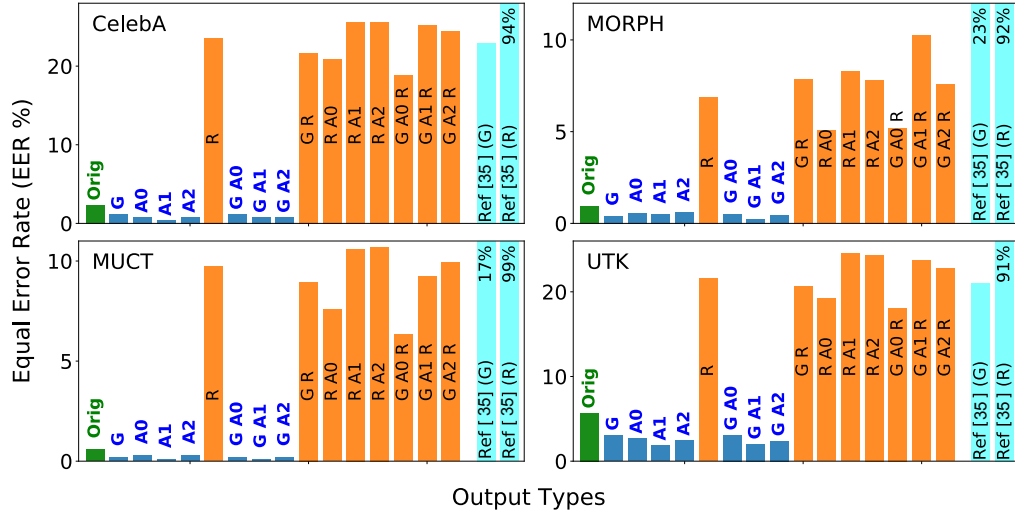
Fig. 6: Performance of the race classifier, R-COTS, on original images as well as different outputs of the proposed model. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1, and A2 (the larger the difference the better). The outputs of PrivacyNet, where the race attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.
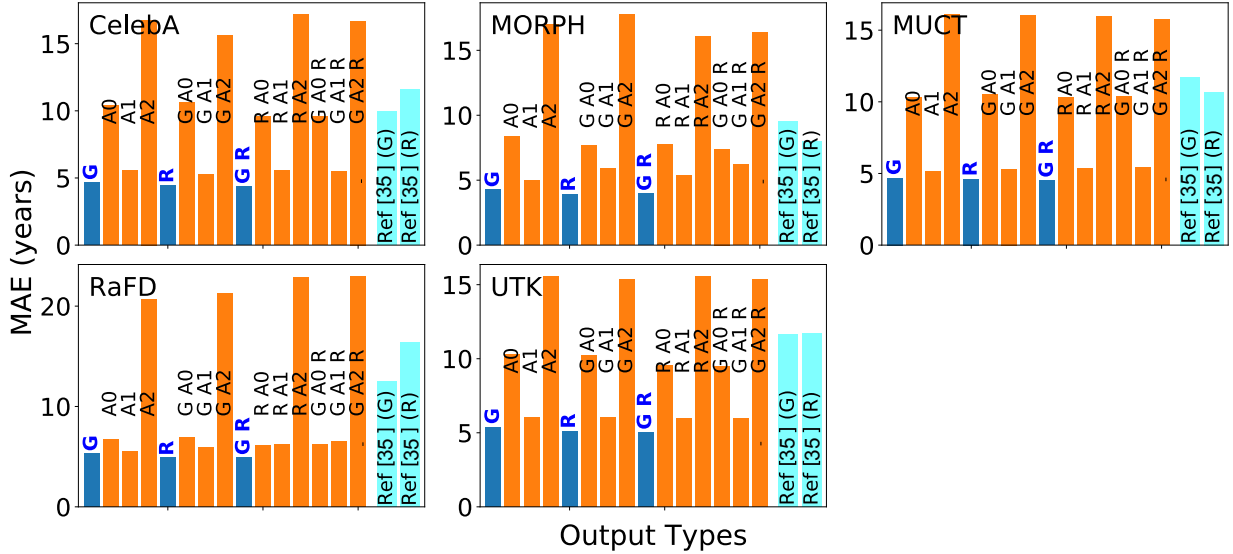

Fig. 7: Change in age prediction of A-COTS on different outputs of the proposed model. This is with respect to the age predicted on original images for CelebA, MUCT and RaFD, and the ground-truth age values for MORPH and UTK-face. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1 and A2. The outputs of PrivacyNet, where the age attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.

9-("Baseline-GAN"), the PrivacyNet model produces more realistic-looking faces images without the matching loss term. However, removing the matching loss term results in a severe decline in matching performance, affecting both the true matching rate and identification accuracy (Figs. 8 and 9). The coefficient λ can be further tuned to control the trade-off between the performance of face-matching and obfuscating the soft-biometric attributes.

In addition to the ROC curves, we have also plotted the Cumulative Match Characteristics (CMC) [97], as shown in Fig. 9. According to the CMC curves, the results of PrivacyNet match very closely with the CMC curves obtained from the original images in all cases, which shows that PrivacyNet retains matching utility.

It is worth noting that Ref. [35] has more favorable CMC curves than the other methods evaluated in this study. A plausible explanation is that Ref. [35] aligns and normalizes its inputs to a reference face image, which significantly reduces the intra-class variations. This reduction of intra-class variations increases the number of true positives. However, it also increases the number of false positives, thereby deteriorating the *ROC* performance. One may argue that the difference in performance could be due to the different training datasets that were used to train our model and that

of Ref. [35], and, perhaps, re-training Ref. [35] would be necessary for a fair comparison. However, we note that we used the original model for Ref. [35], which was constructed from a carefully curated dataset, and the original authors of [35] recommended against retraining.

5 CONCLUSION

In this work we designed PrivacyNet, which is a deep neural network model for imparting multi-attribute privacy to face images including age, gender, and race attributes. PrivacyNet utilizes a Semi-Adversarial Network (SAN) module combined with Generative Adversarial Networks (GANs) to perturb an input face image, where certain attributes are obfuscated selectively, while other face attributes are preserved. Most importantly, the matching utility of face images from these transformations is preserved. Experimental results using three unseen face matchers as well as three unseen attribute classifiers show the efficacy of our proposed model in perturbing such attributes, while the matching utility of face images is not adversely impacted.

Although generating visually realistic images was not the primary objective of this work, we note that the modified images from the proposed model may have some artifacts. As a result, a human observer might be able to distinguish between perturbed face images and non-modified ones. We intend to study the effect of perturbations on human observers in future work and plan to design solutions capable of creating more realistic-looking face images while satisfying the objective of this work, viz., maintaining good face matching performance.

REFERENCES

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

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

[3] D. S. Trigueros, L. Meng, and M. Hartnett, "Face recognition: From traditional to deep learning methods," *arXiv preprint arXiv:1811.00116*, 2018.

[4] G. L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," in *Proceedings of 12th International Conference on Image Analysis and Processing.*, 2003, pp. 115–120.

[5] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1925–1937, 2011.

[6] M. Grgic, K. Delac, and S. Grgic, "SCface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[7] H. Qezavati, B. Majidi, and M. T. Manzuri, "Partially covered face detection in presence of headscarf for surveillance applications," in *4th International Conference on Pattern Recognition and Image Analysis*, 2019, pp. 195–199.

[8] T. Kwon and H. Moon, "Biometric authentication for border control applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 8, pp. 1091–1096, 2008.

[9] J. S. del Rio, D. Moctezuma, C. Conde, I. M. de Diego, and E. Cabello, "Automated border control e-gates and facial recognition systems," *Computers & Security*, vol. 62, pp. 49–72, 2016.

[10] A. K. Bobak, A. J. Dowsett, and S. Bate, "Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills," *PloS one*, vol. 11, no. 2, p. e0148148, 2016.

[11] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[12] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[13] J. H. Chen and H. Huang, "Face recognition using AAM and global shape features," in *IEEE International Conference on Robotics and Biomimetics*, 2008, pp. 824–827.

[14] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[15] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] 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.

[17] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial soft biometrics for recognition in the wild: Recent works, annotation and cots evaluation," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 7, 2018.

[18] M. Nixon, P. Correia, K. Nasrollahi, T. Moeslund, A. Hadid, and M. Tistarelli, "On soft biometrics," *Pattern Recognition Letters*, vol. 68, no. 2, pp. 218–230, December 2015.

[19] 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*, vol. 40, no. 02, pp. 332–351, 2018.

[20] D. A. Reid, M. S. Nixon, and S. V. Stevenage, "Soft biometrics; human identification using comparative descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1216–1228, 2014.

[21] D. Martinho-Corbishley, M. S. Nixon, and J. N. Carter, "On categorising gender in surveillance imagery," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–6.

[22] E. S. Jaha, "Augmenting gabor-based face recognition with global soft biometrics," in *7th IEEE International Symposium on Digital Forensics and Security (ISDFS)*, 2019, pp. 1–5.

[23] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification," *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 739–777, 2011.

[24] C. Dwork and V. Feldman, "Privacy-preserving prediction," in *Conference on Learning Theory*, S. Bubeck, V. Perchet, and P. Rigollet, Eds., vol. 75, 2018, pp. 1693–1702.

[25] C. Dwork, "Differential privacy and the US census," in *38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, D. Suciu, S. Skritek, and C. Koch, Eds., 2019, p. 1.

[26] C. Dwork, C. Ilvento, and M. Jagadeesan, "Individual fairness in pipelines," in *1st Symposium on Foundations of Responsible Computing*, A. Roth, Ed., vol. 156, 2020, pp. 7:1–7:22.

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

[28] 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.

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

[30] B. Sadeghi and V. N. Boddeti, "Imparting fairness to pre-trained biased representations," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.
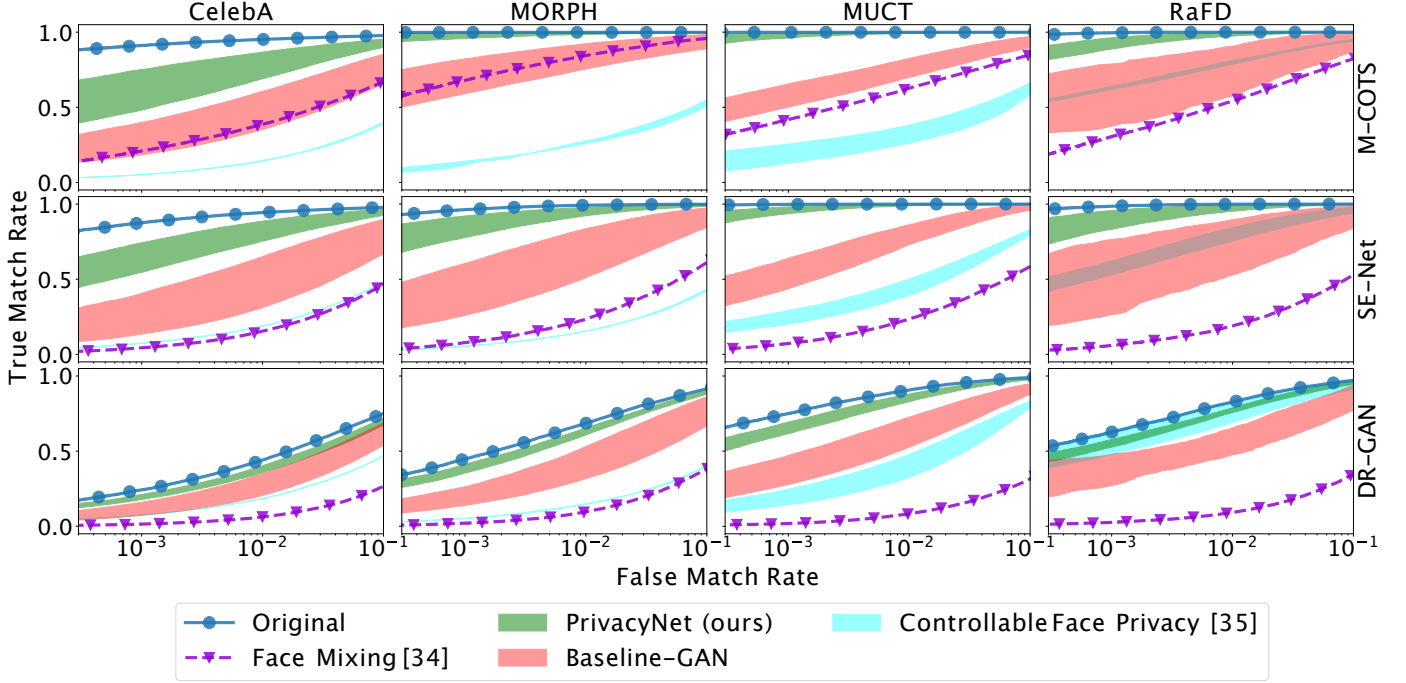
Fig. 8: ROC curves showing the performance of unseen face matchers on the original images for PrivacyNet, the baseline-GAN model, face mixing [34] approach and the controllable face privacy [35] method. The results show that ROC curves of PrivacyNet have the smallest deviation from the ROC curve of original images suggesting that the performance of face matching is minimally impacted, which is desired.
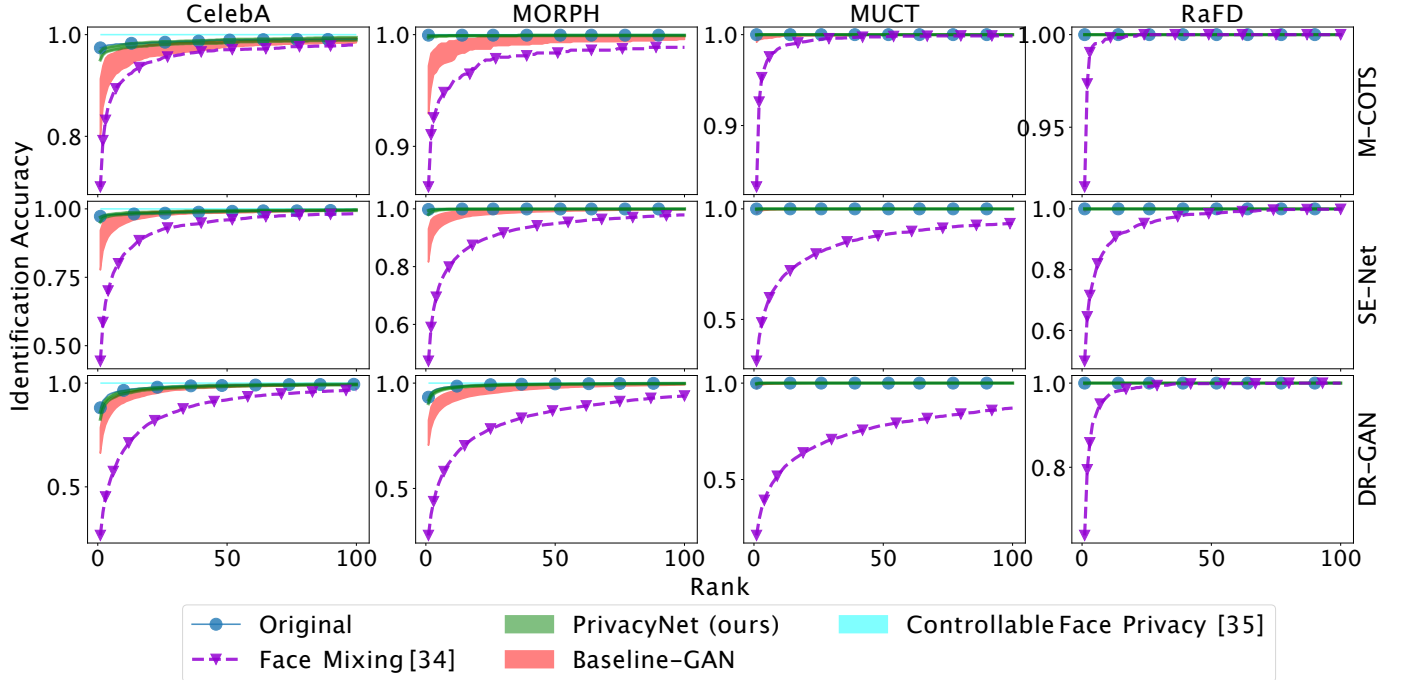

Fig. 9: CMC curves showing the identification accuracy of unseen face matchers on the original images. Also shown is the range of CMC curves for the PrivacyNet model and the baseline-GAN model, along with that of the face mixing [34] and controllable face privacy [35] approaches. It must be noted that in cases where the results of PrivacyNet or GAN are not visible, the curves overlapped with the CMC curve of the original images: this means that there was no change in matching performance at all (which is the optimal case). The results confirm that transformations made by PrivacyNet preserve the matching utility of face images.

[31] B. Medcn, P. Peer, and V. Struc, "Selective face deidentification with end-to-end perceptual loss learning," in *2018 IEEE International Work Conference on Bioinspired Intelligence*, 2018, pp. 1–7.

[32] 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*, 2018, pp. 640–656.

[33] C. Jasserand-Breeman, "Reprocessing of biometric data for law enforcement purposes: Individuals' safeguards caught at the interface between the GDPR and the 'police' directive?" Ph.D. dissertation, 2019.

[34] 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.

[35] 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.

[36] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-Adversarial Networks: Convolutional autoencoders for imparting privacy to face images," in *11th IAPR International Conference on Biometrics*, 2018.

[37] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing k-facial attributes via adversarial perturbations," *27th International Joint Conference on Artificial Intelligence*, 2018.

[38] J. Suo, L. Lin, S. Shan, X. Chen, and W. Gao, "High-resolution face fusion for gender conversion," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 2, pp. 226–237, 2011.

[39] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 495–505.

[40] V. Mirjalili, S. Raschka, and A. Ross, "FlowSAN: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, pp. 1–1, 2019.

[41] V. Mirjalili, S. Raschka, and A. Ross, "Gender Privacy: An ensemble of Semi Adversarial Networks for confounding arbitrary gender classifiers," in *9th IEEE International Conference on Biometrics: Theory, Applications and Systems*, 2018.

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

[43] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4920–4928.

[44] W. Cao, V. Mirjalili, and S. Raschka, "Rank-consistent ordinal regression for neural networks," *arXiv preprint arXiv:1901.07884*, 2019.

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

[46] 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*, 2019, pp. 116–124.

[47] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

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

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

[50] M. Castrillón-Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Descriptors and regions of interest fusion for in-and cross-database gender classification in the wild," *Image and Vision Computing*, vol. 57, pp. 15–24, 2017.

[51] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *IEEE Conference on Biometrics Theory, Applications and Systems*, 2016, pp. 1–8.

[52] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using Ranking-CNN for age estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5183–5192.

[53] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," in *Computer Vision and Pattern Recognition Workshop*, 2006.

[54] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *2nd International Conference on Learning Representations*, 2014.

[55] W. Oleszkiewicz, P. Kairouz, K. Piczak, R. Rajagopal, and T. Trzciński, "Siamese generative adversarial privatizer for biometric data," in *Computer Vision*, 2019, pp. 482–497.

[56] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[57] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations*, 2015.

[58] O. Poursaeed, I. Katsman, B. Gao, and S. Belongie, "Generative adversarial perturbations," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4422–4431.

[59] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.

[60] Y. Su, G. Sun, W. Fan, X. Lu, and Z. Liu, "Cleaning adversarial perturbations via residual generative network for face verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2597–2601.

[61] S. Taheri, M. Salem, and J.-S. Yuan, "RazorNet: Adversarial training and noise training on a deep neural network fooled by a shallow neural network," *Big Data and Cognitive Computing*, vol. 3, no. 3, 2019.

[62] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 284–293.

[63] A. Graese, A. Rozsa, and T. E. Boult, "Assessing threat of adversarial examples on deep neural networks," in *15th IEEE International Conference on Machine Learning and Applications*, 2016, pp. 69–74.

[64] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yadav, "Some research problems in biometrics: The future beckons," *12th IAPR International Conference on Biometrics*, 2019.

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

[66] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig, "Controllable invariance through adversarial feature learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 585–596.

[67] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations," *Applied Intelligence*, pp. 1–18, 2019.

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

[69] P. C. Roy and V. N. Boddeti, "Mitigating information leakage in image representations: A maximum entropy approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2586–2594.

[70] B. Sadeghi, R. Yu, and V. Boddeti, "On the global optima of kernelized adversarial representation learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7971–7979.

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

[72] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014, pp. 2672–2680.

[73] Z. Lu, Z. Li, J. Cao, R. He, and Z. Sun, "Recent progress of face image synthesis," in *4th IAPR Asian Conference on Pattern Recognition*, Nov 2017, pp. 7–12.

[74] S. Raschka and V. Mirjalili, *Python Machine Learning, 3rd Ed.* Birmingham, UK: Packt Publishing, 2019.

[75] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[76] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *European Conference on Computer Vision*, 2018, pp. 417–432.

[77] X. Song, M. Shao, W. Zuo, and C. Li, "Face attribute editing based on generative adversarial networks," *Signal, Image and Video Processing*, pp. 1–9, 2020.

[78] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4030–4038.

[79] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He, "GeneGAN: Learning object transfiguration and attribute subspace from unpaired data," *British Machine Vision Conference*, 2017.

[80] M. Li, W. Zuo, and D. Zhang, "Deep identity-aware transfer of facial attributes," *arXiv preprint arXiv:1610.05586*, 2016.

[81] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VG-GFace2: A dataset for recognising faces across pose and age," in *International Conference on Automatic Face and Gesture Recognition*, 2018.

[82] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*, Oct 2017, pp. 2242–2251.

[83] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017, pp. 5967–5976.

[84] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[85] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[86] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[87] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *34th International Conference on Machine Learning*, 2017.

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

[89] Z. Liu *et al.*, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[90] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Automatic Face and Gesture Recognition*, 2006, pp. 341–345.

[91] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," *Pattern Recognition Association of South Africa*, 2010.

[92] 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.

[93] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

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

[96] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[97] B. DeCann and A. Ross, "Relating ROC and CMC curves via the biometric menagerie," in *Biometrics: Theory, Applications and Systems*, 2013, pp. 1–8.