Gender Privacy Angular Constraints for Face Recognition

Zohra Rezgui[1], Nicola Strisciuglio[1], Raymond Veldhuis[1,2]
{z.rezgui, n.strisciuglio, r.n.j.veldhuis}@utwente.nl,
[1] DMB Group, University of Twente
[2] IIK Department, Norwegian University of Science and Technology, Gjøvik, Norway

Abstract—Deep learning-based face recognition systems produce templates that encode sensitive information next to identity, such as gender and ethnicity. This poses legal and ethical problems as the collection of biometric data should be minimized and only specific to a designated task. We propose two privacy constraints to hide the gender attribute that can be added to a recognition loss. The first constraint relies on the minimization of the angle between gender-centroid embeddings. The second constraint relies on the minimization of the angle between gender specific embeddings and their opposing gender-centroid weight vectors. Both constraints enforce the overlapping of the gender specific distributions of the embeddings. Furthermore, they have a direct interpretation in the embedding space and do not require a large number of trainable parameters as two fully connected layers are sufficient to achieve satisfactory results. We also provide extensive evaluation results across several datasets and face recognition networks, and we compare our method to three state-of-the-art methods. Our method is capable of maintaining high verification performances while significantly improving privacy in a cross-database setting, without increasing the computational load for template comparison. We also show that different training data can result in varying levels of effectiveness of privacy-enhancing methods that implement data minimization.

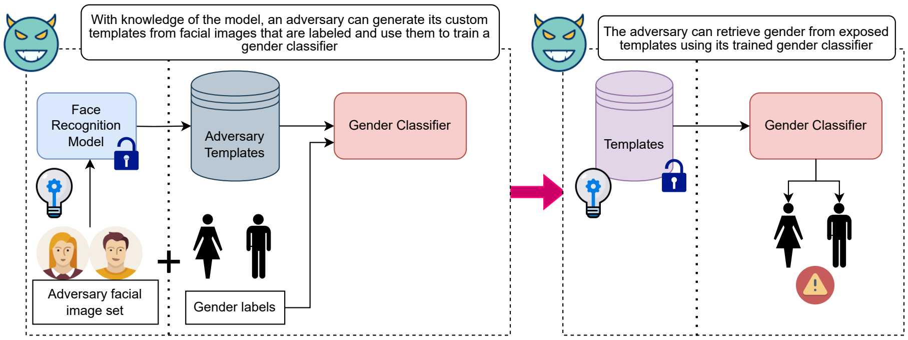
Index Terms—Privacy-enhancing techniques, soft-biometric privacy, gender classification, face recognition.

1 Introduction

Deep learning has been revolutionary for face recognition. CNNs in particular, have enabled the training and convergence of algorithms with large complexity allowing the learning of features that are highly discriminative for the face recognition task. This breakthrough resulted in face recognition systems that were progressively more effective at recognizing faces even in challenging scenarios, including changes in lighting, pose, and expression [1], [2], [3], [4]. Next to containing information that is highly useful for the face verification and identification tasks, the features learned by deep-learning based face recognition systems entangle a variety of other information auxiliary to identity. Previous studies have shown that features extracted from the last layers of face recognition networks can be transferable to other tasks, such as gender, age or ethnicity classification [5], [6] as well as the classification of other fine-grained attributes such as hairstyle or the shape of eyebrows [7]. This entanglement between identity and auxiliary soft biometric attributes present in the facial templates poses privacy risks. For instance, in the event that such templates or their source model are exposed, an adversary can train classifiers that would undertake profiling of subjects based on demographic or other highly sensitive information. This can be problematic as the subject may have not consented to the processing of their biometric information for profiling tasks. We illustrate this risk in Figure 1. From a legal perspective, the disclosure of such information stemming from a face recognition model poses a potential issue to both the model developer and the party responsible for storing the templates. This is due to the fact that such information leakage runs counter to the data minimization principle outlined in the General Data Protection Regulation (GDPR)[1].

The correlation between soft biometrics and identity

[1] The data minimization principle states : *"Personal data shall be: adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"*.


With knowledge of the model, an adversary can generate its custom templates from facial images that are labeled and use them to train a gender classifier

The adversary can retrieve gender from exposed templates using its trained gender classifier

Fig. 1: Threat illustration: Face recognition features contain discriminative information on gender and can be used to train a gender classifier.

can also contribute to the demographic unfairness of biometric systems, which is a topical research problem [8], [9]. Face recognition systems are biased toward demographic categories, meaning they produce more errors for certain categories than others. This is usually due to skewed distributions of different categories in the training data [8] which cause the neural networks to overfit on the dominating category. The entanglement between demographic attributes and identity can further exacerbate this issue. In fact, if the embeddings for face verification are easily separable by a demographic attribute, they can potentially form significantly different distributions for different categories of the attribute. This makes the verification step prone to generating different error rates by category.

To remediate the aforementioned issues, a few works emerged that apply privacy-enhancing techniques at the template level. While some require a training procedure [10], [11], [12], others are training-free and rely on the shuffling of information in the templates rather than removing the sensitive information [13]. However, finding an optimal trade-off between privacy and face verification performance remains a difficult challenge. As an additional observation, previous studies often lack sufficient assessment of the generalizability of their approaches as they do not conduct evaluations on diverse, independent datasets from the training set. In some cases, the evaluation data for privacy-preserving approaches comes from the same source as the training data [10], [11] which does not allow for a real-life scenario evaluation. We present a more detailed overview of such works in Section 2.

In this paper, we focus on protecting the gender[2] attribute as it is easily learned from facial templates [6], [14]. Our proposed method takes advantage of the hyperspherical nature of the feature space used in many face recognition systems [4], [15], [16], [17]. We finetune a face recognition model by passing the feature vectors through a shallow network that projects them onto an unbiased feature space. We introduce angular constraints that when added to a recognition loss, consist in overlapping the distributions of the gender categories while maintaining the verification performance.

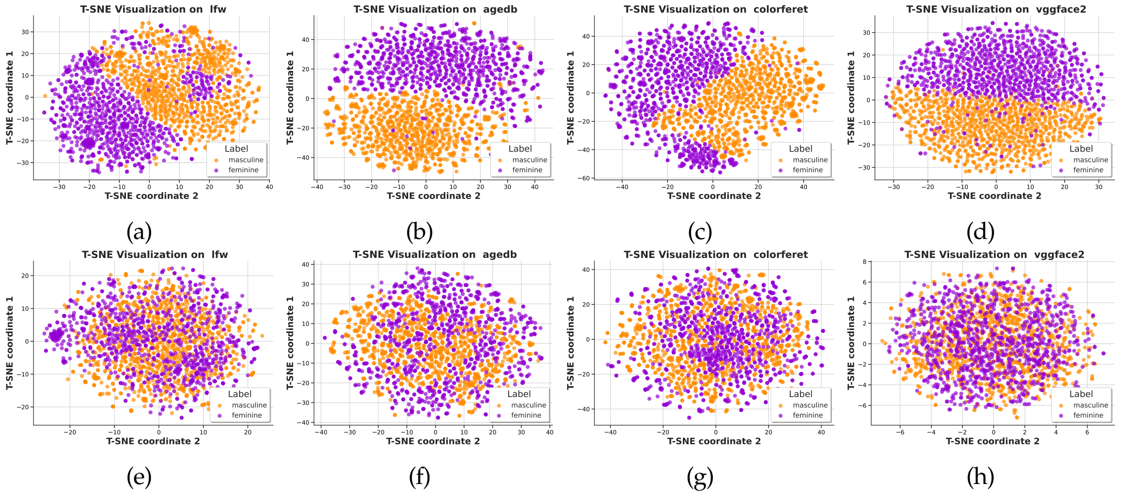
The advantages of our method are that it is geometrically interpretable, easy-to implement, effectively transforms gender-discriminative features to gender-neutral representations while upholding an acceptable verification performance with the same dimensionality. A fundamental point of distinction from prior methods involving finetuning [11], is that we do not depend on a given gender classifier during training which lowers the computational burden of training of our method and makes it independent of a specific decision boundary. We make the source code of our implementation publicly available[3].

We summarize our contributions as the following:

1) We propose light-weight and geometrically interpretable constraints to enhance the privacy of face recognition systems.
2) Our method is designed to align gender distributions, thus fortifying the templates against

[2] What we refer to as gender is the perceived gender expression ”masculine” or ”feminine” that is attributed to facial images by annotators of the cited datasets.

[3] https://github.com/ZohraRezgui/genderPrivacy



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 2: T-SNE visualizations on the original embeddings of ArcFace (first row) and after applying our method with the loss $20L_{p_1} + 1L_{p_2}$ trained on the AgeDB dataset (second row). Every column corresponds to the source dataset of the embeddings.

any potential exploitation to train a robust gender classifier. It is specifically tailored to impede the capacity of unforeseen classifiers to learn any gender-related features.

3) Our method does not require a gender classifier during training and focuses instead on the training of a shallow network that imparts minimal additional computational load to the initial face recognition network.
4) We provide comprehensive cross-database results using various face recognition models and evaluate the effect of the training data on the performance of the method to an extent that was rarely observed in previous work.

2 Related work

Most of the privacy-enhancing techniques that focus on the face modality are image-based. The earliest methods rely on fusing specific parts of a face or morphing faces from different categories of the soft biometric attribute [18], [19], [20]. In [18], they determine the most gender-discriminant face components and use image fusing to choose the closest facial components from the opposing gender for a particular subject. Likewise, [19] perform a transformation of the gender expression of the facial image as a privacy method. They use a spectrum of morphing parameters to generate numerous versions of the input image with varying gender confidence levels.

Other methods rely on adversarial perturbations to fool gender classifiers into making wrong predictions without fooling face recognition systems [20], [21]. For instance, [21] show that gradient-based adversarial attacks on a gender classifier are in some cases not transferable to face recognition systems. Therefore, the images are perturbed to result in a false gender prediction of a gender classifier with imperceptible distortion of the images and negligible decrease of verification performance of face recognition systems. Other works such as [22], [23], [24], [25], [26] use GANs to alter the appearance of the facial image making it imitate characteristics of a different category of the soft biometric attribute. However, while these methods perform well on facial images, they do not necessarily work on the template level as shown in [11].

Therefore, a few works emerged that focus on enhancing soft biometric privacy on the template level [11], [12], [13], [14], [27]. The authors in [11] finetune a face recognition model by training a 3-layer network with a modified triplet loss that incorporates constraints aiming to fool an adversary ethnicity or gender classification layer that is also trained in parallel to the face recognition model. This makes the representation learning dependent on the convergence state of the adversary classification network in different stages of the training.

More recently, [13] proposed an approach based on shuffling blocks of the information encoded in the templates. At the moment of template comparison, shuffled references and probes are realigned based on the Hungarian algorithm. While this method has a more general approach to privacy, it does not tackle the learned bias in the face recognition system that generates the templates and therefore is considered a data protection approach instead of a data minimization approach.

The methods in [14] and [12] implement data minimization based on the identification then suppression of sensitive information. They allow control on which type of attribute to protect and the amount of information that can be suppressed from the templates. Both methods are easily reproducible, are based on an intuitive approach and provide competitive results on the privacy aspect.

While both *data minimization* and *data protection* approaches aim to hinder the retrieval of soft biometric information, the data protection approach does not aim to eliminate such information from the templates. Instead, it performs operations on the templates to block the access to such information. On the other hand, the data minimization approach aims to effectively remove the sensitive information from the stored templates [12], [28].

In this paper, we compare our method to [12], [14], both data minimization methods like our solution. Additionally, we compare to [13], a data protection method. We describe these methods in detail in Sections 2.1, 2.2 and 2.3 respectively.

2.1 IVE: Incremental Variable Elimination

The incremental variable elimination (IVE) algorithm introduced in [10] is based on estimating feature importance for the classification of the targeted attribute. Feature importance is estimated with the decrease in node impurity measures for tree-based classifiers. Following that, the most important features for the classification of the targeted attribute are iteratively eliminated.
The authors claim to suppress features that were discriminative for gender as well as age. While the method is intuitive, its main drawback is the significant loss of information due to reducing the dimension of the templates. This impacts negatively on the utility of the templates for their intended verification task. Indeed, in order to achieve an acceptable level of sensitive attribute suppression, they report that 400 to 500 features had to be eliminated out of 512 features resulting in an equal error rate (EER) 4 times higher on the training data.

In this paper, we executed the IVE algorithm on different datasets to suppress gender and we report cross-database results for comparison to our method.

2.2 Multi-IVE

The authors in [12] propose an improvement on the IVE algorithm. Instead of eliminating the features from the original feature space of the face recognition model, they first perform a transformation of the features by projecting them on the domains generated from a principal component analysis (PCA) or an independent component analysis (ICA). This modification allows maintaining the same dimension of the original embeddings. The feature suppression based on feature importance estimation is then performed on this new domain. Finally, the feature vectors are projected back onto the original domain with inverting the PCA or ICA. This way, the dimension of the feature space remains the same. They employed two settings when applying this method. The first setting does not exclude any principal component in the PCA/ICA from the elimination process. The second setting tries to indirectly minimize loss in verification performance by locking the first k principal components $k = (3{,}5)$. Furthermore, they adapt the IVE algorithm to suppress three soft biometric attributes (gender, age and ethnicity) simultaneously. However, to compare with our results, we run Multi-IVE solely to suppress gender and we do not consider the other attributes as it can bias the performance of the algorithm on the gender suppression.

While both methods are intuitive approaches, they are not explicitly trained to maintain a high verification performance. We instead train the embeddings with a recognition loss and a privacy loss simultaneously. We also note that we train and evaluate both the IVE and Multi-IVE methods rigorously across multiple datasets and report the results on datasets that are completely independent from the training dataset unlike in [10].

2.3 PE-MIU

In [13], the authors introduce the privacy-enhancing minimum information units method (PE-MIU). This method does not require training. It is based on partitioning a feature vector into several blocks that are then randomly shuffled. The sensitive information is not removed from the templates but rendered inaccessible due to the random order of the blocks. The block size parameter controls the trade-off between privacy and verification performance. The authors report that using a block size of 16 features provides nearly perfect privacy for gender while having minimal effect on verification performance. During comparison, the Hungarian algorithm is used to assign an order to the blocks of the probe that is similar to the blocks of the reference. For mated comparisons, the block assignment is often successful and results in a high similarity score. For non-mated comparisons, the blocks are not assigned correctly but this often results in a low similarity score which is suitable. Due to the step of block assignment, the time needed for comparisons is considerably longer than on unprotected templates. We run this method across multiple datasets using a block size of 16 features which results in 32 blocks for every template. Furthermore, we include run-time evaluations for all methods in Section 5.3 to assess their suitability for real-world applications.

3 Proposed Method

The templates that are generated by state-of-the-art face recognition systems suffer from a strong entanglement between identity-relevant information and gender information. In the first row of Figure 2, a t-distributed stochastic neighbor embedding (T-SNE) plot of the features obtained from an ArcFace model [4] reveals a high gender separability across several datasets. While in many ways the presence of gender information can facilitate the identity verification, it is prone to resulting in different distributions by gender causing unfair decisions for one of the categories. Furthermore, the presence of gender information can be exploited by a privacy adversary to infer the gender attribute on a large scale.

3.1 Adversary's Knowledge

As illustrated in Figure 1, we consider the scenario where the privacy adversary gains access to the feature extractor that generated the templates and has access to the templates. We also suppose that the adversary has access to a large dataset of facial images annotated with gender labels. The adversary would then use this dataset to train a set of gender classifiers that we describe in Section 4.3. Finally, they would feed the templates to these classifiers in order to retrieve the gender predictions. The goal of gender privacy-enhancing methods on the template level is to reduce the separability of the gender. The adversary with the ability to create labeled templates, should not be able to train a reliable gender classifier that can be used to infer the gender from the stored templates. Gender information in the templates is either removed through data minimization or made inaccessible through data protection after privacy-enhancing methods are applied. In this case, the templates become inadequate for accurate classification or for training an effective gender classifier. Successful implementation of the privacy-enhancing method should result in the balanced accuracy of the gender classifier approaching 50%, equivalent to the performance of a random binary classifier. We propose a training method that is based on a constrained recognition loss. In the following sections, we first introduce the trainable parameters then the composition of the proposed loss.

3.2 Architecture and training parameters of the gender privacy-preserving layers

For training our method, we feed the embeddings extracted from a pre-trained face recognition network into a shallow network consisiting of two fully connected layers as presented in Figure 3 in order to obtain the private features. We use a Leaky ReLU as an activation function between these two layers to allow for a non-linear projection of the embeddings. Afterwards, the embeddings are passed into a last fully connected layer to estimate the class identity weights necessary to calculate the logits for the recognition loss. This layer does not have any bias parameters accordingly to the losses described in the Section 3.3.

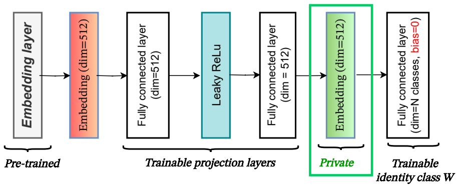

Embedding layer
Embedding (dim=512)
Fully connected layer (dim=512)
Leaky ReLu
Fully connected layer (dim = 512)
Embedding (dim=512)
Fully connected layer (dim=N classes, bias=0)
Pre-trained
Trainable projection layers
Private
Trainable identity class W

Fig. 3: Illustration of the trainable layers.

3.3 Composition of the training loss

The proposed loss has three components. The first component is a normalized softmax loss that has an objective of maintaining the recognition performance.

The two other components are weighted privacy constraints minimizing angles between gender-centroid features as well as angles between gender-centroid features and their opposed gender-centroid identity class weights. We dissect the formulation of these three loss components in the following subsections.

3.3.1 Normalized softmax loss

This component takes the role of ensuring that the recognition performance stays high despite the privacy constraints. Following the modifications to the softmax function outlined in [29], the feature vectors obtained from the embedding layer of a face recognition network preceding the calculation of the logits, lay on a hypersphere. The aforementioned procedure is accomplished by imposing that bias $b_j = 0$ in that layer and $l2-$normalizing both the learnable class weights $||W_j|| = 1$ as well as the feature vectors $||x_i|| = 1$. This step guarantees that the features lay on the unit hypersphere of a given dimension d which in its turn, allow a straightforward calculation of cosine similarity between the feature vectors and their corresponding identity class weights via their inner product.

In [30], the authors propose the following improvement. The feature vectors are scaled to a fixed number s after normalization. This loss is equivalent to ArcFace [4] with margin $m = 0$. The following equation describes the cross-entropy loss to be minimized with this addition.

$$L_r = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{\sum_{j=1}^{n}e^{s\cos\theta_j}} \tag{1}$$

3.3.2 Our privacy constraints

Building on the improvements in [4], [30] for the recognition loss, we formulate two angular constraints to ensure that the gender distributions overlap. The ensemble of the angles we minimize during training is illustrated in Figure 4.

In order to confuse a gender classifier, the feature distributions of each category have to be as similar as possible. The first constraint L_{p_1} involves solely the embeddings. The distance between the distributions is approximated by the angle between the average feature vectors in each category. Eventually, this angular distance is minimized during training. For every batch, we select separately the feature vectors for each gender category. After l_2- normalization, we calculate the average feature vectors for the masculine category $\overline{x}_m$ and for the feminine category $\overline{x}_f$ and then perform $l2-$ normalization. As a final step, we calculate the angle θ_{p1} between these two vectors via the $\arccos$ function on their inner product. In order to improve estimation of the mean vectors per gender, the batches are gender-balanced. The pseudo-code to calculate the following L_{p_1} is given in Algorithm 1.

$$L_{p_1} = \theta_{p1} = \arccos(\langle\overline{x}_f, \overline{x}_m\rangle) \tag{2}$$

The second constraint L_{p_2} involves both the embeddings and the weights of the identity classes. As the normalized softmax loss L_r is minimized during training, the identity class weights and the embeddings are updated such that each embedding forms the smallest angle possible with the identity class weight vector corresponding to its ground truth identity. The constraint L_{p_2} is added to guide the updates of the identity class weights and the feature vectors simultaneously by enforcing that the average masculine identity vector gets as close as possible to the average feminine feature vector and vice versa.

For every batch, we select separately the feature vectors for each gender category and we also select separately the l_2- normalized weights of the identity classes associated with each gender category. We calculate the average identity weight vector for each gender category and then l_2-normalize it to bound it to the surface of the unit hypersphere. Finally, we compute the angles between every feature vector and the average identity vector associated with its opposite gender category. The averages of these angles θ_{p_2} and θ_{p_3} are then minimized during training. We provide the pseudo-code for calculating the following constraint in Algorithm 2.

$$\begin{aligned} L_{p_2} = &\frac{1}{n_f}\sum_{i=1}^{n_f}\arccos(\langle x_{if}, \overline{W}_m\rangle) \\ &+\frac{1}{n_m}\sum_{i=1}^{n_m}\arccos(\langle x_{im}, \overline{W}_f\rangle) \end{aligned} \tag{3}$$

The final loss to be minimized during training is given in the equation below with α and β as hyperparameters for increasing or decreasing the magnitudes of the privacy constraints:

$$L = L_r + \alpha L_{p_1} + \beta L_{p_2} \tag{4}$$

4 Experiment settings

4.1 Datasets

In [10] and [11], the authors use one dataset for training and evaluating their privacy-preserving approach. In

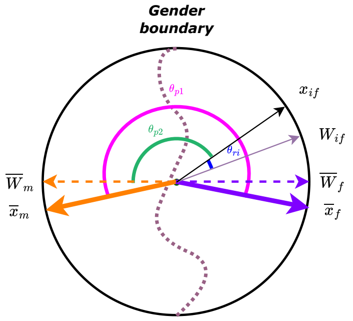


Fig. 4: 2D Illustration of the different angles on the hypersphere that we minimize during training. x_i refers to the feature vector of a sample i, W_i refers to the weight vector corresponding to its identity class, $\overline{W}_m$ and $\overline{W}_f$ correspond to the average weight vectors for the masculine and feminine identity classes respectively and $\overline{x}_m$ and $\overline{x}_f$ correspond to the average feature vectors for each gender category respectively.

Algorithm 1 Calculate privacy constraint L_{p_1}

Input:

i_m,i_f: male and feminine indices,
n_m,n_f: sizes of male and feminine samples,
X: feature matrix of shape (n,d) with n: batch size and d: feature space dimension

Output: $L_{p_1} = \theta_{p_1}$, the angle between the average feminine feature vector and the average male feature vector.

```
$\mathbf{x_f} \leftarrow \frac{1}{n_f}\sum_{i=1}^{n_f}\mathbf{X}_{if\cdot}$
$\mathbf{x_m} \leftarrow \frac{1}{n_m}\sum_{i=1}^{n_m}\mathbf{X}_{im\cdot}$
$\overline{\mathbf{x_f}} \leftarrow \frac{\mathbf{x_f}}{\|\mathbf{x_f}\|_2}$
$\overline{\mathbf{x_m}} \leftarrow \frac{\mathbf{x_m}}{\|\mathbf{x_m}\|_2}$
$L_{p_1} = \theta_{p_1} \leftarrow \arccos(\langle\overline{\mathbf{x_m}},\overline{\mathbf{x_f}}\rangle)$
**return** $L_{p_1}$
```

Algorithm 2 Calculate privacy constraint L_{p_2}

Input:

i_f,i_m: Indices of feminine and masculine feature vectors,
j_f,j_m: Indices of feminine and masculine identity weight vectors,
n_f,n_m: Sizes of feminine and masculine samples in the batch of feature vectors,
k_f, k_m: Numbers of feminine and masculine identities,
X: Feature matrix of shape (n,d) with n: batch size and d: feature space dimension,
W: Identity class weight matrix of shape (m,d) with m: number of identity classes and d: feature space dimension

Output: L_{p_2}: Privacy constraint based on angles θ_{p_2} and θ_{p_3} between the feature vectors and the centroid identity weight vector of the opposing gender category.

```
$\mathbf{W_f} \leftarrow \frac{1}{k_f}\sum_{j=1}^{k_f}\mathbf{W}_{j_f\cdot}$
$\mathbf{W_m} \leftarrow \frac{1}{k_m}\sum_{j=1}^{k_m}\mathbf{W}_{j_m\cdot}$
$\overline{\mathbf{W}}_f \leftarrow \frac{\mathbf{W_f}}{\|\mathbf{W_f}\|_2}$
$\overline{\mathbf{W}}_m \leftarrow \frac{\mathbf{W_m}}{\|\mathbf{W_m}\|_2}$
$\theta_{p_2} \leftarrow \frac{1}{n_f}\sum_{i=1}^{n_f}\arccos(\langle\overline{W}_m^T,\mathbf{X}_{if\cdot}\rangle)$
$\theta_{p_3} \leftarrow \frac{1}{n_m}\sum_{i=1}^{n_m}\arccos(\langle\overline{W}_f^T,\mathbf{X}_{im\cdot}\rangle)$
$L_{p_2} \leftarrow \theta_{p_2} + \theta_{p_3}$
**return** $L_{p_2}$
```

[12], the authors use separate datasets to train and evaluate their method however, they do not evaluate the verification task and the privacy task simultaneously on the evaluation dataset. Instead, they pick one dataset to evaluate the gender suppression and another to evaluate the verification performance. To have a large overview of the generalization ability of our method and the methods from [10], [12], [13] we reproduce, we use four facial datasets. The following datasets are alternated for training and evaluation and all evaluations are performed simultaneously for the verification and the privacy tasks: The Labeled Faces in the Wild (LFW) dataset [31] consists of 13,233 images in unconstrained conditions of 5,749 identities. AgeDB [32] contains 16,516 images of 570 identities in uncontrolled conditions with a large variation in age. ColorFeret [33] contains 11,338 images of 994 identities collected under controlled conditions. We also randomly select a gender-balanced sample of 15,000 images from the VGGFace2 train set [34] of 5,000 identities.

The samples from LFW, AgeDB, and VGGFace2 encompass a wide array of images taken in real-world conditions, exhibiting substantial variability in terms of pose, lighting, and demographic characteristics, including age, ethnicity, and gender expression.

The faces of these images are detected and aligned with an MTCNN face detection algorithm [35] and finally they are resized to 112x112 pixels. We give in Table 1 the total number of images and gender distribution for these datasets. We alternate using these datasets to train the privacy-enhancing methods and evaluate the performance on the remaining ones. We note that in contrary to what is reported in [10] and [11], whenever we use one of these datasets for training a method, we do not include in the evaluation set to avoid bias that might be resulting from overfitting.

4.2 Pre-trained face recognition models

We run our experiments on three state-of-the art face recognition models, namely ArcFace[4] [4], SphereFace[5] [15], [17] and ElasticFace[6] [16] that are trained using distinct angular losses. The ArcFace model used is trained on the VGGFace2 dataset [34] with an IResNet50 architecture while the SphereFace and ElasticFace models used are both trained on the MS-Celeb-1M dataset [36] with an IResNet100 architecture. All of the models take images of 112x112 pixels and output 512-dimensional embeddings.

4.3 Gender classifiers

To evaluate the gender separability of the embeddings before and after applying privacy-enhacing methods, we use a 3-fold cross-validation setting. For each dataset used for evaluation, we form 3 folds where in each fold, the train set and the test set do not have overlapping identities. For every fold, an ensemble of gender classifiers is trained on the train set and evaluated on the test set. This ensemble of gender classifiers consist of two linear classifiers, namely a linear SVM and a logistic regression, and one non-linear classifier, an SVM with an RBF kernel. The composition of the folds in terms of number of images and number of identities is given in Table 1.

4.4 Evaluation metrics and model selection

We use the average balanced accuracy that we refer to as ACC_G of the gender classifiers across the 3-folds to evaluate the gender classification performance. The balanced accuracy is defined as:

$$ACC_G = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5}$$

where the numbers TP, TN, FP, FN refer to true positive, true negative, false positive and false negative predictions respectively. We report the details regarding the partition of the folds in Table 1.

In order to have results that describe reliably the impact of privacy enhancing techniques on verification, all verification evaluations are performed following the standard protocole 1 for benchmark on the LFW dataset in [37] where 6,000 pairs (3,000 mated and 3,000 non-mated) are compared using the Euclidian distance. The same procedure was used in [32] to create age-invariant verification protocols. The most challenging protocol is the one where the pairs have 30 years of age difference (AgeDB-30). This protocol is the most widely used to report verification performance on the AgeDB dataset [4] therefore we use it to guarantee comparability. Similarily, we use the same protocol to generate the verification pairs (6,000 pairs) for the other datasets (ColorFeret and the sampled VGGFace2). We choose the equal error rate (EER) to report verification performance, that we refer to as EER_V.

To select the models with the best trade-off between the two tasks we use the privacy gain (PG) - identity loss (IL) criterion (PIC) introduced in [38]:

$$PIC = PG - IL \tag{6}$$

$$\text{with} \begin{cases} PG &= \dfrac{(1-ACC_G)-(1-ACC^*_G)}{(1-ACC^*_G)} \\ \\ IL &= \dfrac{EER_V - EER^*_V}{EER^*_V} \end{cases} \tag{7}$$

with the couples $(ACC^*_G,\ EER^*_V)$ and $(ACC_G,\ EER_V)$ designating the gender classification and verification performances on the embeddings before and after the privacy-enhancing method respectively.

The higher PIC gets, the better privacy-utility trade-off we obtain. In the case where the identity loss is greater than the privacy gain, this metric yields negative values. We note however that this metric calculates relative improvements in privacy and face verification performances with regards to the original embeddings. Therefore, if the original embeddings are not highly discriminative for gender or obtain near perfect verification performance, the metric is likely to yield negative values even if the obtained privacy and verification performances are satisfactory.

5 Experiments

For IVE, we ran the method with various total number of eliminations from the feature vectors and with different training datasets. The total number of

[4] https://github.com/deepinsight/insightface
[5] https://github.com/ydwen/opensphere
[6] https://github.com/fdbtrs/ElasticFace

Dataset		**Total**	**Train**			**Test**		
			Fold 1	**Fold 2**	**Fold 3**	**Fold 1**	**Fold 2**	**Fold 3**
ColorFeret	N(% feminine)	11 286 (35.39)	7 524 (35.39)	7 525 (35.40)	7 523 (35.37)	3 762 (35.38)	3 761 (35.36)	3 763 (35.42)
	N_{ids} (% feminine)	994 (40.54)	663 (40.57)	663 (40.57)	662 (40.48)	331 (40.48)	331 (40.48)	332 (40.66)
LFW	N(% feminine)	13 233 (22.42)	8 822 (22.42)	8 822 (22.42)	8 822 (22.42)	4 411 (22.42)	4 411 (22.42)	4 411 (22.42)
	N_{ids} (% feminine)	5 749 (17.13)	3 836 (25.78)	3 831 (25.68)	3 831 (25.71)	1 913 (25.72)	1 918 (25.70)	1 918 (25.72)
AgeDB	N(% feminine)	15 698 (40.76)	10 436 (40.71)	10 462 (40.76)	10 498 (40.81)	5 262 (40.86)	5 236 (40.76)	5 200 (40.65)
	N_{ids} (% feminine)	567 (38.80)	377 (38.73)	379 (38.79)	378 (38.88)	190 (38.95)	188 (38.83)	189 (38.62)
VGGFace2 subset	N(% feminine)	15 000 (50.00)	9 994 (50.85)	9 991 (50.87)	9 993 (50.86)	4 996 (50.86)	4 999 (50.83)	4 997 (50.85)
	N_{ids} (% feminine)	5 000 (50.00)	3 334 (52.97)	3 333 (52.99)	3 333 (52.99)	1 666 (53.00)	1 667 (52.97)	1 667 (52.97)

TABLE 1: Overview of the number of images N and identities N_{ids} in total and in the folds setting used for the evaluation of gender classification performance.

eliminations ranges from only 20 eliminated features to 500. For each training set, we selected the resulting elimination algorithm that provided the highest PIC value. Across the training sets used, the highest PIC values correspond to a total elimination of 500 features from the ArcFace templates compared to an elimination of 400 to 500 features from ElasticFace templates and 300 to 400 features from SphereFace templates depending on the training set used.

Similarly for Multi-IVE, we varied the type of intermediate transformation domain (PCA or ICA), the total number of eliminations as well as the number of locked principal components in the transformation domain ($k = 0, 3, 5$). For each training set used, we select parameters that correspond to the highest PIC value. In all cases, the highest PIC value was associated to a total of 120 eliminations in the PCA domain with $k = (3{,}5)$ for the ArcFace templates. For the ElasticFace templates, the optimal number of eliminations ranges from 276 to 432 eliminations in the PCA domain with $k = (3{,}5)$. As for the SphereFace templates, 81 to 354 eliminations are optimal in in the PCA domain with $k = (0, 3, 5)$ depending on the training set. For all models, more eliminations come with an even higher expense on the verification performance.

As for PE-MIU, we run it and report all the results using a block size of 16 features resulting in templates of 32 blocks. When it comes to our proposed losses, we minimize them by training the privacy finetuning layers for 100 epochs with a learning rate of 0.01. The scale factor for the recognition loss L_r is set to $s = 64$. A batch size of 128 images is used with a roughly balanced number of images per gender. When it comes to the privacy weight factors α and β, we set $\alpha = 20$ as it gives a magnitude to L_{p_1} that is comparable to that of L_r and set $\beta \in \{0{,}1\}$. We note that higher values of β resulted often in convergence problems. We also varied the training sets, each of the datasets is used as a training set and we also formed pair combinations of datasets LFW, ColorFeret and AgeDB. When it comes to model selection, we evaluate the performance of all the saved parameters every 10 epochs and we select the model that is associated with the highest PIC value.

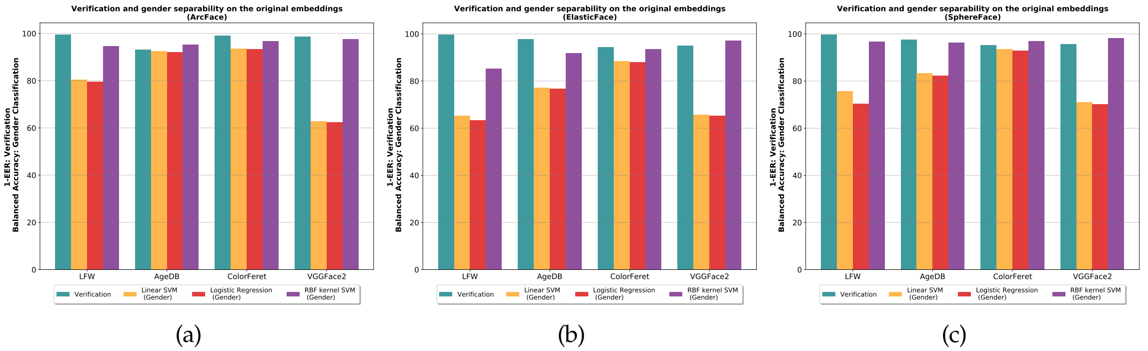
Before applying any privacy-preserving technique, we investigated the original embeddings when they are extracted from different datasets. We see from Figure 5 that the features obtained from images in the VGGFace2 dataset are not linearly separable regardless of the feature extractor used as ACC_G does not exceed 71% for linear SVM and logistic regression classifiers but are highly separable with a non-linear classifier reaching an ACC_G of 98.20% for SphereFace features.

AgeDB and ColorFeret are associated with more linearily separable features, in particular ColorFeret with an ACC_G exceeding 80% using both linear classifiers. The most linearly separable datasets are ColorFeret, followed by AgeDB, then LFW and finally VGGFace2. However, all of the datasets are easily separable using an RBF kernel SVM classifier that achieves a performance ACC_G exceeding 80% in all cases. We can speculate that the differences in separability among the datasets are caused by the different levels with which these dataset distributions vary from the training data distributions of the face recognition systems. In addition to disparities at the dataset level, we also notice that ElasticFace seems to result in less gender separable features across all datasets compared to ArcFace and SphereFace.

These observations indicate that the gender separability of the embeddings vary from one dataset to another as well as from one face recognition model to another and depends on the type of classifier that is used.

5.1 Cross-database evaluation

We illustrate in Figure 6, the gender classification and verification performances of all the methods with



(a) (b) (c)

Fig. 5: Verification performance and gender classification performance using three classifiers (Linear SVM, Logistic Regression and an SVM with an RBF kernel) on the original embeddings from the pre-trained face recognition models. Zooming may be necessary for the best viewing of the plots..

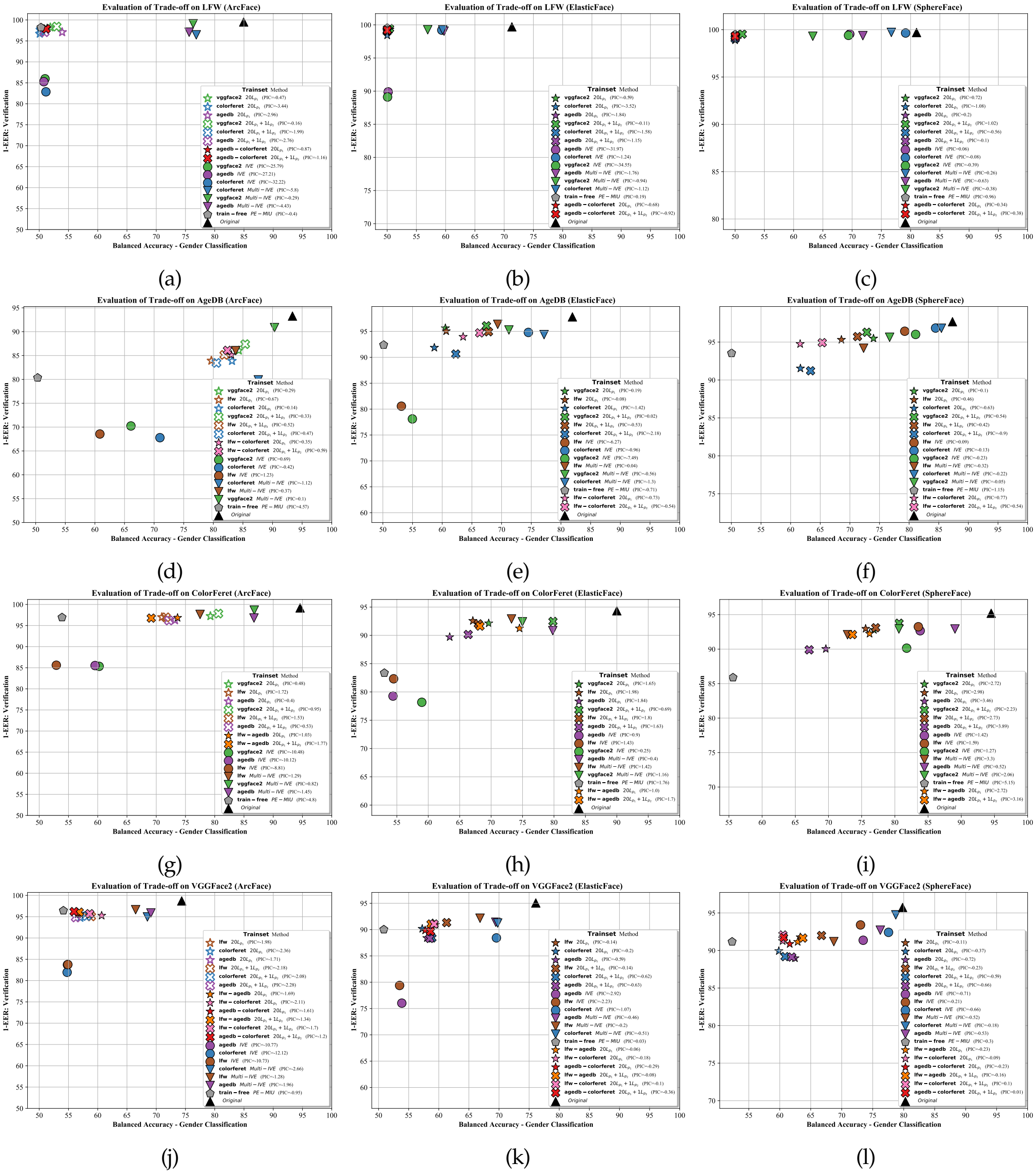
various training sets on the evaluation sets. We exclude the results where the training set of the privacy-preserving method is the same as the evaluation set.

For features extracted using ArcFace, we notice that our methods achieve a near ideal trade-off on LFW and VGGFace2 datasets next to the PE-MIU shuffling approach. On AgeDB and ColorFeret, PE-MIU achieves the best trade-off with ACC_G lower than 55% and an EER_V lower than 20%. On these datasets, our methods have a much lower privacy with our best ACC_G of 69.15% on ColorFeret and of 79.67% on AgeDB while maintaining in all cases a EER_V less than 17% . However, on all four datasets, Multi-IVE and IVE result in either a worse privacy level with a comparable verification performance to our methods or a better privacy with a significantly deteriorated verification performance. For instance, on the AgeDB dataset, IVE achieves an ACC_G of 60.84% but with an EER_V of 31.45% which is better than ours in terms of privacy but hinders extremely the verification task.

For ElasticFace features, we notice a similar ideal trade-off on the LFW dataset for PE-MIU and our methods where the best of our methods achieves an ACC_G of 50.41% with an EER_V of 0.5%. PE-MIU achieves a similar level of privacy but with an EER_V of 0.46%. For the remaining datasets, our methods are more effective than Multi-IVE in terms of privacy and less effective than PE-MIU and IVE but tend to achieve higher verification performance. For instance, on ColorFeret, PE-MIU has a near total privacy with an ACC_G of 53.01% and an EER_V of 16.65% which is 3.47 times higher than the original EER_V of 4.8% while our best method achieves an ACC_G of 63.41% with an EER_V of 10.29%. However, we note that despite not being consistently the best at enhancing privacy all our best methods achieve consistently an ACC_G less than 65% across all datasets while maintaining an EER_V equal or lower than 10.29%.

As for the SphereFace features, similarly to ArcFace and ElasticFace features, our methods are as successful in achieving an ideal trade-off on LFW as the data protection approach PE-MIU. Both sets of methods achieve nearly an ACC_G of 50% with only a negligible deterioration of verification performance. On the other hand, IVE and Multi-IVE best privacy results achieve an ACC_G of 69.38% and 63.28% respectively. In the remaining datasets, our methods best results supersede IVE and Multi-IVE while maintaining an EER_V equal or lower than 10% but they are superseded by PE-MIU which in the case of ColorFeret, achieves an ACC_G of 55.61% but results in an EER_V of 14.12%. Nevertheless, our methods best results consistently achieve an ACC_G lower than 68% across all datasets.

The results on ColorFeret and AgeDB show that for these two datasets, identity and gender tend to be highly entangled, in particular for ColorFeret. It is more difficult in the case of these datasets to remove gender information without severely impacting the verification performance. In contrary to the other data minimization approaches, our method includes a recognition loss term L_r in the training loss that explicitly forces the network to retain the verification performance from decreasing significantly. Multi-IVE implicitly attempts to maintain the identity-relevant



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

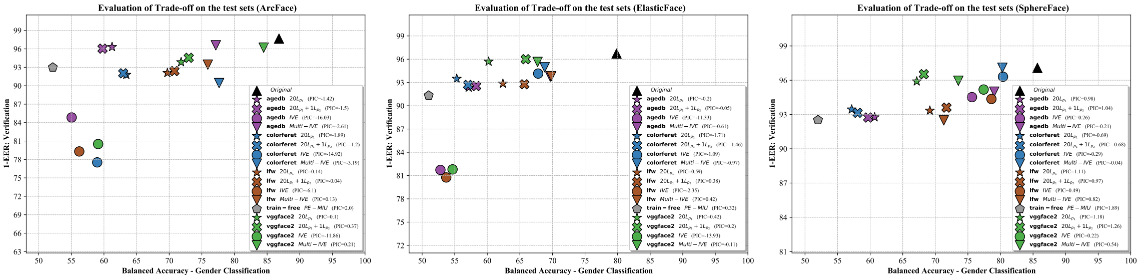
(i)

(j)

(k)

(l)

Fig. 6: The above plots summarize the effect of the training set with different methods on the privacy-utility trade-off evaluated on different datasets. The training set appears in bold in the legend preceded by the loss or method used. Each column corresponds to a distinct face recognition model. Zooming may be necessary for the best viewing of the plots.



(a) (b) (c)

Fig. 7: Figure showing the average performance per training set and method on the remaining evaluation sets. Zooming may be necessary for the best viewing of the plots.

information in the embeddings by excluding a number of principal components from elimination in the transformed PCA or ICA domain. IVE only executes the privacy-enhancement by suppressing gender-related features. The recognition loss that is included in our method could be the reason why the gender classification performance does not always decrease as drastically as with the other methods, due to the high entanglement between gender and identity for certain datasets such as ColorFeret.

We also note that overall, the choice of the training data has an impact on the performance of the methods. Using our methods, training on ColorFeret and AgeDB has the tendency to produce better results on each other while VGGFace2 seems to be the least suitable training data for our methods in terms of privacy. Combinations of datasets are in some cases beneficial as the combination of LFW and AgeDB when evaluating on the ColorFeret dataset or the combination of AgeDB and ColorFeret when evaluating on VGGFace2 dataset for the ArcFace features.

In Figure 7, we plotted the average verification performance and gender classification on the evaluation sets. We excluded the combination of training sets used in our methods to guarantee that all methods appearing in the figure share the same training set and evaluation sets. For our methods, the gender classification performance varies from an averageACC_G of 55.31% to 72.99% with an average EER_V consistently below 10% ranging between 3.47% and 8.21% across all feature extractors. When it comes to IVE, the performances vary depending on the training sets with a severe impact on verification performance. It achieves an average ACC_G ranging from 52.81% to 80.39% with an average EER_V from 3.7% to 22.47%. Multi-IVE achieves worse results from a privacy point of view but tends to have a higher verification performance when using most training sets. Its achieves an average ACC_G ranging from 67.72% to 84.47% and an average EER_V ranging from 2.91% to 9.57%. Finally, for PE-MIU, it achieves near optimal privacy results with an average ACC_G consistently below 55% and its average EER_V ranges from 7.02% to 8.67%.

We retain that our methods are able to minimize the impact on the verification performance while obtaining a significant privacy gain compared to methods Multi-IVE and IVE. IVE only tackles the privacy aspect and therefore, causes a substantial loss in verification performance while Multi-IVE results in limited privacy due to limiting the elimination of a number of principal components in the PCA domain. While PE-MIU supersedes our methods based on the reported performances, it is crucial to note that PE-MIU has a higher risk of being compromised due to the fact that the sensitive information is not removed as is the case in IVE, Multi-IVE and our methods. Instead, the sensitive information is only shuffled.

To qualitatively assess the performance of our method, we show in Figure 2 the t-distributed stochastic neighbor embedding (T-SNE) 2-D visualizations of our feature vectors after training our method with $\alpha = 20$ and $\beta = 1$. We used the best model with the ArcFace backbone trained on the AgeDB dataset as it gives the best average trade-off on all the remaining datasets. We can see from the Figure that the overlap of the gender distributions is apparent in the second row compared to a clear separation of the gender distributions in the first row.

5.2 Sensitivity analysis of the privacy factor α

α	11	14	17	**20**	23	26	29
Linear SVM	0.8532	0.1544	**0.0016**	-	0.5347	**8.0052e-06**	0.0575
LogReg	0.4926	0.5104	**0.0396**	-	**0.0023**	**5.1781e-08**	**0.0172**
RBF SVM	0.4671	0.0878	**0.0098**	-	0.5614	**8.7602e-04**	0.0529
ACC_G (%)	81.55	81.82	**82.79**	81.50	81.09	**80.46**	81.14

TABLE 2: Sensitivity analysis of the α privacy factor. The values in the rows 1-3 correspond to the p-value of the Wilconxon signed test with the null hypothesis that the predictions from the classifiers are not different when $\alpha = 20$ and when α taking the values presented in the column headers. The last row corresponds to the ACC_G averaged across the classifiers.

To see the impact of the α parameter on the privacy gain, we performed a sensitivity analysis with $\alpha \in \{11,14,17,23,26,29\}$. We did a Wilcoxon signed test to compare the performance of the gender classifiers when $\alpha = 20$ and when α takes values from the set.

The statistical test compares the predictions of the gender classifiers on the AgeDB dataset with ArcFace features. We chose to perform this experiment on AgeDB as it is the most challenging dataset in terms of privacy gain for ArcFace features. All the models generating the features are trained on the same dataset ColorFeret. We report the p-values of the Wilcoxon signed per classifier and value of α as the ACC_G averaged across all classifiers in Table 2. We notice that the differences are significant for $\alpha = 17$ and $\alpha = 26$ with p-values < 0.05. The ACC_G is the highest for $\alpha = 17$ and is the lowest for $\alpha = 26$. This shows that α is a sensitive parameter despite not consistently resulting in a significant privacy gain when it increases. This could be explained by the fact that the composed loss has two main components; the recognition component and the privacy component that are two tasks competing against each other.

5.3 Computational time analysis

We assess the suitability of the aforementioned methods in a real-world situation by quantifying the computational time required for generating privacy-enhanced templates and executing comparisons. The template generation step includes pre-processing for the facial image, the running the original feature extractor and applying the privacy enhancement method to obtain the final template. The comparison step refers only to the computation of the Euclidian distance between two templates.

These runtime measures are performed on the ArcFace features, using consistently the best model for each privacy-enhancing method. We note that for IVE, the privacy-enhanced template has only 12 features. We report such measures in Table 3 where we can see that the average time to generate the template is approximatively the same for all methods, except for Multi-IVE where it is roughly 2.3 times slower than the other methods. This is due to the complicated steps in Multi-IVE that require at least three different steps next to the generation of the pre-privacy template, namely a projection onto the transformed domain, elimination of sensitive features (120 eliminations) in the transformed domain then a reverse projection onto the original domain.

We note that despite the additional layers that we train on top of the initial templates, our method does not add a significant computational burden in order to obtain the privacy-enhanced templates. When it comes to the time needed for computing one-to-one comparisons using Euclidian distance, we see that PE-MIU stands out as a substantially heavy method. Comparison between IVE templates is slightly faster than comparison between original templates due to the reduced size of the templates after IVE.

While the remaining methods have similar computational time for template comparison to that of the original templates, PE-MIU is 1444 times slower. This is largely due to the assignment of the blocks between the reference and the probe, which is a crucial part to calculate reliable comparison scores, especially for mated pairs. However, due to its significant computational load during comparison, it is unlikely that PE-MIU can be implemented in a practical application. This is especially true in situations where one-to-many comparisons are necessary for the purpose of identification.

6 Conclusion

In this paper, we finetune a face recognition system with the aim to enhance gender privacy in facial templates. We propose two constraints that act on both the gender-specific feature vectors and the learnable identity class weights. These constraints are intuitive and take advantage of the hyperspherical nature of the feature space in state-of-the-art face recognition systems and are effective for training a shallow network on top of the embedding layer of a pre-trained face recognition model. Our findings demonstrate that the inclusion of said constraints significantly

Method	Template generation	Comparison score
IVE	67.4	0.0073
Multi-IVE	**155**	0.0106
PE-MIU	67.4	**14.3**
Ours	67.5	0.0101
Original	67.4	0.0099

TABLE 3: Average computational time in **milliseconds (ms)** of privacy-enhancing methods using ArcFace as the pre-privacy feature extractor. Runtime is estimated with the *timeit* Python module and is averaged over 1000 iterations. These measures are run on an **Intel® Core™ i7-10750H CPU with 2.60GHz**.

improves privacy while preserving the face verification performance with no additional computational burden unlike other methods. We also highlight the impact of the choice of the training data for privacy-enhancing techniques. Additionally, we provide an extensive evaluation protocole that emphasizes the importance of performing the evaluations on several datasets that were not used for training the privacy-enhancing method. Future work is required to assess further the impact of the separability of the training data on the effectiveness of privacy-enhancing techniques.

Acknowledgement

This work was supported by the PriMa project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 860315.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

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

[3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[5] G. Ozbulak, Y. Aytar, and H. K. Ekenel, "How transferable are cnn-based features for age and gender classification?" in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2016, pp. 1–6.

[6] A. Acien, A. Morales, R. Vera-Rodriguez, I. Bartolome, and J. Fierrez, "Measuring the gender and ethnicity bias in deep models for face recognition," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings 23*. Springer, 2019, pp. 584–593.

[7] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "Beyond identity: What information is stored in biometric face templates?" in *2020 IEEE international joint conference on biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[8] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[9] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.

[10] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Suppressing gender and age in face templates using incremental variable elimination," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[11] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 2158–2164, 2020.

[12] P. Melzi, H. O. Shahreza, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, J. Fierrez, S. Marcel, and C. Busch, "Multi-ive: Privacy enhancement of multiple soft-biometrics in face embeddings," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 323–331.

[13] P. Terhörst, K. Riehl, N. Damer, P. Rot, B. Bortolato, F. Kirchbuchner, V. Struc, and A. Kuijper, "Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units," *IEEE Access*, vol. 8, pp. 93 635–93 647, 2020.

[14] P. Terhörst, M. Huber, J. N. Kolf, I. Zelch, N. Damer, F. Kirchbuchner, and A. Kuijper, "Reliable age and gender estimation from face images: Stating the confidence of model predictions," in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019, pp. 1–8.

[15] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[16] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1578–1587.

[17] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller, "Sphereface revived: Unifying hyperspherical face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[18] 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, 2010.

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

[20] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in *2017 IEEE International joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 564–573.

[21] Z. Rezgui, A. Bassit, and R. Veldhuis, "Transferability analysis of adversarial attacks on gender classification to face recognition: Fixed and variable attack perturbation," *IET biometrics*, vol. 11, no. 5, pp. 407–419, 2022.

[22] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *2018 International Conference on Biometrics (ICB)*. IEEE, 2018, pp. 82–89.

[23] V. Mirjalili, S. Raschka, and A. Ross, "Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers," *IEEE Access*, vol. 7, pp. 99 735–99 745, 2019.

[24] ——, "Privacynet: semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[25] Q. Li, Y. Liu, and Z. Sun, "Age progression and regression with spatial attention modules," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 378–11 385.

[26] Q. Li, W. Wang, C. Xu, and Z. Sun, "Learning disentangled representation for one-shot progressive face swapping," *arXiv preprint arXiv:2203.12985*, 2022.

[27] B. Bortolato, M. Ivanovska, P. Rot, J. Križaj, P. Terhörst, N. Damer, P. Peer, and V. Štruc, "Learning privacy-enhancing face representations through feature disentanglement," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 495–502.

[28] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, "An overview of privacy-enhancing technologies in biometric recognition," *arXiv preprint arXiv:2206.10465*, 2022.

[29] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.

[30] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[31] 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," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[32] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

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

[34] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[35] J. Xiang and G. Zhu, "Joint face detection and facial expression recognition with mtcnn," in *2017 4th international conference on information science and control engineering (ICISCE)*. IEEE, 2017, pp. 424–427.

[36] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 87–102.

[37] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[38] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations," *Applied Intelligence*, vol. 49, pp. 3043–3060, 2019.

Zohra Rezgui received her engineering degree in Statistics and Information Analysis from the University of Carthage (Tunisia) in 2019. Since 2020, she is a Ph.D candidate at the University of Twente (The Netherlands) as part of the PriMa project (Privacy Matters). Her research interests include privacy-enhancing techniques to reduce the risk of biometric profiling in facial images and templates.

Nicola Strisciuglio is Associate Professor of Computer Vision and Machine Learning at the University of Twente (The Netherlands). He obtained a joint Ph.D degree cum laude from the University of Groningen (The Netherlands) and the University of Salerno (Italy), in 2016. He is general co-chair of the APPIS conference series, and has been Program Chair of CAIP 2019. He currently serves as associate editor of Pattern Recognition, and Pattern Recognition Letters. His research interests include robustness and generalization in machine learning and computer vision, with focus on the identification and exploitation of bias.

Raymond Veldhuis graduated from the University of Twente, The Netherlands, in 1981. He received the Ph.D. degree from Nijmegen University on a thesis entitled Adaptive Restoration of Lost Samples in Discrete-Time Signals and Digital Images, in 1988. From 1982 to 1992, he was a Researcher with Philips Research Laboratories, Eindhoven, in various areas of digital signal processing. From 1992 to 2001, he was involved in the field of speech processing. He currently holds part-time professor positions at the University of Twente, Enschede, The Netherlands, and at NTNU, Gjøvik, Norway. His main research topic is machine learning for biometrics, with a focus on face recognition and biometric template protection. The research is both applied and fundamental.