Enhancing Soft Biometric Face Template Privacy with Mutual Information-Based Image Attacks

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

Abstract

The features learned by deep-learning based face recognition networks pose privacy risks as they encode sensitive information that could be used to infer demographic attributes. In this paper, we propose an image-based solution that enhances the soft biometric privacy of the templates generated by face recognition networks. The method uses a reliable mutual information estimation and simulates a minimization step of the mutual information between the features and the target variable. We comprehensively assess the effectiveness of our approach on the gender classification task by formulating two distinct evaluation settings: one for evaluating the performance of the approach's ability to fool a given gender classifier and another for evaluating its ability to hinder the separability of the gender distributions. We conduct an extensive analysis, considering varying levels of perturbation. We show the potential of our method as a privacy-enhancing method that preserves the verification performance as well as a strong single-step adversarial attack.

1. Introduction

Prior research has highlighted the discriminative nature of the features extracted from face recognition networks. These features have proven to be useful for tasks beyond face recognition, such as gender, age, or ethnicity classification, as demonstrated by previous studies [1, 15]. This intertwining of identity information with additional soft biometric attributes within facial templates raises legal and ethical concerns regarding privacy that could lead to profiling and unfairness. To address these concerns, several approaches to enhance the privacy at the template level emerged. Most of these approaches rely on finetuning the parameters of the face recognition network or perform feature suppression and shuffling directly on the templates [10, 19, 20].

We propose an approach, illustrated in Figure 1 that constitutes an adversarial attack on the templates to enhance soft biometric privacy, by perturbing solely the images without any modification of the face recognition network's parameters or any suppression of features from the templates.

While adversarial attacks are typically used to fool classification models, in this paper we show that they can be leveraged to counter the bias in the features of face recognition networks. In particular, we show that our method can help obfuscate gender-related information by manipulating these feature representations using an adversarial attack on images. We propose a new single-step adversarial attack based on minimizing the mutual information between features and the gender attribute that can be executed in a black-box or white-box manner. We demonstrate that injecting controlled imperceptible noise into images with our method can improve feature-level privacy in face recognition systems.

The main contributions of this work are the following:

1. We propose a novel privacy-enhancing mutual information-based adversarial attack with a black-box and a white-box versions that targets the information at the template level of face recognition systems.
2. We provide extensive results using two evaluation settings: The first setting (evaluation setting "A") evaluates the attack ability to fool a gender classifier while the second (evaluation setting "B") assesses the attack ability to align the gender distributions.
3. The attack is applied to enhance gender privacy on the template level of a face recognition network and we show competitive results to state-of-the-art methods in terms of the trade-off between privacy and face verification performance.

2. Related Work

Several adversarial attacks have been developed to study the vulnerabilities of deep learning-based and other machine learning classification models. The main categories of such attacks depend on the amount of knowledge about the classification model that an attacker is assumed to have. In the first category, the attacker performs a white-box attack, where all of the parameters of the classification model are exposed [6, 8, 9].

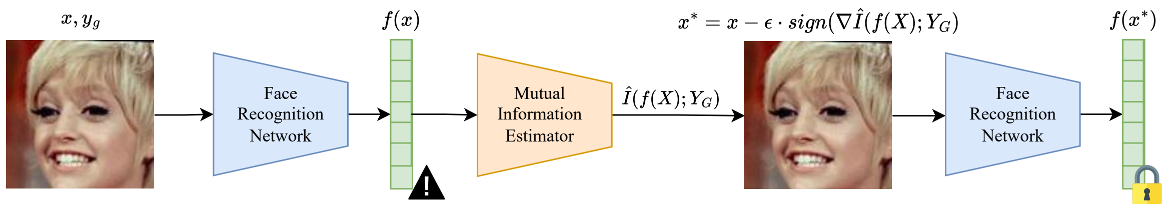


Figure 1. Overview of the proposed method. $\hat{I}(f(X);Y_G)$ designates the estimated mutual information between the features obtained from a face recognition network's embedding layer $f(X)$ and the target gender label Y_G.

Such attacks are typically very successful in inducing misclassifications of the adversarial images. In the second category, while the attacks are typically not as effective as the white-box attacks, they have realistically more chances to occur as the attacker can corrupt the classification outcome with no knowledge of the model's architecture and parameters [2, 16]. The most effective white-box attacks use the gradient information to simulate in one-step [6] or iteratively [8, 9] a gradient ascent behaviour on the targeted model's loss. Meanwhile, several black-box attacks attempt to approximate the targeted model's gradient then apply the gradient-based attacks from the white-box category [2, 16].

In [18], the authors propose to use adversarial attacks as a privacy preserving approach by showing that attacking a gender classification network using two white-box attacks has little effect on the verification performance of a face recognition network with a similar backbone architecture. Other works rely on editing images so that they contain more visual characteristics of a different category from the sensitive attribute while simultaneously preserving the face recognition performance using generative adversarial networks [11, 12] or face morphing [21].

However, these works only boost privacy of the images against image classification models, do not delve into the privacy aspect on the template level. Templates generated by deep-learning based face recognition networks are considered sensitive as they encode features that can be used to train soft biometric classifiers such as gender classifiers [15]. A few methods focus on the soft biometric privacy of templates generated by face recognition systems. These are based on the suppression or the shuffling of the features in the templates [10, 19, 20] or rely on training and fine-tuning a face recognition network to remove the sensitive information [5, 13]. Our approach stands out as it employs images to enhance the privacy of soft biometrics at the face recognition network template level.

3. Attacks

3.1. Our mutual information-based attacks

We propose a one-step adversarial attack that we refer to as Mutual Information based Fast Gradient Sign Method (MI-FGSM). It perturbs images X such that $\hat{I}(f(X);Y_G)$, the estimated mutual information between the feature distribution obtained by a face recognition network $f(X)$ and the ground-truth gender variable Y_g is minimized. The attack is a blackbox attack because it does not require knowledge of any gender classifier's parameters. It protects the templates generated by the face recognition network from unknown classifiers by limiting the amount of information that the templates can reveal about the gender variable. The gradient used in the attack is that of the mutual information estimator with regards to the input image and not the gradient of the classifier's loss as in the Fast Gradient Sign Method (FGSM) [6]. The attack simulates the behaviour of a one step of gradient descent on the mutual information. The attack is performed on a clean sample x using a controlled level of perturbation that we denote as ϵ.

An MI-FGSM adversarial image denoted as x^* is generated as follows:

$$x^* = x - \epsilon \cdot sign(\nabla \hat{I}(f(X);Y_G)) \tag{1}$$

Given that not all of the features employed in this attack are anticipated to have significant relevance to gender, we explore the potential benefits of learning a separate set of gender-specific features derived from the initial features and employing them in the attack instead. Consequently, we introduce an alternative white-box version of the attack, that we refer to as MI-FGSM-MLP. In this attack, the mutual information is estimated between the gender labels Y_G and gender-specific features denoted as $f_{MLP}(X)$ obtained via a multi-layer perceptron (MLP) gender classifier that we train on the original face recognition features and their respective gender labels. The reliance on the gender-specific features from the MLP classifier characterizes this attack as a white-box attack. The MLP classifier used is composed of two fully connected layers of 512 and 1 units respectively, separated by a LeakyReLU activation function, the logits layer of 1 unit is followed by a sigmoid function and is trained by minimizing the binary cross entropy between the ground truth labels and the probabilities resulting from the sigmoid.

An MI-FGSM-MLP adversarial image is obtained as follows:

$$x^* = x - \epsilon \cdot sign(\nabla \hat{I}(f_{MLP}(X);Y_G)) \tag{2}$$

3.2. Reference attacks

As reference attacks, we consider two methods, the first is the original Fast Gradient Sign Method attack [6] as it is a one-step attack with a controlled perturbation, similar to our attacks. This attack is a white-box attack as it requires the knowledge of the parameters of the attribute classifier. It simulates a gradient ascent step to maximize the loss of a targeted classifier. To perform the attack, we use the sign of the gradient of the MLP classifier's cross-entropy loss $H(p,\hat{p})$ with p and $\hat{p}$ denoting the class labels and estimated class probability distribution respectively.

Using FGSM, the adversarial image is generated according to the following equation:

$$x^* = x + \epsilon \cdot sign(\nabla H(p,\hat{p})) \tag{3}$$

We also consider a random attack, referred to as RNDATK, where the sign tensor that is multiplied by the perturbation level ϵ is randomly generated, the resulting image is obtained via the following equation:

$$x^* = x + \epsilon \cdot sign \cdot M \tag{4}$$

where M is a Bernoulli matrix with parameter $p = 0.5$.

4. Experiment settings

4.1. Mutual Information Estimator

To estimate the mutual information between high dimensional feature vectors $f(X)$ and the ground-truth gender distribution Y_g, given the conditional probability distribution $p(y_g|f(x))$, we consider a Contrastive Log-ratio Upper Bound to mutual information (CLUB) defined in [3] as follows:

$$\begin{aligned} I_{CLUB}(f(X);Y_G) := \mathbb{E}_{p(f(x),y_g)}[\log p(y_g|f(x))] - \\ \mathbb{E}_{p(f(x))}\mathbb{E}_{p(y_g)}[\log p(y_g|f(x))] \end{aligned} \tag{5}$$

However, in our case, $p(y_g|f(x))$ is unknown therefore we use a Variational Contrastive Log-ratio Upper Bound estimator $vCLUB$ defined in [3]. The variational upper bound to the mutual information between the features and the gender variable is given by the following equation:

$$\begin{aligned} I_{vCLUB}(f(X);Y_G) := \mathbb{E}_{p(f(x),y_g)}[\log q_\theta(y_g|f(x))] - \\ \mathbb{E}_{p(f(x))}\mathbb{E}_{p(y_g)}[\log q_\theta(y_g|f(x))] \end{aligned} \tag{6}$$

The variational distribution $q_\theta(y_g|f(x))$, an approximation of the conditional distribution $p(y_g|f(x))$ is obtained by a neural network with parameters θ. The unbiased estimator of the variational upper bound in Equation 6 is the following:

$$\begin{aligned} \hat{I}_{vCLUB}(f(X);Y_G) = \frac{1}{N}\sum_{i=1}^{N}[\log q_\theta(y_{g_i}|f(x_i)) - \\ \frac{1}{N}\sum_{j=1}^{N}\log q_\theta(y_{g_j}|f(x_i))] \end{aligned} \tag{7}$$

If the neural network has sufficient capacity and is parameterized properly, it can result in a reliable variational distribution q_θ which in return, allows us to have a reliable upper bound of the mutual information as defined in Equation 5.

The architecture used in our experiments consists of two fully connected layers with 256 and 2 units respectively, separated by a Rectified Linear Unit (ReLU) activation function. For further details about the theoretical derivation of the upper bound, please refer to the work in [3].

4.2. Face Recognition Model and Datasets

We use a IResNet50 ArcFace [4] model trained on the VGGFace2 dataset as the targeted face feature extractor. We use three datasets to perform and evaluate the attack: The Labeled Faces in the Wild (LFW) dataset [7] consists of 13,233 images in unconstrained conditions of 5,749 identities. AgeDB [14] contains 16,516 images of 570 identities with a large variation in age. ColorFeret [17] contains 11,338 images of 994 identities collected under controlled conditions.

In a first step, we analyze the effectiveness of the attack with different perturbation levels using a gender-balanced subset of 5,000 samples for each dataset. In a second step, to evaluate the privacy utility trade-off we use the totality of the datasets for both the gender classification and the face verification tasks.

4.3. Workflow

To ensure a reliable estimation of mutual information, we initially train mutual information estimators between the feature vectors and the target variable for each dataset separately. As a first step to evaluate the attacks, we chose a range of perturbation levels of [0.05, 0.3], with an incremental step of 0.05, applied to a small subset of each dataset consisting of 5,000 gender-balanced images. This allows us to observe variations in gender classification as the perturbation levels increase.

Based on these observations we investigate a sensitive range of perturbation levels inside the initial range to further analyze the attacks. In a final step, we select the most effective perturbation level denoted as ϵ^*, and use it to generate adversarial images for the totality of the datasets. These larger datasets are then used to evaluate the overall gender classification performance and the face verification performance.

5. Evaluation settings

We describe in Table 1, the two evaluation settings we use to evaluate our approach as an adversarial attack and as a privacy-enhancing method. The first setting, referred to as "A", concerns the evaluation of the approach as an adversarial attack. In this setting, the classification model used to evaluate the attack is the same MLP classifier defined in Section 3.1. This classifier is trained on features generated from clean images only. In the first setting, the decision boundary is fixed because it is estimated by the MLP on clean features. In this case, we do not expect the

Evaluation Setting	Evaluation Model	Training data
A:Attack	MLP	Clean
B:Privacy	3-Fold Cross Validation - Logistic regression - Linear SVM - RBF kernel SVM	Adversarial

Table 1. Description of the evaluation settings used to evaluate the effectiveness of the attacks both as adversarial attacks and privacy preserving approaches.

method to achieve necessarily a closer accuracy to 50% because the method does not target a specific fixed decision boundary. It aims to move the gender distributions closer to each other which would not necessarily land on the decision boundary estimated from the clean features. The goal of this setting is to evaluate our attack's ability to fool the MLP classifier as an adversarial attack and a higher rate of false predictions is preferable.

In the second setting, referred to as "B", we are more concerned with evaluating the separability of the features computed for the adversarial images. In this context, we assume that a person interested in inferring the gender, can retrain gender classifiers using features from the adversarial images. In this setting, the decision boundary is updated and therefore correctly assesses the separability of the two distributions. Therefore, this evaluation setting is more useful for privacy evaluation and similar settings have been used previously to evaluate privacy [10, 20]. We evaluate the gender classification performance by calculating the average accuracy of three different classifiers: a logistic regression classifier, a linear SVM as well as an RBF-kernel SVM. We cross-validate these classifiers on the features generated from adversarial images and the closer the accuracy will be to 50%, the more private the templates will be as the templates would lay closer to the decision boundary and therefore are harder to separate by gender.

6. Results

6.1. Impact of Mutual Information Estimator Training on the Attack

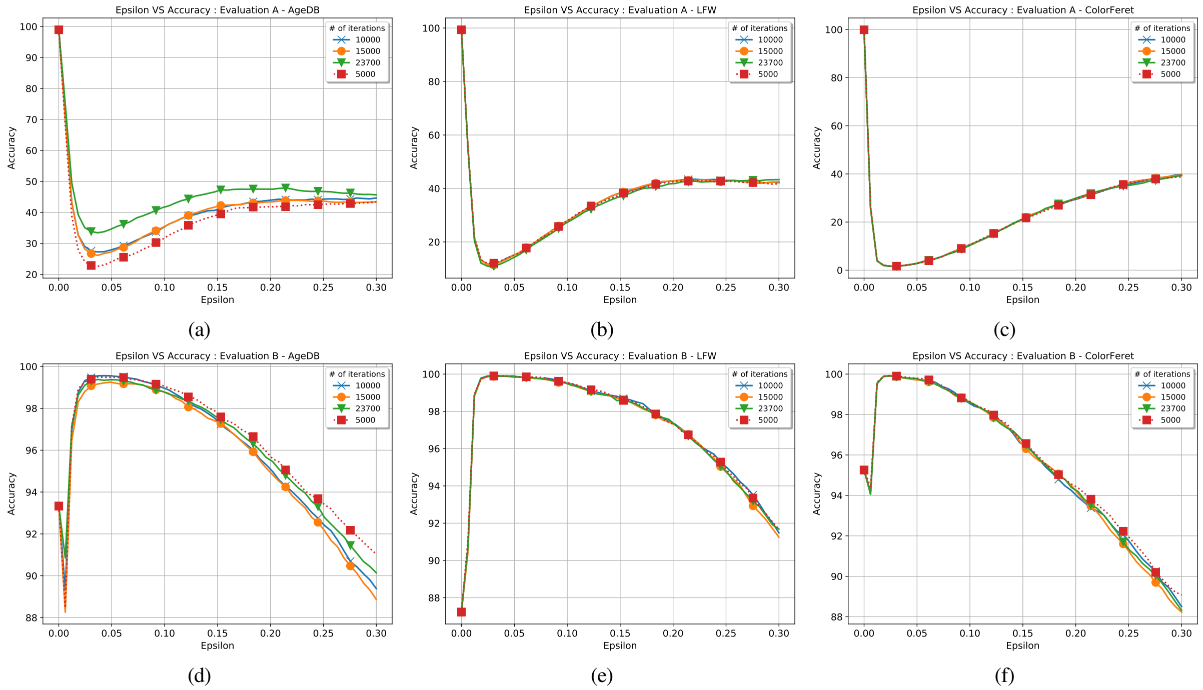
In our experiments, the mutual information estimator is parameterized by a neural network that requires training. We train this estimator on the features of each dataset separately. We compare the performance of the attack in both attack and privacy evaluation setting using different training iterations of the mutual information estimator. This is a useful experiment to determine when the estimator is performant depending on how many iterations it was trained for.

We notice that for the evaluation setting "A", the attack seems to reduce considerably the accuracy of the MLP classifier, particularly at a perturbation magnitude lower than 0.05. For the LFW and the ColorFeret dataset, there is not any apparent difference between the estimators at different training steps. However, for the AgeDB dataset, the MLP classification performance is lowest at 5,000 training iterations and highest at 23,700 training iterations for the MI estimator. Interestingly, in the privacy evaluation setting "B", the accuracy is highest for the 5,000 training iterations but lowest for the 15,000 training iterations for AgeDB. For the LFW and the ColorFeret datasets, a similar behaviour is apparent especially for epsilons neighbouring 0.30. However, the difference between estimators is not as apparent. A better performance on the privacy evaluation setting "B" is more desirable. The results in this setting suggest that a MI estimator trained for a number of iterations larger than 15,000 could potentially hinder the performance of the attack. This could be due to an overfitting on certain batches of the data after 15,000 iterations. Therefore we consider 15,000 to be a sufficient number of iterations to train the MI estimators for each dataset before using them in the attacks. Nevertheless, the results in both settings highlight the stability of the estimator as it consistently yields similar results regardless of variations in training iterations.

6.2. Attack Comparison

6.2.1 Default perturbation range

The first observation we can make from Figures 2 and 3 is that the results in the attack evaluation setting "A" differ from the privacy evaluation setting "B" in a visible manner, not only when we are evaluating our MI based FGSM attacks but also for the reference attacks. In fact, in the attack evaluation setting "A", we see that all the attacks except for the random attack significantly reduce the accuracy of the MLP classifier, in particular for a very low range of perturbation levels that are lower than 0.05. We note also that while our MI-based attacks result in a lower accuracy of the MLP than the original FGSM for two of the subsets LFW and ColorFeret, the opposite behaviour is observed for the AgeDB subset. We also notice that our black-box attack using solely the face recognition features to minimize the mutual information is better than the white-box attack with the gender-specific features. Overall, our black box attack achieves competitive results compared to the white box FGSM and MI-FGSM-MLP attacks in most datasets. As for the privacy evaluation setting "B", we notice that past a very low level of perturbation, all of the attacks except the random attack result in a higher average accuracy of the independent classifiers than the average clean accuracy. This could be due to the classifiers detecting and learning a noise in the feature vectors that is resulting from the attack that make them more discriminative for the gender classification task. For this reason, we are interested in the low perturbation level where the accuracy is reduced. Therefore we performed the attack on a lower perturbation range that consists of 50 uniform steps until the maximum level of 0.03.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2. Evolution of the performance of the gender classification from the features of the MI-FGSM adversarial attack with increasing levels of perturbation and varying training iterations of the MI estimator. The first row corresponds to the attack evaluation setting “A” and the second row corresponds to the privacy evaluation setting “B”.

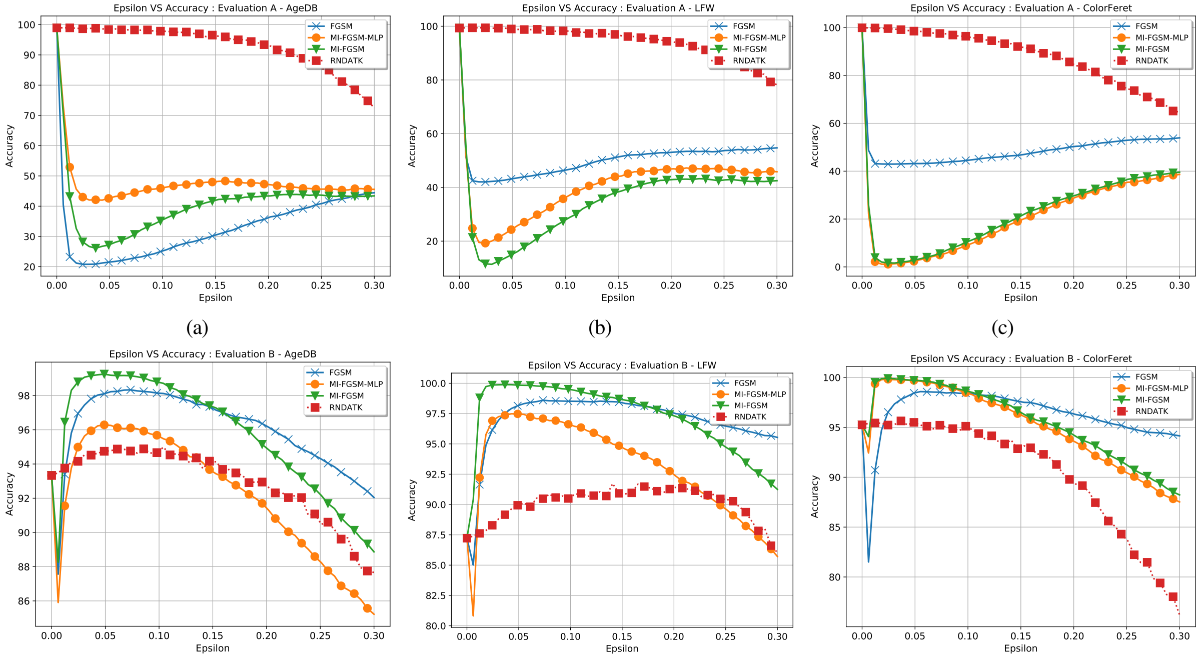
6.2.2 Sensitive perturbation range

We observe in Figure 4, that for the attack evaluation setting “A”, a perturbation level as low as 0.015 is sufficient for our MI-based FGSM attacks and the original FGSM attack to reach their lowest respective accuracies. However, in the privacy evaluation setting “B”, the most randomized accuracies are all associated with MI-based attacks regardless of the dataset. It is important to highlight that while the gap between MI-based attacks and the FGSM attack performance is small for the AgeDB dataset, it is more considerable for the LFW and ColorFeret dataset where the difference in the lowest accuracies achieved between the MI-FGSM method and the FGSM method is as high as 7 points on LFW with MI-FGSM reaching 75% accuracy compared to 82% with FGSM and 4 points on the ColorFeret dataset where MI-FGSM achieves 75% accuracy and FGSM reaches 79% accuracy. Overall, the results show that the MI-based attacks are more promising in the privacy evaluation setting “B” than the FGSM attack.

6.2.3 Impact of mixing clean and adversarial templates on privacy

We observe from Figure 5 that for the AgeDB dataset, including clean images in the privacy evaluation setting “B” does indeed help achieve more privacy when the adversarial images are perturbed with the maximum perturbation $\epsilon = 0.3$ for both MI-based attacks. Indeed, storing a combination of the templates from clean and adversarial images results in more confusion in the classifiers, despite the significant degradation in the adversarial images. The most randomized gender classification performance (highest privacy) is achieved with the MI-FGSM-MLP attack when the percentage of clean images in the set is 37.5% , this results in an accuracy of 80.57% instead of 85.29% with no clean images in the set. For the MI-FGSM attack, the percentage of clean images in the set of 50% achieves an accuracy of 83.31% compared to 89.03% with no clean images in the set and 93.32% on a completely clean set. When it comes to LFW and ColorFeret datasets, while we observe more privacy similar to AgeDB as we include progressively clean images in the set of adversarial images, both MI-based attacks with the optimal perturbations and with an only adversarial image set still achieve the best privacy with an accuracy of ∼75% for both datasets, reducing the clean accuracy by ∼12 points for LFW and ∼20 points for ColorFeret. Meanwhile, mixing clean images with adversarial images attacked with the maximum perturbation achieves a lowest accuracy of 77.43% for LFW when using MI-FGSM-MLP and 78.93% when using MI-FGSM. Similarly for ColorFeret, this combination of clean and adversarial images only reduce the accuracy by at most 15 points using both MI-based attacks with the maximum perturbation.

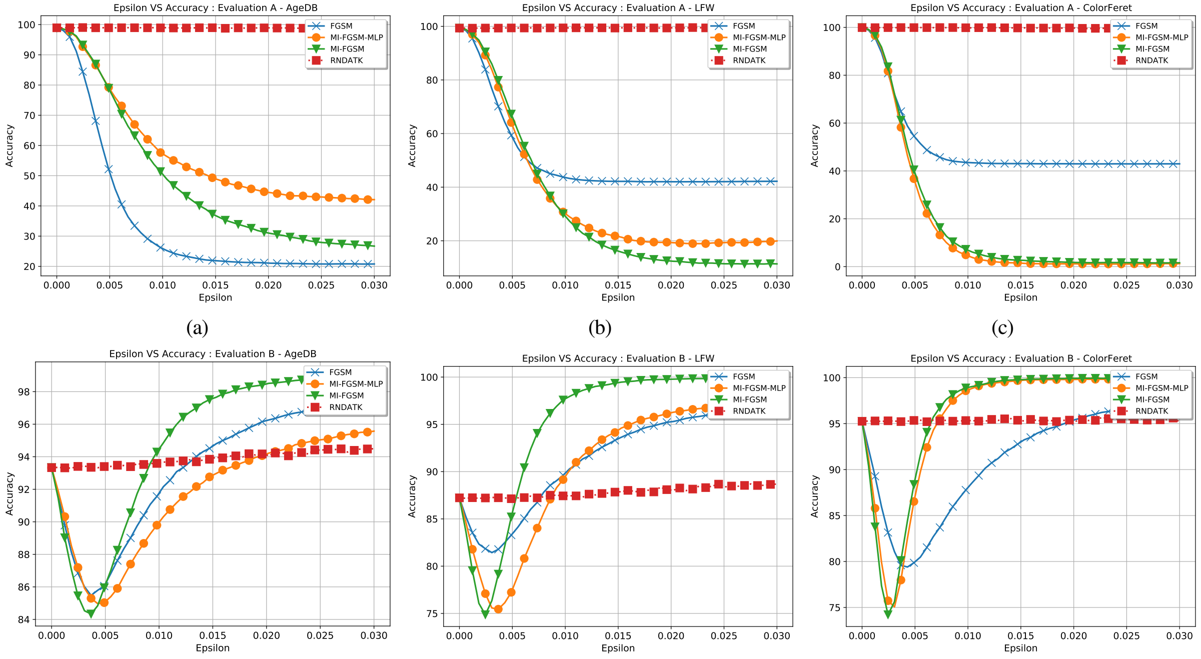
For all datasets, we notice that the optimal perturbation is associated with the best privacy outcome when no clean images are



(a) (b) (c)

(d) (e) (f)

Figure 3. Comparison of different attacks using the default range of levels of perturbation in both attack and privacy evaluation settings. The first row corresponds to the attack evaluation setting "A" and the second row corresponds to the privacy evaluation setting "B".



(a) (b) (c)

(d) (e) (f)

Figure 4. Comparison of different attacks using the sensitive range of levels of perturbation in both attack and privacy evaluation settings. The first row corresponds to the attack evaluation setting "A" and the second row corresponds to the privacy evaluation setting "B".

included in the set. This finding confirms that the optimal perturbation is effective to reduce the mutual information between the templates and the gender attribute while remaining undetectable by classifiers at the template level. This prevents the classifiers

Method	Dataset					
	LFW		AgeDB		ColorFeret	
	Gender Accuracy (%)	EER (%)	Gender Accuracy (%)	EER (%)	Gender Accuracy (%)	EER (%)
Reference (No Privacy)	84.96	0.483	93.36	5.076	94.60	0.822
Ours (MI-FGSM)	75.93	**0.616**	**87.02**	**6.003**	76.46	**1.173**
IVE (2019) [19]	**75.32**	1.56	89.92	10.88	85.14	1.67
Multi-IVE (2023) [10]	75.62	2.92	87.42	10.91	**75.22**	3.07

Table 2. Comparison of our MI-FGSM attack as a privacy-enhancing approach to state-of-the-art privacy methods.

in the privacy evaluation setting "B" from overfitting due to excessive noise introduced at the template level through image perturbation as is the case with the maximum perturbation.

6.3. Privacy-Face Verification Trade-Off

To evaluate the privacy-utility trade-off, we compare our approach with two privacy enhancing methods based on feature elimination. Incremental Variable Elimination (IVE) method [19] estimates feature importance for targeted attribute classification by iteratively eliminating features based on their contribution. However, despite being an intuitive and effective privacy approach, it suffers from significant information loss, as it reduces template dimensionality. Multi-IVE [10] is an improved version of IVE. It projects features onto PCA or ICA-generated domains, preserving dimensionality, and then performs feature suppression on this transformed domain while locking a number of the first principal components. We reproduce IVE and Multi-IVE methods according to the best parameters described in the papers, using several elimination steps. We make sure that for each dataset, we perform the training of the feature importance estimators using a different dataset than the evaluation set and we select the best closest performance in terms of privacy to our results. For the Multi-IVE method, we used the PCA generated domain with 5 locked principal components as advised in [10].

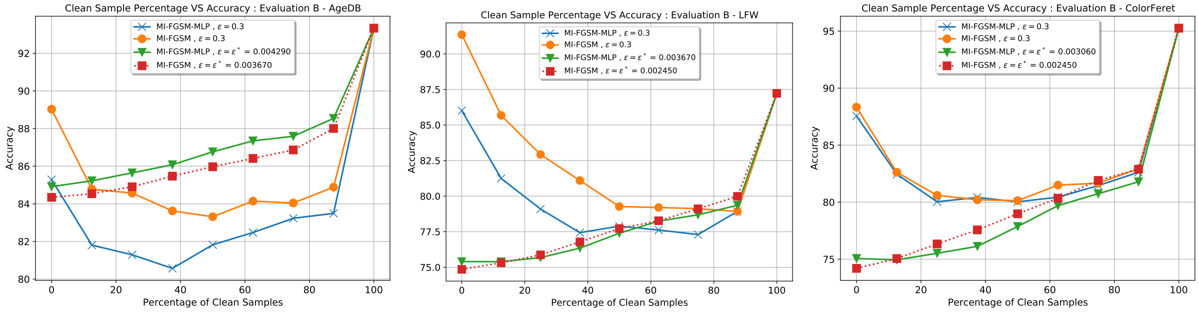
We evaluate using the privacy evaluation setting "B", the MI-FGSM attack on the totality of the images from every dataset to quantify the trade-off between privacy and face verification performance. We use the average balanced accuracy, instead of the plain accuracy of the gender classifiers across the 3-folds to evaluate the gender classification performance as the totality of the datasets are not all gender balanced. In order to have results that describe reliably the impact of privacy enhancing techniques on verification, all verification evaluations on LFW and ColorFeret are performed following the standard protocol 1 for benchmark on the LFW dataset in [7] where 6,000 pairs (3,000 mated and 3,000 non-mated) are compared using the Euclidian distance. As for AgeDB, we make sure that the pairs selected for comparison have an age gap of 30 years to reproduce the most challenging protocol "AgeDB-30" defined by the dataset's authors in [14].

We notice that the MI-FGSM method significantly enhances privacy with very minimal deterioration in verification performance. The efficiency of the attack is the highest on the ColorFeret dataset with close to 18 points of difference between the balanced accuracy before and after the attack. It is followed by LFW with around 9 points and AgeDB with around 6 points. On the verification part, AgeDB's EER increases the most with close to 1 point from before to after the attack. In comparison to the state-of-the-art methods, we notice from Table 2 that for similar levels of privacy, our method is better at minimizing the impact on the verification performance. For instance, in the case of AgeDB where gender and identity are highly entangled, we notice that the EER after applying the Multi-IVE method doubles from 5.076% to 10.88% for a comparable gender accuracy to our method at ~ 87 %.

In Figure 6, we plotted the t-SNE visualization of the features from clean and adversarial images generated with the optimal perturbation ϵ^* as well as the maximum perturbation we used in our experiments $\epsilon = 0.3$. We also show, for the assessment of the attack perceptability, a sample from the LFW dataset with different levels of perturbation. Based on a qualitative assessment of the images, we cannot see any discernible deterioration for the image when using the optimal perturbation, as opposed to the maximum perturbation, where the noise is quite visible. However, when it comes to the t-SNE visualization, we notice that the gender distribution of the features changes considerably from the clean to the attacked samples. Using the clean samples, the t-SNE graph shows a clear separability of the feature vectors associated to different gender categories. Meanwhile, using adversarial samples, the features appear to be more overlapping. However, the T-SNE visualization of the feature vectors generated with the maximum perturbation $\epsilon = 0.3$ fails to capture the high separability we observed in the privacy evaluation in Section 6.2.2.

7. Conclusion

In this paper, we proposed a privacy-enhancing mutual information-based adversarial attack, that is performed in a single step and targets the gender information contained in templates of face recognition systems to enhance their privacy. We defined two distinct evaluation protocols for both the attack and the privacy aspects and we thoroughly analyzed the attack performance using varying levels of perturbation. We showed that this attack is effective both as an adversarial attack and as a privacy-preserving method. We identified the optimal perturbation level yielding the greatest reduction in the gender classification performance and ensured that the perturbation is extremely small and therefore, imperceptible to the human eye. We further demonstrated that for certain data distributions, it may be beneficial from a privacy standpoint, to store MI-FGSM adversarial templates with higher perturbation levels alongside clean templates rather than only using MI-FGSM adversarial templates as this can result in more randomized gender distributions. Furthermore, we provided both a black-box and a white-box version of our attack and demonstrated that using the black-box version is sufficient to outperform the FGSM attack, and yields better verification performances for comparable



(a) (b) (c)

Figure 5. Impact of combining clean image templates with adversarial image templates on the gender classification performance in the privacy evaluation setting "B". The maximum perturbation $\epsilon = 0.3$ and optimal perturbation ϵ^* per dataset are compared using MI-FGSM and MI-FGSM-MLP attacks.

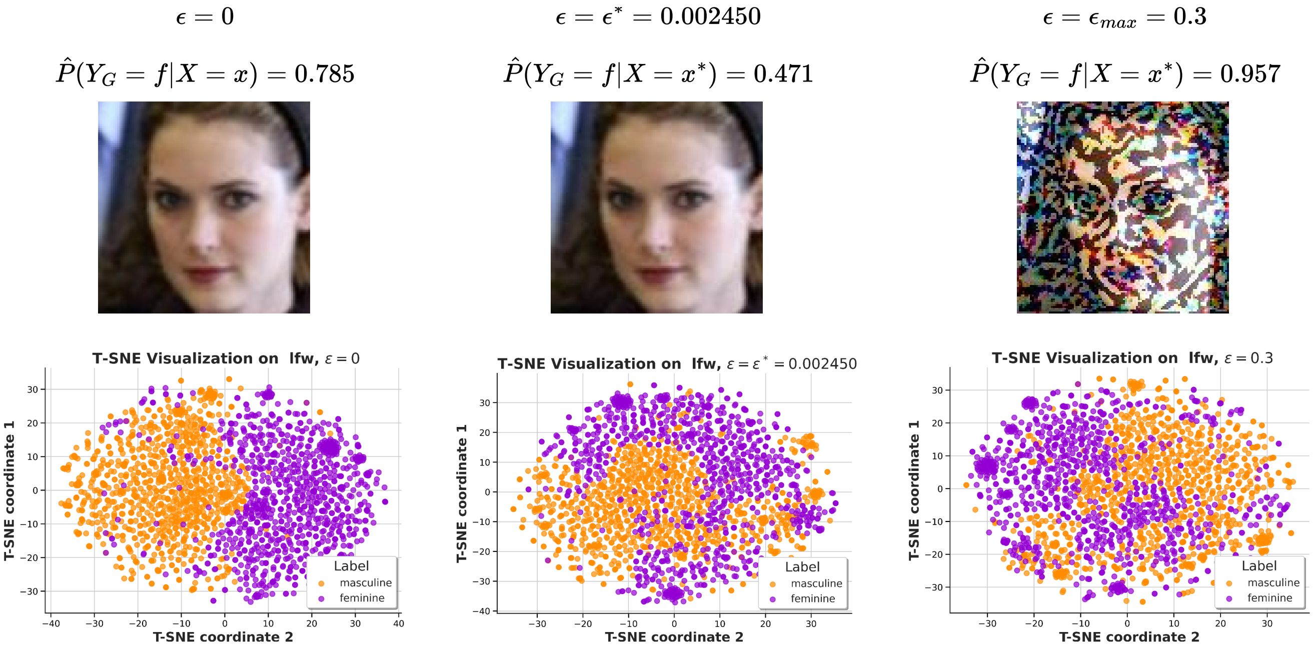


Figure 6. Visualization of a sample from the LFW dataset with different perturbations (top row) and T-SNE visualizations of features generated from LFW dataset images with different perturbations using the MI-FGSM attack (bottom row). $\epsilon = 0$ indicates clean image features. The "feminine" ("f") category probabilities shown are the averages of classifier probabilities in evaluation set "B".

levels of privacy of state-of-the-art privacy-enhancing methods. Future research could combine our attack with model-based privacy enhancing methods to improve privacy without compromising verification performance.

Acknowledgements

We thank Rezwan Hasan and Richard Guest for helpful discussion. 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] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Ivan Bartolome, and Julian 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*, pages 584–593. Springer, 2019. 1

[2] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial*

intelligence and security, pages 15–26, 2017. 2

[3] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020. 3

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 3

[5] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 330–347. Springer, 2020. 2

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 3

[7] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric 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. 3, 7

[8] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 1, 2

[9] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 1, 2

[10] Pietro Melzi, Hatef Otroshi Shahreza, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Sébastien Marcel, and Christoph 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*, pages 323–331, 2023. 1, 2, 4, 7

[11] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. FlowSAN: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019. 2

[12] Mirjalili, Vahid and Raschka, Sebastian and Ross, Arun. PrivacyNet: semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020. 2

[13] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020. 2

[14] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 3, 7

[15] Gokhan Ozbulak, Yusuf Aytar, and Hazim Kemal Ekenel. How transferable are cnn-based features for age and gender classification? In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2016. 1, 2

[16] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 2

[17] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000. 3

[18] Zohra Rezgui, Amina Bassit, and Raymond Veldhuis. Transferability analysis of adversarial attacks on gender classification to face recognition: Fixed and variable attack perturbation. *IET biometrics*, 11(5):407–419, 2022. 2

[19] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Suppressing gender and age in face templates using incremental variable elimination. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 1, 2, 7

[20] Philipp Terhörst, Kevin Riehl, Naser Damer, Peter Rot, Blaz Bortolato, Florian Kirchbuchner, Vitomir Struc, and Arjan Kuijper. Pe-miu: A training-free privacy-enhancing face recognition approach based on minimum information units. *IEEE Access*, 8:93635–93647, 2020. 1, 2, 4

[21] Shunxin Wang, Una M. Kelly, and Raymond N.J. Veldhuis. Gender obfuscation through face morphing. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2021. 2