Learning privacy-enhancing face representations through feature disentanglement

Blaž Bortolato[1], Marija Ivanovska[1], Peter Rot[1,2], Janez Križaj[1], Philipp Terhörst[3], Naser Damer[3], Peter Peer[2], Vitomir Štruc[1]

[1] Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana, Slovenia

[2] Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, Ljubljana, Slovenia

[3] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany

***Abstract*— Convolutional Neural Networks (CNNs) are today the de-facto standard for extracting compact and discriminative face representations (templates) from images in automatic face recognition systems. Due to the characteristics of CNN models, the generated representations typically encode a multitude of information ranging from identity to soft-biometric attributes, such as age, gender or ethnicity. However, since these representations were computed for the purpose of identity recognition only, the soft-biometric information contained in the templates represents a serious privacy risk. To mitigate this problem, we present in this paper a privacy-enhancing approach capable of suppressing potentially sensitive soft-biometric information in face representations without significantly compromising identity information. Specifically, we introduce a Privacy-Enhancing Face-Representation learning Network (PFRNet) that disentangles identity from attribute information in face representations and consequently allows to efficiently suppress soft-biometrics in face templates. We demonstrate the feasibility of PFRNet on the problem of gender suppression and show through rigorous experiments on the CelebA, Labeled Faces in the Wild (LFW) and Adience datasets that the proposed disentanglement-based approach is highly effective and improves significantly on the existing state-of-the-art.**

I. Introduction

Contemporary face recognition systems rely predominantly on convolutional neural networks (CNNs) to extract face representations that can be used for identity recognition. Despite the fact that the CNNs used in such systems are trained to generate representations that only encode identity, recent research has shown that information about soft-biometrics attributes, such as age, gender or ethnicity, is also encoded in these representations [3], [4], [21], [23]. Because this information can be extracted from the computed representations and may potentially be misused (for user profiling, discrimination, etc.) it represents a considerable privacy risk. Researchers are, therefore, increasingly looking into ways of suppressing soft-biometric information in face representations computed by modern CNN-based models.

The task of suppressing specific facial attributes in face representations (or templates) is often referred to as soft-biometric privacy-enhancement. The main challenge when designing such techniques is to suppress as much of the soft-biometric information as possible without compromising identity cues needed for recognition. Because the information about soft-biometric attributes and identity is highly entangled in CNN-based face representations trying to suppress one inevitably affects the other.

B. Bortolato, M. Ivanovska and P. Rot are first authors with equal contributions.

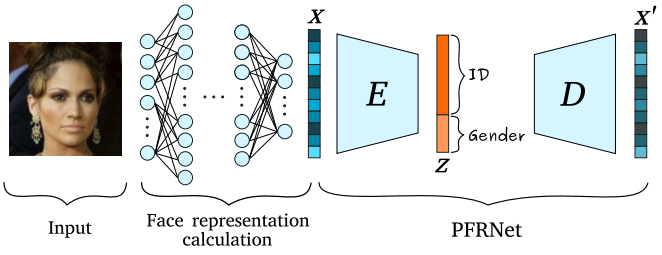


Fig. 1. We address the problem of soft-biometric privacy-enhancement and present PFRNet - a neural network capable of suppressing soft-biometric attributes, such as gender, in face representations without compromising identity cues. PFRNet disentangles the input face representations in such a way that attributes and identity are encoded separately in the latent space.

Due to this entanglement existing research in the area of soft-biometric privacy enhancement focused mostly on variable elimination techniques and feature-space transformations that have a greater impact on soft-biometric attributes than on identity, e.g., [28], [29]. While these techniques were shown to enhance the level of soft-biometric privacy in face representations to some extent, they still do not address the central problem, that is, that the information in the face representation is highly entangled.

In this paper we try to address this gap and propose a Privacy-enhancing Face-Representation learning Network (PFRNet) that approaches the task of suppressing soft-biometric attributes in face representations by disentangling face representations so that distinct variables encode attribute and identity information, as illustrated in Fig. 1. Thus, different from existing techniques in this area, we do not try to modify face representations to improve the identity-to-attribute information ratio, but propose a more general approach that separates soft-biometric attributes from identity. We design PFRNet as an autoencoder and devise loss functions that force different parts of the latent representations of PFRNet to encode only specific characteristics of the input representations (attributes or identity). We evaluate PFRNet in rigorous experiments on the CelebA, Lebeled Face in the Wild (LFW) and Adience datasets and show that it is capable to effectively suppress soft-biometric attributes (on the example of gender) in face representations, while still ensuring highly competitive face recognition performance.

The model generalizes well across different datasets and convincingly outperforms the existing state-of-the-art.

Our key contributions in this work are:

- We introduce PFRNet, a neural-network-based model for soft-biometric privacy-enhancement of face representations (templates) that sets a new state-of-the-art on multiple face datasets in terms of attribute suppression. PFRNet is applicable to face representations of arbitrary CNNs and is not limited to a specific face model.
- We propose a learning objective that efficiently disentangles the face representations and results in distinct encodings for identity and gender in the latent space.
- We show that feature disentanglement is an effective way for soft-biometric privacy-enhancement of face representations that mitigates many of the issues encountered with prior techniques proposed in this area.

II. Related work

Existing techniques for soft-biometric privacy enhancement in face recognition systems can broadly be divided into one of two groups: *i)* techniques that aim to conceal soft-biometric attributes at the image-level [20], [17], and *ii)* techniques that try to suppress attribute information at the face representation (or template) level [28], [29]. In this section we present a brief overview of the most competitive techniques from the two groups. More information on the general topic of biometric privacy can be found in recent literature, such as [25], [6], [11], [16], [8], [15], [22], [9].

Image-level techniques. Most of the recent research on soft-biometric privacy at the image level tries to suppress attribute information in images so that machine-learning models, such as CNNs, are unable to automatically infer facial attributes, while human observers perceive only minimal (or no) changes in the appearance of the modified images [2], [26], [17], [20], [19]. Mirjalili *et al.* [17], for example, present one such approach that utilizes a so-called Semi-Adversarial neural Network (SAN), an autoencoder that maps the input face image into a visually similar, but perturbed image capable of fooling a predefined gender classifier. In their follow-up work [19], the authors extend this approach to arbitrary gender classifiers. Techniques from [2], [26] approach the problem of soft-biometric privacy with adversarial perturbations. These methods typically add adversarial noise to images, such that predefined attribute classifiers are forced to make classification mistakes.

A common characteristic of image-level techniques is that they rely on pretrained attribute classifiers to learn the image perturbation required to ensure soft-biometric privacy. Whether these methods generalize to arbitrary classifiers (e.g., not CNN-based) and classification models trained on the perturbed images is still unclear from the existing literature. Our approach, on the other hand, though belonging to the group of template-level techniques, is able to suppress gender attribute information regardless of the classifier used and even in cases when the suppressed data is used as the basis for training an attribute classification model.

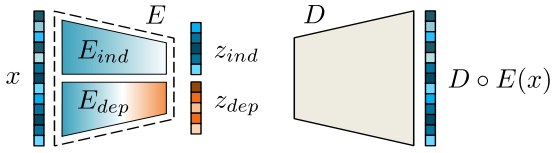


Fig. 2. Illustration of the architecture of PFRNet. The model maps the initial face representations x into disentangled latent representations z_{ind} and z_{dep}, where the first encodes identity information and the second encodes information about soft-biometric attributes.

Template-level techniques. Different from the methods discussed above, template-level techniques try to suppress attribute information in the face representations (or templates) and not the input face images. Because the information in these representations is highly compressed (compared to the original input images) and, therefore, somewhat easier manipulated, existing techniques from this group were shown to generalize well to arbitrary attribute classifiers. Terhörst *et al.* [28], for example, proposed an Incremental Variable Elimination (IVE) algorithm to suppress information about age and gender in face representations. The algorithm is based on a decision tree ensemble that scores each variable in the face representation with respect to its importance for a specific recognition task. Variables most affecting attribute classification are then excluded from the representation. Another technique [29], called Cosine-Sensitive Noise (CSN) transformation tried to ensure soft-biometric privacy by adding a specific type of noise to the face representations, such that the attribute information was masked more than identity. Different from the presented techniques, our approach does not attempt to mask or alter attribute information to have a better ratio between soft-biometric cues and identity. Instead it tries to disentangle the face representations, so that attribute information can be removed. As we show in Section IV, such an approach leads to highly competitive performance.

III. Methodology

In this section we present our privacy-enhancing approach for suppression of selected soft-biometric attributes in biometric templates. The approach is based on a novel Privacy-enhancing Face-Representation learning Network (PFRNet).

A. Overview of PFRNet

We design PFRNet as an autoencoder that consists of a two-path encoder E and a single-path decoder D, as shown in Fig. 2. E comprises two separate encoders E_{ind} and E_{dep}. The first E_{ind} maps the face representation x (commonly generated by a CNN face recognition model, such as FaceNet [27], VGGFace [24] or VGGFace2 [1]) into a latent vector z_{ind}, which preserves identity cues, while greatly reducing the amount of information related to selected soft-biometric attributes, such as gender. Thus, the latent representations z_{ind} can ideally be used in biometric systems for identity recognition without privacy-related concerns regarding the misuse of soft-biometric information. The second encoder E_{dep} maps the original face representation x into a latent vector z_{dep} that encodes soft-biometric attributes

only. The complete latent representation of x generated by the encoder E is a concatenation of the latent representations z_{ind} and z_{dep}, i.e., $z = z_{ind} \oplus z_{dep}$. This latent representation can in principle be reconstructed by the decoder function $D : z \rightarrow x$. In this paper, we train PFRNet to suppress gender information, but in general the same concept can also be extended to other soft-biometric attributes.

The properties of z_{ind} and z_{dep} discussed above follow from the learning objective devised for PFRNet. Specifically, the model makes use of three different loss functions. The first ensures a good reconstruction of the input data, the second suppresses gender information in z_{ind} and the third loss forces the distributions of z_{dep} for male and female subjects to be as different as possible. The second and third loss functions require attribute labels for the training data, while the first relies on self-supervision.

B. Autoencoder

The basic property of a generic autoencoder is that it encodes the input data x into a latent representation z and then reconstructs the input data x from z, or formally: $D(E(x)) = x$. Both the encoder E and the decoder D are commonly implemented with neural networks and the parameters of the networks are learned by minimizing a reconstruction loss, such as

$$\mathcal{L}_0 = ||x - D \circ E(x)||_{L_2}, \tag{1}$$

over some training data. Here, the operator $||.||_{L_2}$ denotes the L_2 norm. The relation $D \circ E(x) = x$ implies that all information in x is also stored in the latent representation $z = E(x)$. Thus, the latent representation z can be viewed as a reparametrization of x. Since PFRNet is designed as an autoencoder, we use the above reconstruction objective as one of the loss terms used in our training procedure.

C. Removing gender information

A straight-forward way to remove gender information from z_{ind} is to require that the distribution $Q(z_{ind}, f)$ of the latent vectors z_{ind} of female (f) subjects is as similar as possible to the distribution $Q(z_{ind}, m)$ of the latent vectors z_{ind} of male (m) subjects. We incorporate this requirement in the learning procedure of PFRNet through a sampling based approach. During training we first sample batches of n_g data points from each of the distributions $Q(z_{ind}, f)$ and $Q(z_{ind}, m)$ and then compute α-order moments from the sampled data, i.e.:

$$\langle z_{ind}^{\alpha} \rangle_g \equiv \frac{1}{n_g} \sum_{j=1}^{n_g} (z_{ind,j})^{\circ \alpha}, \tag{2}$$

where $^{\circ}$ stands for the Hadamard (or element-wise) power operation and $g \in \{m, f\}$. Here, the value of α defines the moment type, $\alpha \in \{0, 1, 2, \ldots, \alpha_{max}\}$.

In order to align the latent distributions of male and female subjects generated by PFRNet, we define the following loss:

$$\mathcal{L}_{\alpha} = ||\langle z_{ind}^{\alpha} \rangle_f - \langle z_{ind}^{\alpha} \rangle_m||_{L_2}, \tag{3}$$

which we aim to minimize during training. The loss is minimized when the corresponding moments of both male and female distributions are equal: $\langle z_{ind}^{\alpha} \rangle_m = \langle z_{ind}^{\alpha} \rangle_f$.

D. Gender discrimination

Because the latent representation z_{ind} is gender-free in the ideal case, all of the gender information from x needs be encoded in the latent representation z_{dep} to enable the decoder D to reconstruct the input data. To ensure that the information pertaining to gender is stored in z_{dep} and suppressed in z_{ind}, we introduce a third loss function that explicitly forces the distributions $Q(z_{dep}, f)$ and $Q(z_{dep}, m)$ to be as different as possible.

Similarly as described in the previous section, we again first sample batches of n_g data points from the distributions $Q(z_{dep}, f)$ and $Q(z_{dep}, m)$ and then compute β-order moments from the sampled data, for $\beta \in \{0, 1, \ldots, \beta_{max}\}$, i.e.:

$$\langle z_{dep}^{\beta} \rangle_g \equiv \frac{1}{n_g} \sum_{j=1}^{n_g} (z_{dep,j})^{\circ \beta}, \tag{4}$$

where $^{\circ}$ again denotes the Hadamard power operation and $g \in \{m, f\}$. To push the male and female latent distributions $Q(z_{dep}, f)$ and $Q(z_{dep}, m)$ apart, we formulate the following Gaussian shaped loss:

$$\mathcal{L}_{\beta} = \exp\left\{ -\frac{|\langle z_{dep}^{\beta} \rangle_f - \langle z_{dep}^{\beta} \rangle_m|^2}{2\ \sigma_{\beta}^2} \right\}. \tag{5}$$

The loss function has a free parameter σ_{β} that defines the shape of the Gaussian. Note that the loss is minimized when the difference between the corresponding moments of the male and female distributions is as large as possible.

E. Training objective

Combining the loss functions from Eq. (1), (3) and (5) we obtain the overall learning objective of PFRNet, which takes the following form

$$\mathcal{L} = \mathcal{L}_0 + \sum_{\alpha=1}^{\alpha_{max}} \lambda_{\alpha} \mathcal{L}_{\alpha} + \sum_{\beta=1}^{\beta_{max}} \lambda_{\beta} \mathcal{L}_{\beta}. \tag{6}$$

λ_{α} and λ_{β} represent Lagrange multipliers that balance the impact of the individual loss terms. To learn the parameters of PFRNet, we minimize the loss on suitable training data.

IV. Experiments

A. Datasets and experimental setup

Datasets. Three publicly available face datasets are used to assess the performance of PFRNet: CelebA [13], Labeled Faces in the Wild (LFW) [7] and Adience [5]. Images from these datasets were captured in unconstrained settings and exhibit significant variability across various factors (e.g., pose, age, gender, facial expression, image quality, etc.) that are known to affect the facial representations computed by modern face recognition models. The datasets were selected for the experiments, because they represent standard (and challenging) benchmarks for evaluating face recognition

TABLE I

OVERVIEW OF THE EXPERIMENTAL DATASETS AND SETUP.

Database	#Images	#Subjects	Purpose	Labels/Variability
CelebA [13]	202,599	10,117	Train/Test	ID, 40 attributes
LFW [7]	13,233	5,749	Test	ID, gender
Adience [5]	19,370	2,284	Test	ID, age, gender

techniques and/or attribute-prediction models and ship with the attribute labels required by our training procedure. We use aligned versions of the datasets for our experiments and crop the facial regions from all images in such a way that the eyes are located at approximately the same locations. We then rescale the images to a fixed size of 224×224 and use the rescaled faces to extract the facial representations needed for the evaluation. A high-level comparison of the three datasets is given in Table I and a few representative example images from the datasets are presented in Fig. 3.

Face representations. To generate face representations x for the experiments, we use the VGGFace2 model from [1]. The model represents a ResNet-50 CNN model trained on a datasets of more than 3.3 million faces corresponding to $9000+$ identities and achieves state-of-the-art performance on challenging face datasets, such as IJB-A, IJB-B and IJB-C [14]. The model is chosen for the experiments due to its outstanding performance and the fact that a (pre-trained) open-source version is readily available[1]. Using VGGFace2, we extract 512 dimensional representations from the input face images and use the computed representations as the basis for training and evaluating PFRNet.

Experimental setup. We split the CelebA dataset into two disjoint image sets and make sure the subjects in the two sets do not overlap. The first sets comprises 70% of CelebA data (i.e., $141,819$ images) and is used to train PFRNet, whereas the second set consists of the remaining 30% of CelebA data (i.e., $60,780$ images) and is used as test data for the performance evaluation. Images from LFW and Adience are employed exclusively for testing purposes and are used to assess the generalization capabilities of PFRNet. To have a consistent evaluation setup across all datasets, we organize the (test) images of each dataset into 5 folds for both, face recognition as well as gender recognition experiments. For the gender recognition experiments, we make sure that each fold has a balanced number of male and female subjects and that the identities are evenly distributed across all of the folds. For the face recognition experiments we conduct verification experiments and again construct 5 folds for each dataset. We include an equal number of image pairs corresponding to matching and non-matching identities in each fold. Note again that the goal of our experiments is to evaluate the ability of PFRNet to conceal potentially sensitive gender information without compromising face recognition performance. We, therefore, do not follow the predefined experimental protocols defined for the datasets, but define a consistent protocol for all datasets that allows us to compare the performance of PFRNet in a common setting.

[1] https://github.com/WeidiXie/Keras-VGGFace2-ResNet50

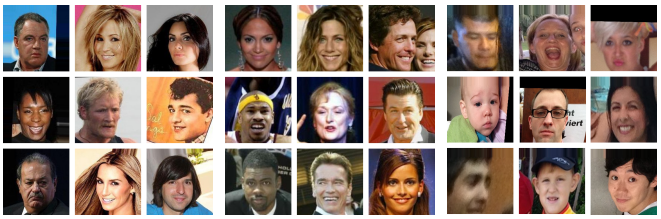


Fig. 3. Example images from: CelebA (left), LFW (middle), and Adience (right). Images from all datasets are cropped so that the eyes are always in approximately the same location. Note the difference in visual appearance.

TABLE II

SUMMARY OF PFRNET ARCHITECTURE.

PFRNet part	Layer	Input size	Output size
E_{ind}	FC+ReLU	512	512
	FC+ReLU	512	512
	FC+ReLU	512	512
	FC	512	496
E_{dep}	FC+ReLU	512	256
	FC+ReLU	256	128
	FC+ReLU	128	64
	FC	64	16
D	FC+ReLU	512	512
	FC+ReLU	512	512
	FC+ReLU	512	512
	FC	512	512

B. Measuring performance

We follow established methodology [28], [29] and report separate error measures for both relevant tasks:

- for gender recognition we report the fraction of incorrectly classified images (fic), and
- for face recognition we report equal error rates (eer) computed in face verification experiments.

To have a single scalar measure for the experiments, we also adopt the so-called *privacy-gain identity-loss coefficient* (PIC) proposed in [29]. PIC is defined as:

$$PIC = \frac{fic' - fic}{fic} - \frac{eer' - eer}{eer}, \tag{7}$$

where fic' and eer' were computed from the attribute suppressed representations, whereas the errors fic and eer were calculated from the original (unmodified) face representations. Positive PIC values imply that the privacy gain is higher than the potential loss in face recognition performance and higher values indicate better performance.

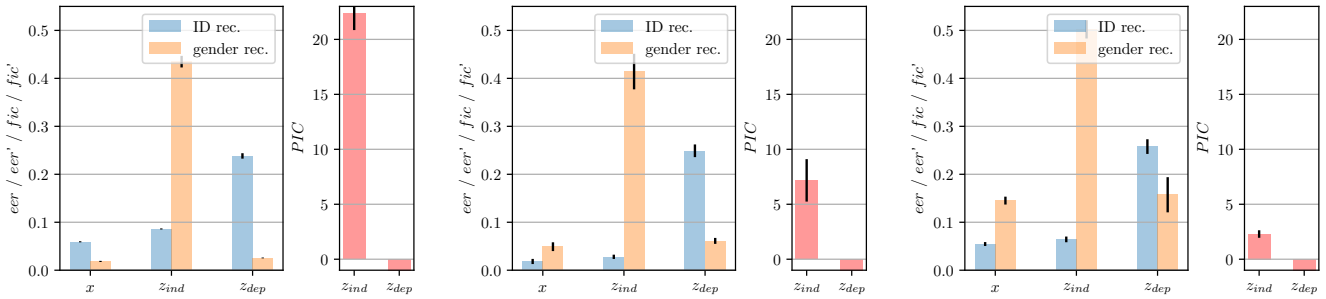
C. Training details

Training procedure. We train PFRNet on face representations of $141,819$ images from the CelebA dataset. These representations are split between a set of $127,637$ samples that are used for the actual training of our model, and a set of $14,182$ validation samples that help spot overfitting issues during training. We use the overall learning objective from Eq. (6) and optimize for the first two moments of the loss terms $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$, i.e., $\alpha_{max} = 2$ and $\beta_{max} = 2$. The parameter σ_β is set to $1/\sqrt{2}$. We use the Adam [10] optimizer with an initial learning rate of 10^{-4} for the training

TABLE III

EVALUATION RESULTS ON CELEBA, LFW AND ADIENCE. RESULTS ARE PRESENTED IN THE FORM $\mu \pm \sigma$ - COMPUTED OVER 5 FOLDS.

Dataset	x		z_{ind}			z_{dep}		
	fic	eer	fic'	eer'	PIC	fic'	eer'	PIC
CelebA	0.018 ± 0.001	0.059 ± 0.002	0.435 ± 0.012	0.086 ± 0.001	22.378 ± 1.530	0.026 ± 0.001	0.238 ± 0.006	-2.656 ± 0.174
LFW	0.049 ± 0.009	0.018 ± 0.005	0.414 ± 0.037	0.028 ± 0.004	7.174 ± 1.931	0.061 ± 0.006	0.249 ± 0.013	-13.457 ± 3.769
Adience	0.145 ± 0.008	0.056 ± 0.004	0.502 ± 0.019	0.064 ± 0.006	2.292 ± 0.343	0.157 ± 0.037	0.258 ± 0.015	-3.526 ± 0.410



(a) CelebA results (b) LFW results (c) Adience results

Fig. 4. Evaluation results on (a) CelebA, (b) LFW and (c) Adience - results are presented in the form of average error rates (eer, fic) and standard errors computed over 5 experimental folds. The results show that the latent representation z_{ind} generated by PFRNet retains the majority of the discriminative information needed for face recognition, while it significantly reduces the amount of gender information compared to the initial face representation x. The second part of the latent representation z_{dep} retains most of the gender information, but also encodes identity to some extent. The PFRNet generalizes well across dataset and ensures efficient feature disentanglement. Positive PIC scores are also observed on all three datasets with the attribute suppressed representations z_{ind}. The figure is best viewed in color.

procedure. The fully connected (FC) layers comprising PFRNet are initialized from a uniform distribution $\mathcal{U}(-\sqrt{k}, \sqrt{k})$, where $k = 1/\rho$ and ρ denotes the number of input features into the FC layers. We balance the contribution of the individual loss terms from Eq. (6) using $\lambda_\alpha = 0.01, \forall\alpha$, $\lambda_\beta = 0.01, \forall\beta$, which were found empirically to produce useful face representations. We make no additional effort to optimize these values for better performance. The learning procedure is stopped early if the learning objective on the validation data does not improve for more than 100 epochs.

Model architecture. The initial architecture of PFRNet used in the experiments is shown in Table II. The two-path encoder (comprising E_{ind} and E_{dep}) is designed in such a way that the dimensionality of the combined latent representation $z_{ind} \oplus z_{dep}$ is the same as the dimensionality of the original face representations $x \in \mathbb{R}^d$, $d = 512$. E_{ind} generates a 496-dimensional latent representation z_{ind} for identity recognition, while E_{dep} compresses the gender information into a compact 16-dimensional representation z_{dep}. Most of the space in the latent representation is allocated for z_{ind}, which needs to encode as much of the discriminative information of x (for identity recognition) as possible, while a smaller part of the latent space is left for encoding gender in z_{dep} - a more detailed analysis of the impact of latent space dimensionality is presented in Section IV-E. The decoder D has four FC layers with intermediate ReLU activations.

Note that the two-path encoder architecture was selected due to the overall learning objective of PFRNet. Because two distinct loss functions are applied on different parts of the latent representation, separate encoder parts are used to produce the two latent representations z_{ind} and z_{dep}. This architectural detail ensures that there is no interaction between the loss functions of z_{ind} and z_{dep} and offers an additional mechanism for disentangling gender from identity in the latent representations.

Training complexity. The training procedure takes between 1 and 2 hours on a GeForce GTX 1070 Ti GPU to converge. Once trained, our model is able to generate latent representations from the input vectors x at an average speed of 0.008 ms/image on GPU in batch mode (with a batch size of 100) or 0.4 ms/image in real-time mode.

D. Performance evaluation

PFRNet performance. In order to evaluate the performance of PFRNet we first conduct face verification and gender recognition experiments on CelebA, LFW and Adience. For the verification experiments, we compare face representations of genuine and impostor pairs using the cosine similarity. For the gender-recognition experiments, we train a Logistic Regression (LR) classifier on the CelebA training data and then use the trained model in all following experiments.

In Table III and Fig. 4 we compare results generated based on the original face representation x, the gender-suppressed representation z_{ind} and the latent representation encoding gender z_{dep}. Using the original representations x as the basis for the experiments, we observe average values of eer/fic of $5.9\%/1.8\%$ on CelebA, $1.8\%/4.9\%$ on LFW and $5.6\%/14.5\%$ on Adience for the face-verification/gender-recognition experiments. Verification performance degrades slightly on all three datasets when the attribute-suppressed representations are used, resulting in eer' values of 8.6% on CelebA, 2.8% on LFW and 6.4% on Adience. The error rates for gender recognition performance, on the other hand,

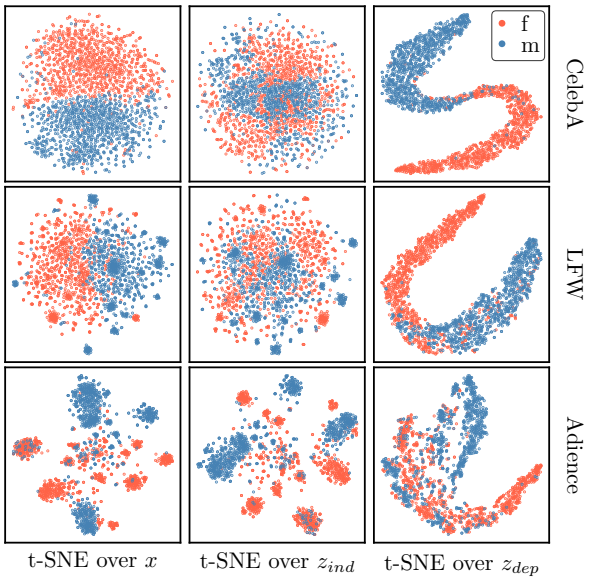


Fig. 5. t-SNE visualization of different face representations (x, z_{ind}, z_{dep}) for three experimental datasets (CelebA, LFW, Adience) in 2D. The data samples are colored with respect to gender: male - blue (1000 samples), female - orange (1000 samples). The plots show that male and female subjects are reasonably well separated in x-space, the distributions mix in z_{ind} and are again well separable in z_{dep}-space. Best viewed in color.

TABLE IV

EFFECT OF z_{dep} AND z_{ind} DIMENSIONALITY. PIC VALUES ARE PRESENTED IN THE FORM: $\mu \pm \sigma$, COMPUTER OVER 5 FOLDS.

Model	$dim(z_{ind})$	$dim(z_{dep})$	PIC (CelebA)	PIC (LFW)	PIC (Adience)
PFRNet $D1$	255	1	24.743 ± 1.860	6.690 ± 1.338	2.478 ± 0.423
PFRNet $D2$	248	8	24.575 ± 1.801	7.063 ± 1.482	2.233 ± 0.427
PFRNet $D3$	192	64	25.261 ± 2.035	7.459 ± 1.795	2.370 ± 0.525
Average	$dim(z) = 256$		24.860 ± 1.899	7.071 ± 1.538	2.360 ± 0.458
PFRNet $D4$	511	1	21.192 ± 1.772	7.087 ± 1.099	2.611 ± 0.205
PFRNet $D5$	496	16	22.378 ± 1.530	7.174 ± 1.931	2.292 ± 0.343
PFRNet $D6$	384	128	22.797 ± 1.714	7.196 ± 1.579	2.373 ± 0.443
Average	$dim(z) = 512$		22.084 ± 1.688	6.911 ± 1.413	2.268 ± 0.272
PFRNet $D7$	1023	1	22.513 ± 1.967	7.566 ± 1.933	2.140 ± 0.203
PFRNet $D8$	992	32	18.403 ± 1.487	6.612 ± 1.683	1.973 ± 0.206
PFRNet $D9$	768	256	19.207 ± 1.680	7.152 ± 1.974	1.837 ± 0.302
Average	$dim(z) = 1024$		20.041 ± 1.711	7.110 ± 1.863	1.983 ± 0.237

reach close to random values on all three datasets with averages fic' scores of 43.5% on CelebA, 41.4% on LFW and 50.2% on Adience. The experiments with z_{dep} show degradations in both eer as well as fic compared to the original representations x. If we look at the PIC scores, we notice that PFRNet significantly improves the level of soft-biometric privacy for all three datasets, when generating the attribute-suppressed latent representation z_{ind} from x. The smallest value of PIC is observed on Adience. However, this is a consequence of the relatively larger initial errors (eer and fic) obtained with x on this dataset and the fact that PIC measures relative improvements.

Overall, the results of this first series of experiments suggest that PFRNet is able to efficiently disentangle facial representations into two parts, where the first part z_{dep} encodes identity information in a gender agnostic way and the majority of gender information is carried over to the second part of the latent representation z_{dep}. The model also generalizes reasonably well over different datasets, despite

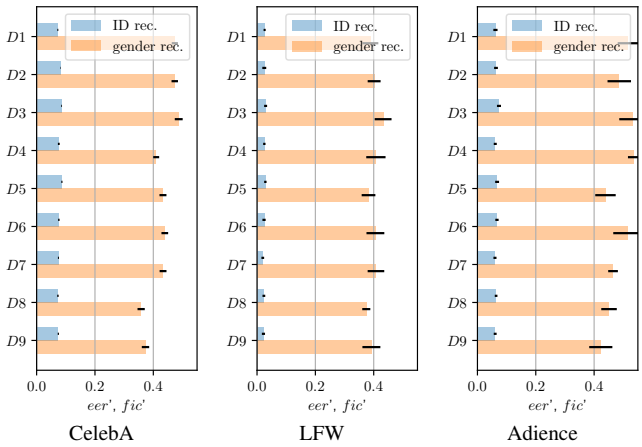


Fig. 6. Impact of the latent-space dimensionality on face verification errors (blue) and gender recognition errors (orange) achieved with z_{ind}. Results are shown in the form of mean error rates and corresponding standard errors computed over five folds for three experimental datasets: CelebA (left), LFW (middle) and Adience (right). The characteristics of the nine models ($D1$ - $D9$) are summarized in Table IV. The figure is best viewed in color.

the fact that the datasets exhibit considerable differences in visual characteristics as illustrated Fig. 3.

Representation visualization. To gain additional insight into the characteristics of the generated latent representations we use t-distributed stochastic neighbour embedding (t-SNE) [30] and visualize the distributions of x, z_{dep} and z_{ind} in 2D in Fig. 5. Here, we use 1000 randomly selected samples for each gender from the three experimental datasets. We color-code the samples with respect to whether they correspond to male or female subjects. The plots show that the male and female distributions are quite well separated in all three datasets when the original face representations x are considered, which shows that gender information is clearly encoded in the data. Male and female distributions mix in the space of attribute-suppressed representations z_{ind} and are again separated in z_{dep}-space. The observed behaviour again shows that PFRNet is able to disentangle gender from identity information and consequently to improve the level of soft-biometric privacy. We note that the distributions for Adience are somewhat different from the distributions of the other two datasets, which is a consequence of the greater variability in appearance of the images in this dataset.

E. Model analysis

Next, we evaluate how different components of the PFRNet model affect the level of soft-biometric privacy.

Dimensionality of latent representations. We first investigate how the dimensionality of the latent representations z_{dep} and z_{ind} affects the level of soft-biometric privacy ensured by PFRNet. To examine this issue, we train 9 different variants of PFRNet with latent-space dimensionalities of either 256, 512 or 1024 and differently sized vectors z_{ind} and z_{dep}. A summary of the latent-space dimensionality of the trained models ($D1$ through $D9$) is given in Table IV.

From the results in Table IV and Fig. 6 we see that the gender suppressed representation z_{ind} generated by PFRNet is relatively robust with respect to the dimensionality of the

TABLE V

IMPACT OF LOSS TERMS ON PIC SCORES.

Model	α_{max}	β_{max}	PIC (CelebA)	PIC (LFW)	PIC (Adience)
PFRNet $L1$	0	0	-0.039 ± 0.071	-0.380 ± 0.215	-0.308 ± 0.117
PFRNet $L2$	1	0	22.123 ± 1.783	6.286 ± 1.417	-0.135 ± 0.244
PFRNet $L3$	2	0	20.364 ± 1.800	7.351 ± 1.664	2.228 ± 0.220
PFRNet $L4$	3	0	22.032 ± 1.585	7.208 ± 1.795	-0.133 ± 0.142
PFRNet $L5$	4	0	20.663 ± 1.513	6.260 ± 1.559	-0.028 ± 0.187
PFRNet $L6$	0	1	4.694 ± 0.528	2.484 ± 0.692	-0.171 ± 0.164
PFRNet $L7$	0	2	3.986 ± 0.317	1.428 ± 0.769	-0.084 ± 0.137
PFRNet $L8$	0	3	1.552 ± 0.115	0.677 ± 0.324	-0.251 ± 0.132
PFRNet $L9$	0	4	3.706 ± 0.387	1.141 ± 0.588	-0.179 ± 0.159
PFRNet $L10$	1	1	18.189 ± 1.631	5.618 ± 1.481	2.056 ± 0.430
PFRNet $L11$	2	2	22.378 ± 1.530	7.174 ± 1.931	2.292 ± 0.343
PFRNet $L12$	3	3	20.209 ± 1.682	8.070 ± 1.897	2.317 ± 0.315
PFRNet $L13$	4	4	17.743 ± 1.630	7.027 ± 1.426	2.058 ± 0.122

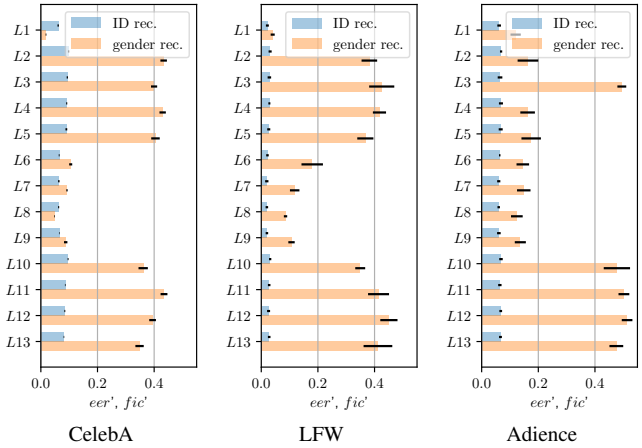


Fig. 7. Impact of loss terms on face verification errors (blue) and gender recognition errors (orange) achieved with z_{ind}. Results are shown in the form of mean error rates and corresponding standard errors computed over five folds for three experimental datasets: CelebA (left), LFW (middle) and Adience (right). The characteristics of the loss combination (L1-L13) are summarized in Table V. The figure is best viewed in color.

latent space, but also with respect to relative size of the latent representations z_{ind} and z_{dep}. Even in the most extreme cases, where only a single variable is allocated for z_{dep} (models $D1$, $D4$, $D7$), the model is still able to suppress gender information in z_{ind} to a considerable extent on all datasets. When the size of z_{dep} is increased, PIC values generated from z_{ind} increase on CelebA, but this trend is not consistent across the remaining two datasets.

Interestingly, the smallest latent-space dimensionality (see models $D1$-$D3$ and results for Averages) and consequently the smallest dimensionalities of z_{ind} result in higher levels of soft-biometric privacy than larger dimensionalities on all three datasets. We also observe a trend that the PIC values decrease as the dimensionality of the latent space gets larger, but the differences here are minimal. This seems to be a consequence of slightly better attribute suppression at this dimensionality (on average), since the verification performance is mostly stable across different dimensionalities of z_{ind}.

Impact of loss functions. Next, we examine the contribution of the individual loss terms in the learning objective from Eq. (6). Specifically, we change the values of α_{max} and β_{max} from 0 to 4 and observe results for different combinations of the two parameters in Table V and Fig. 7. Here, a value of $\alpha_{max} = 0$ indicates that $\mathcal{L}_\alpha$ is not included in the learning objective and similarly a value of $\beta_{max} = 0$ suggests that $\mathcal{L}_\beta$ is not used during training. We report only results for experiments with the gender suppressed representations z_{ind}.

As can be seen, the model is not able to suppress any gender information when only the reconstruction term is considered during the training procedure - see results for $L1$. Without the gender-discrimination term $\mathcal{L}_\beta$ (see $L2$-$L5$), PFRNet is still able to suppress gender information in z_{ind} on CelebA and LFW, but does not generalize well to the Adience data. If we remove the gender-concealing loss $\mathcal{L}_\alpha$ (models $L6$-$L9$), the results show small privacy gains compared to the original face representations x on CelebA and LFW, but the overall performance is not really competitive. When both $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$ are considered (models $L10$-$L13$), results exhibit a certain level of variability, but with 2 statistical moments considered for both loss terms, i.e., $\alpha_{max} = 2$ and $\beta_{max} = 2$, the most consistent results are achieved across the three datasets. These results show that all loss terms are important and contribute to both the performance as well as generalization ability of PFRNet.

F. Comparison with state-of-the-art

In the last series of experiments we compare PFRNet with state-of-the-art models from the literature that aim at increasing the level of soft-biometric privacy at the template level. Specifically, we implement the following two techniques and compare them to PFRNet: the Incremental Variable Elimination (IVE) technique [28] and the Cosine Sensitive Noise (CNS) transformation [29]. We optimize both competing techniques on CelebA for optimal performance to ensure a fair comparison to PFRNet and then evaluate all methods using the setup from the previous sections. Specifically, the IVE parameters, number of steps and number of elimination per step (set here to 60 and 5, respectively), and the similarity preserving parameter θ of CNS (set here to 0.8), are chosen to maintain a low eer on the training data [28], [29].

From Table VI we see that PFRNet and IVE both result in positive PIC values on all three datasets, while CNS generates values close to zero on LFW and Adience. Overall, PFRNet is clearly the top performer and outperforms IVE, the runner up, by a factor of close to 2 on CelebA and a factor of close to 6 on LFW in PIC terms. The difference in performance is even more evident on the more challenging Adience dataset, where PFRNet results in 10 times higher PIC scores than the IVE approach. We also compare the methods graphically in Fig. 8. Here, we show fic' scores on the x-axis, eer' scores on the y-axis and encode the value of PIC in the radius of a circle centered at (fic', eer'). Thus, locations closer to the upper right corner of the graph and larger circles indicate better performance. The results of this comparison again point to the competitive performance of PFRNet for the task of soft-biometric privacy.

It needs to be noted that the CNS approach did not suppress the gender to the degree reported in [29], which can be explained by the different nature of the face representations used for the experiments. While SphereFace [12] used in [29] produces centered feature values that are predominantly not equal to zero, VGGFace2 [1] produces feature values

TABLE VI

COMPARISON TO THE STATE-OF-THE-ART. RESULTS ARE PRESENTED IN THE FORM: $\mu \pm \sigma$ COMPUTED OVER 5 FOLDS FOR ALL SCORES.

Method	CelebA			LFW			Adience		
	fic'	eer'	PIC	fic'	eer'	PIC	fic'	eer'	PIC
PFRNet (ours)	0.435 ± 0.012	0.086 ± 0.001	22.378 ± 1.530	0.414 ± 0.037	0.028 ± 0.004	7.174 ± 1.931	0.502 ± 0.019	0.064 ± 0.006	2.292 ± 0.343
IVE [28]	0.214 ± 0.004	0.086 ± 0.001	10.272 ± 0.780	0.134 ± 0.009	0.031 ± 0.007	1.126 ± 0.831	0.216 ± 0.097	0.071 ± 0.003	0.182 ± 0.134
CNS [29]	0.023 ± 0.001	0.063 ± 0.003	0.212 ± 0.104	0.032 ± 0.002	0.020 ± 0.005	-0.439 ± 0.164	0.135 ± 0.020	0.061 ± 0.004	-0.182 ± 0.116

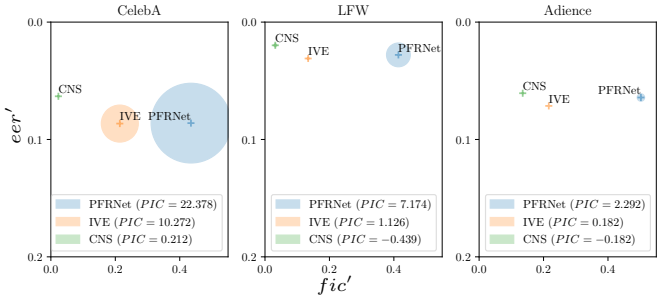


Fig. 8. Comparison of PFRNet with state-of-the-art methods from the literature. fic' scores are shown on the x-axis, eer' scores are shown on the y axis and the size of the circles corresponds to the PIC values achieved. Locations closer to the upper right corner and circles with larger radii indicate better performance in terms of soft-biometric privacy. Note that PFRNet clearly outperforms the competing methods on all three datasets.

that follow a heavy-tailed Poisson-like distribution where many values are in fact equal to zero. This renders the CNS approach less effective in our experiments and points to a potential shortcoming of this approach when applied on face representations, such as the ones produced by VGGFace2.

V. CONCLUSIONS

We have presented PFRNet, a deep learning model capable of suppressing soft-biometric information in biometric templates to a considerable extent, while retaining most of the discriminative information required for identity recognition. We tested the model in rigorous experiments on the CelebA, LFW and Adience datasets and presented comparative results with competing methods from the literature. Our results suggest that the model ensures competitive results and generalizes well across different datasets. As part of our future work, we will examine possibilities for including additional loss terms into the learning objective of PFRNet that encourage the latent representations to discriminate better with respect to identity, which should contribute further to increasing the level of soft-biometric privacy.

ACKNOWLEDGEMENTS

This research was supported in parts by the ARRS Project J2-1734 “Face deidentification with generative deep models”, and ARRS Research Programs P2-0250 (B) “Metrology and Biometric Systems” and P2-0214 (A) “Computer Vision”.

REFERENCES

[1] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGface2: A dataset for recognising faces across pose and age. In *FG*, 2018.
[2] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta. Anonymizing k-facial attributes via adversarial perturbations. In *IJCAI 2018*, 2018.
[3] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? a survey on soft biometrics. *TIFS*, 11(3):441–467, 2015.
[4] P. Dhar, A. Bansal, C. D. Castillo, J. Gleason, P. J. Phillips, and R. Chellappa. How are attributes expressed in face DCNNs? In *FG 2020*, pages 1–8, 2020.
[5] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *TIFS*, 9(12):2170–2179, 2014.
[6] O. Gafni, L. Wolf, and Y. Taigman. Live face de-identification in video. In *ICCV 2019*, 2019.
[7] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, Uni. Mas., 2007.
[8] A. Jourabloo, X. Yin, and X. Liu. Attribute preserved face de-identification. In *ICB 2015*, pages 278–285, 2015.
[9] E. J. Kindt. *Privacy and data protection issues of biometric applications*, volume 1. Springer, 2016.
[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR 2014*, 2014.
[11] T. Li and L. Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *CVPR-W 2019*, 2019.
[12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017.
[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV 2015*, 2015.
[14] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. IARPA Janus Benchmark - C: Face Dataset and Protocol. In *ICB 2018*, pages 158–165, 2018.
[15] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer. k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification. *Entropy*, 20(1):60, 2018.
[16] B. Meden, R. C. Mallı, S. Fabijan, H. K. Ekenel, V. Štruc, and P. Peer. Face deidentification with generative deep neural networks. *IET Signal Processing*, 11(9):1046–1054, 2017.
[17] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *ICB 2018*, pages 82–89, 2018.
[18] V. Mirjalili, S. Raschka, and A. Ross. Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers. *BTAS 2018*, pages 1–10, 2018.
[19] V. Mirjalili, S. Raschka, and A. Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019.
[20] V. Mirjalili and A. Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *IJCB*, 2017.
[21] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.
[22] A. Othman and A. Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *ECCV-W*, pages 682–696, 2014.
[23] C. J. Parde, C. Castillo, M. Q. Hill, Y. I. Colon, S. Sankaranarayanan, J.-C. Chen, and A. J. O’Toole. Face and image representation in deep cnn features. In *FG 2017*, pages 673–680, 2017.
[24] O. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC 2015*, page 6, 2015.
[25] S. Ribaric, A. Ariyaeeinia, and N. Pavesic. De-identification for privacy protection in multimedia content: A survey. *SPIC*, 47, 2016.
[26] A. Rozsa, M. Günther, E. M. Rudd, and T. Boult. Facial attributes: Accuracy and adversarial robustness. *PRL*, 2017.
[27] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR 2015*, 2015.
[28] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. Suppressing gender and age in face templates using incremental variable elimination. In *ICB 2019*, pages 4–7, 2019.
[29] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper. Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. *Applied Intelligence*, pages 1–18, 2019.
[30] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008.

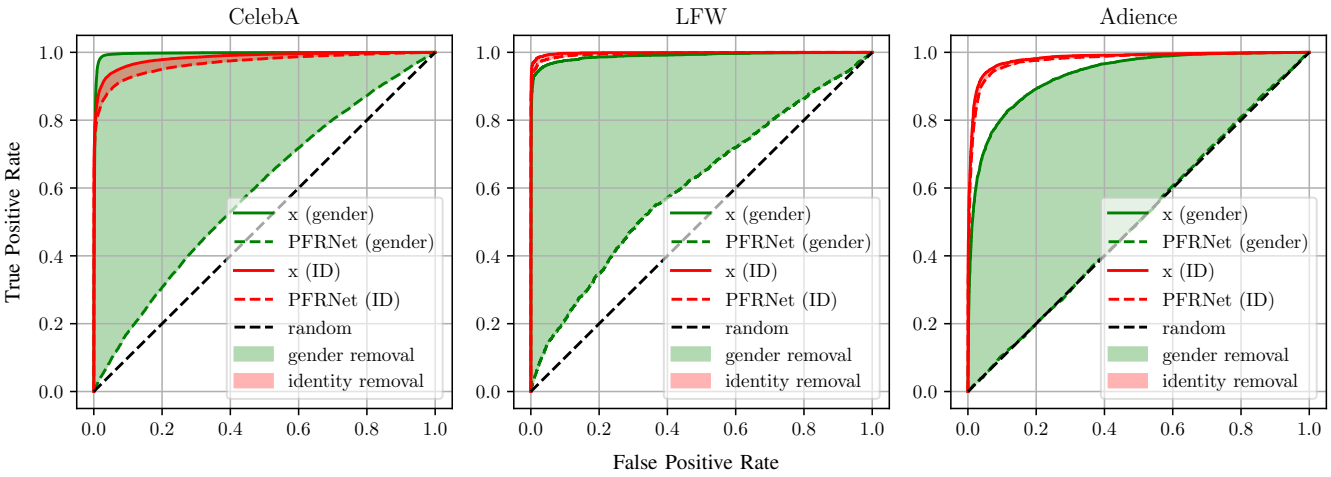


Fig. 9. Complete ROC evaluation of PFRNet on the CelebA, LFW and Adience datasets. The red graphs show face verification performance before (i.e., on x - solid red line) and after (i.e., on z_{ind} - dashed red line) disentanglement with PFRNet. The green graphs show gender recognition performance before (i.e., on x - solid green line) and after (i.e., on z_{ind} - dashed green line) disentanglement with PFRNet. The surface area shaded green corresponds to the amount of suppressed gender information and the shaded surface area shaded red corresponds to the loss of identity information. The dashed black curve indicates random performance. The graphs show that PFRNet performs well across all three datasets. The figure is best viewed in color.

APPENDIX

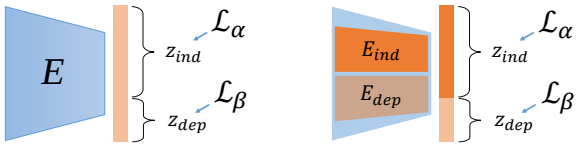
In this Appendix we show some additional results and analyses related to PFRNet.

A. Complete ROC evaluation

In the main part of the paper we evaluated the performance of PFRNet using quantitative performance measures. Specifically, we used the equal error rate (eer) to measure the performance of face verification and the fraction of incorrectly classified images (fic) to evaluate gender recognition performance. These measures allowed us to quantify soft-biometric privacy gains with easy-to-interpret scalar measures. In Fig. 9, we now present complete Receiver Operating Characteristic (ROC) curves for the three experimental datasets to show a more complete picture of the performance of PFRNet across all ROC operating points.

Here, the red curves show the face verification performance with the original face representation x and the dashed red curves show the verification performance after disentanglement (i.e., using z_{ind}). The surface area shaded red between the two curves corresponds to the loss of identity information due to the mapping from x to z_{ind}. The solid green curves show gender recognition performance (female is considered the positive class), whereas the dashed curves show gender recognition performance with the attribute suppressed representation z_{ind}. As can be seen the surface area shaded green that corresponds to the amount of suppressed gender information is significantly larger than the red surface area corresponding to the loss of identity information, which suggests that PFRNet results in improved levels of soft biometric privacy across the entire ROC curve and not only for a selected operating point.

When comparing results across datasets we see a similar setting as in the main part of the paper. PFRNet performs well across all datasets and ensures reasonable levels of generalization despite different visual characteristics of the images from the three datasets.



(a) Single-path encoder (b) Two-path encoder

Fig. 10. Illustration of the architectures compared. The single-path encoder generates a latent representation, on which two loss functions are applied - each loss on a separate part of the latent representation. The two-path encoder uses separate model parts to produce the latent representations for the two losses $\mathcal{L}_\alpha$ and $\mathcal{L}_\beta$.

TABLE VII
SINGLE-PATH ENCODER ARCHITECTURE USED FOR THE EVALUATION.

PFRNet part	Layer	Input size	Output size
E	FC+ReLU	512	512
	FC+ReLU	512	512
	FC+ReLU	512	512
	FC	512	512

B. Impact of model architecture.

In the main part of the paper, we introduced PFRNet with a two-path encoder design instead of simple and straightforward single-path architecture. This design choice was motivated by the fact that it is difficult to penalize different parts of the latent representation using separate loss functions if the model topology is shared by both parts - illustrated in Fig. 10. As can be seen, penalizing one part of the latent representation affects the second part as both parts are generated by the same model E. In a two-path architecture separate parts of the encoder are dedicated for the generation of the two latent-representation parts.

TABLE VIII

COMPARISON OF THE SINGLE-PATH AND TWO-PATH PFRNET VARIANTS AND THEIR PERFORMANCE.

Method	Dataset	z_{ind}			z_{dep}		
		fic'	eer'	PIC	fic'	eer'	PIC
PFRNet - Two-path Encoder	Celeba	0.435 ± 0.012	0.086 ± 0.001	22.378 ± 1.530	0.026 ± 0.001	0.238 ± 0.006	-2.656 ± 0.174
	LFW	0.414 ± 0.037	0.028 ± 0.004	7.174 ± 1.931	0.061 ± 0.006	0.249 ± 0.013	-13.457 ± 3.769
	Adience	0.502 ± 0.019	0.064 ± 0.006	2.292 ± 0.343	0.157 ± 0.037	0.258 ± 0.015	-3.526 ± 0.410
PFRNet - Single-path Encoder	Celeba	0.021 ± 0.002	0.073 ± 0.002	-0.081 ± 0.051	0.023 ± 0.002	0.302 ± 0.006	-3.862 ± 0.137
	LFW	0.119 ± 0.016	0.024 ± 0.004	1.112 ± 0.423	0.043 ± 0.006	0.308 ± 0.010	-17.328 ± 4.542
	Adience	0.393 ± 0.043	0.061 ± 0.006	1.579 ± 0.280	0.149 ± 0.026	0.300 ± 0.008	-4.484 ± 0.331

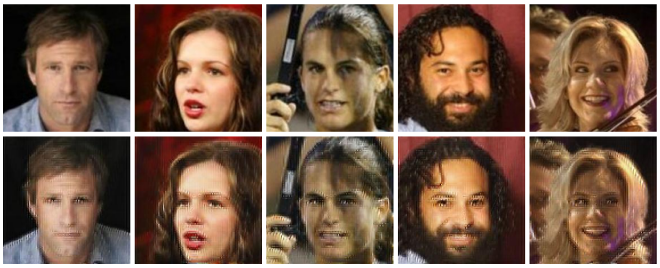


Fig. 11. Visualization of the modified images generated by the SAN-based approach from [18]. The top row shows original images from LFW and the bottom row shows modified images where the gender attribute has been altered. The modified images typically display noise-like patterns around specific facial areas, such as the mouth, eyes or hair that seem to affect the performance of gender-recognition models.

In this section we aim at validating the initial choice of our two-path encoder topology through a series of comparative experiments. To this end, we design and train PFRNet with a single-path encoder and compare it to the two-path model evaluated in the main part of the paper. The single-path encoder used for these experiments consists of four fully connected layers with intermediate ReLU activations, as summarized in Table VII. The decoder is kept unchanged compared to the original PFRNet. The first 496 elements of the 512-dimensional latent representation are selected to encode identity information and the last 16 elements to encode gender. For the loss functions $\alpha_{max} = 2$ and $\beta_{max} = 2$ are used. We again report performance using the same performance measures as in the main part of the paper.

The results in Table VIII show that the latent representation z_{ind} generated by the single-path encoder PFRNet contains slightly more identity information, while gender information is considerably less suppressed compared to z_{ind} vectors generated by the PFRNet with a two-path encoder. This observation is also reflected in the PIC scores where the single-path version of PFRNet produces scores close to 0, whereas the two-path version ensures considerable soft-biometric privacy gains with large positive PIC scores for the three experimental datasets. The results on z_{dep} are comparable with both PFRNet versions.

Overall, the results of this experiment suggest that the two-path architecture has an important role in PFRNet that allows to penalize different parts of the latent representation with separate loss functions and consequently leads to higher levels of soft-biometric privacy.

TABLE IX

COMPARISON WITH IMAGE-LEVEL TECHNIQUES ON LFW.

Method	fic' (or fic)	eer' (or eer)	PIC
Original representation x	0.049 ± 0.009	0.018 ± 0.005	n/a
PFRNet - on z_{ind} (ours)	0.414 ± 0.037	0.028 ± 0.004	7.174 ± 1.931
Mirjalili *et al.* [18]	0.128 ± 0.011	0.030 ± 0.006	0.995 ± 0.513

C. Comparison with image-level techniques

In Section IV-F we compared the proposed PFRNet model to competing techniques from the literature aimed at improving the degree of soft-biometric privacy at the template level. Here, we also compare to a state-of-the-art technique that operates at the image level and tries to ensure soft-biometric privacy by altering the input images in such a way that (typically pre-trained) classifiers fail to recognize a selected soft-biometric attribute, e.g., gender. While the goal of these techniques is to modify the input face images and not the face representations extracted by an arbitrary CNN model, it is not completely clear from the literature if these methods are useful for soft-biometric privacy at the template level.

To explore this issue, we implement the SAN-based approach from Mirjalili *et al.* [18] and train it on the training images of CelebA. We apply the SAN-based model on the test images from LFW and generate modified images as shown in the bottom row of Fig. 11. Next, we use the VGGFace2 model to extract 512 dimensional face representations from the original as well as from the modified images. Once the representations are computed for all images of the 5 experimental folds, we again conduct face verification and gender recognition experiments using the same setup as described in Section IV-D. Note again that we train the gender classifier (i.e., logistic regression) on the face representations of the modified images, so we in fact explore how much gender information is still present in the data and not how well it is masked.

From the results in Table IX we see that compared to the original representation, the approach from Mirjalili *et al.* [18] improves the level of soft-biometric privacy with a positive PIC value around 1. PFRNet, on the other hand, performs significantly better with a PIC value of more than 7 on this datasets, as already reported in the main part of the paper. These results are expected, as PFRNet is designed specifically for the task of ensuring soft-biometric privacy at the template level, whereas the SAN-based approach is aimed at another application domain and designed under different

assumptions. Nevertheless, it is able to outperform the state-of-the-art CNS approach (see Table VI for details) in terms of PIC score, which was proposed originally for soft-biometric privacy enhancement at the template level.