IET Biometrics

IET The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Transferability analysis of adversarial attacks on gender classification to face recognition: Fixed and variable attack perturbation

Zohra Rezgui[1] [iD] | Amina Bassit[1,2] [iD] | Raymond Veldhuis[1,3] [iD]

[1]EEMCS Faculty, Data Management & Biometrics Group, University of Twente, Enschede, The Netherlands

[2]EEMCS Faculty, Services and CyberSecurity Group, University of Twente, Enschede, The Netherlands

[3]Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Gjøvik, Norway

Correspondence

Zohra Rezgui, EEMCS Faculty, Data Management & Biometrics Group, University of Twente, Drienerlolaan 5, 7522 NB, Enschede, The Netherlands.
Email: z.rezgui@utwente.nl

Abstract

Most deep learning-based image classification models are vulnerable to adversarial attacks that introduce imperceptible changes to the input images for the purpose of model misclassification. It has been demonstrated that these attacks, targeting a specific model, are transferable among models performing the same task. However, models performing different tasks but sharing the same input space and model architecture were never considered in the transferability scenarios presented in the literature. In this paper, this phenomenon was analysed in the context of VGG16-based and ResNet50-based biometric classifiers. The authors investigate the impact of two white-box attacks on a gender classifier and contrast a defence method as a countermeasure. Then, using adversarial images generated by the attacks, a pre-trained face recognition classifier is attacked in a black-box fashion. Two verification comparison settings are employed, in which images perturbed with the same and different magnitude of the perturbation are compared. The authors' results indicate transferability in the fixed perturbation setting for a Fast Gradient Sign Method attack and non-transferability in a pixel-guided denoiser attack setting. The interpretation of this non-transferability can support the use of fast and train-free adversarial attacks targeting soft biometric classifiers as means to achieve soft biometric privacy protection while maintaining facial identity as utility.

1 | INTRODUCTION

The cutting-edge advances in deep learning (DL) have made computer vision problems more approachable. However, neural networks can be unreliable on data unseen during the training, which makes their security questionable. In fact, the majority of DL models are vulnerable to *adversarial attacks* that, based on subtle perturbations applied to the clean samples, mislead the classifier with high confidence. There has been a number of studies investigating the vulnerabilities of DL-based machine learning systems to different types of adversarial attacks on the input images. The existing attacks can be partitioned into two categories: *white-box attacks*, where an

adversary has full access to the attacked model's parameters, and *black-box attacks,* where an adversary has no access to such information. Typically, white-box attacks are more powerful than black-box attacks due to their ability to leverage the parameters of the model against its own predictions.

In a real-life scenario, a deployed model's parameters would not be accessible leaving the black-box attacks as the only option to disrupt its predictive performance. To benefit from the strength of white-box attacks, Ref. [2, 3] show that it is possible to target a model, where its parameters are known, and transfer the resulting effects on an unknown model, as long as the two models are trained for the same task. Particularly in the field of biometrics, the effectiveness of these

---

attacks should not be overlooked, given the variety of biometric applications such as forensics and border control where wrong predictions are not tolerated.

Many biometric applications are inter-connected, specifically those related to the face modality. For instance, there is a plethora of studies showing that different face recognition systems can be enhanced with a soft biometric classifier such as a gender classifier [4]. Similarly, deep face recognition features are known to be discriminative for soft biometric classification [5] via transfer learning. This association motivates us to investigate the potential transferability of adversarial attacks on models that share the same input space but were trained independently to perform different tasks. Previous studies on the transferability of adversarial attacks, however, did not include such a hypothesis because they were solely concerned with same-task transferability.

The investigation of this hypothesis could be beneficial in terms of soft biometric privacy. There is usually a trade-off between privacy and utility in privacy applications. Utility specifies how much information we want to keep in the protected data. Indeed, if we consider an adversarial attack to be a privacy mechanism that conceals soft biometric attributes, studying between-task transferability allows us to assess how such a privacy mechanism would impact utility if utility is expressed as the performance of a face recognition classifier. For instance, an effective adversarial attack on a gender classifier that is not transferable to a face recognition system can be used as a privacy protection mechanism to obfuscate gender.

This paper is an extension of work originally presented in BIOSIG 2021 [1], where we investigate the transferability of two adversarial attacks against a gender classifier to a face recognition classifier where both classifiers are independently trained and only share the same input space (facial images) and the same model architecture. We start by providing an overview of the hypothesis of transferability between different tasks given the same input space. We then study the impact of two existing gradient-based attacks and deep feature-based defence on the gender classifier. Subsequently, we use the generated adversarial images along with those resulting from the defence against a pre-trained face recognition model to analyse the transferability of both the attacks and the defence. In Ref. [1], there was a faint transferability between an attack on a gender classifier using a range of fixed perturbations. The verification comparisons were performed over images perturbed with the same perturbation. In this extension, we study the transferability using variable perturbations from two attacks. The comparisons are between images with different perturbations in contrast to what was done in Ref. [1]. We also extend the experiments to a ResNet50 architecture.

Our results, illustrated in Figure 1, support the transferability hypothesis for the fixed perturbation setting but not for the variable perturbation setting. This finding suggests that unlike fixed-perturbation assaults, attacks with variable perturbations may be better suited as privacy protection mechanisms as there is no indication that they are transferrable from a soft biometric classifier and a face recognition classifier.

2 | RELATED WORK

Adversarial attacks have become an active area of research as they expose the design vulnerabilities of DL-based models. Several white-box attacks are gradient-based such as the Fast Gradient Sign Method (FGSM) [6], its iterative version (IFGSM) [7], and Projected Gradient Descent [8]. Unlike the gradient used in backpropagation to train neural networks, the gradient used in those attacks helps determining the nearest perturbation to the input such that the adversarial image is misclassified. Other methods are based on network architecture information, Ref. [9] finds the minimal perturbation possible to an image that would make it misclassified, via projecting inputs on the closest classification hyperplane. Results in Ref. [10] show that DL-based face recognition models, such as VGGFace, are vulnerable to such attacks and to image processing methods that perturb the samples in a perceptible manner. Moreover, Ref. [11] uses Generative Adversarial Network (GAN)-based image editing to change the direction of the predictions of a binary gender classifier. However, the changes in the resulting images are perceptible to the human eye, which contradicts the purpose of adversarial attacks.

To improve the robustness of existing DL models, many defence approaches have been proposed to withstand these attacks. Ref. [6, 12] show that incorporating adversarial samples with the training data increases the attacked model's robustness but such an approach can be resource-demanding. In practice, the model is trained over a diverse training set, where it learns to correctly classify the clean samples and, at the same time, it rectifies the predictions of the adversarial samples. Ref. [13] enhances the classifier's predictions by targeting each class and partitioning it into several sub-classes, assuming that only a few of them are sensitive to adversarial attacks. Subsequently, the different predictions of the sub-classes are aggregated via voting. Other approaches are based on input reconstruction such as Ref. [14] by using a denoising auto-encoder on the adversarial images in order to remove the perturbations. This method has been improved in Ref. [15] by using a U-Net architecture for the denoiser and defining the reconstruction loss based on the deep features of the classifier.

While white-box attacks are effective on known machine learning models, it was shown in Ref. [16] that the resulting adversarial images can be effective against unknown models. The literature refers to such phenomenon as *attack transferability* where the attacked model is called *surrogate model* and the model to which the attack is transferred is called *target model*. Ref. [16] shows that adversarial attacks are transferable between the same models and between different models performing the same task, whether these models are differentiable (such as DNNs) or non-differentiable (such as Support Vector Machines). Ref. [2] analyses the level of complexity of the surrogate model in an attempt to justify the transferability effectiveness; a surrogate model that has a low variance loss function is more transferable than a model with a high variance loss function. In order to ameliorate transferability across different neural networks performing the same task, Ref. [17] modifies the IFGSM attack by randomly resizing the images at each
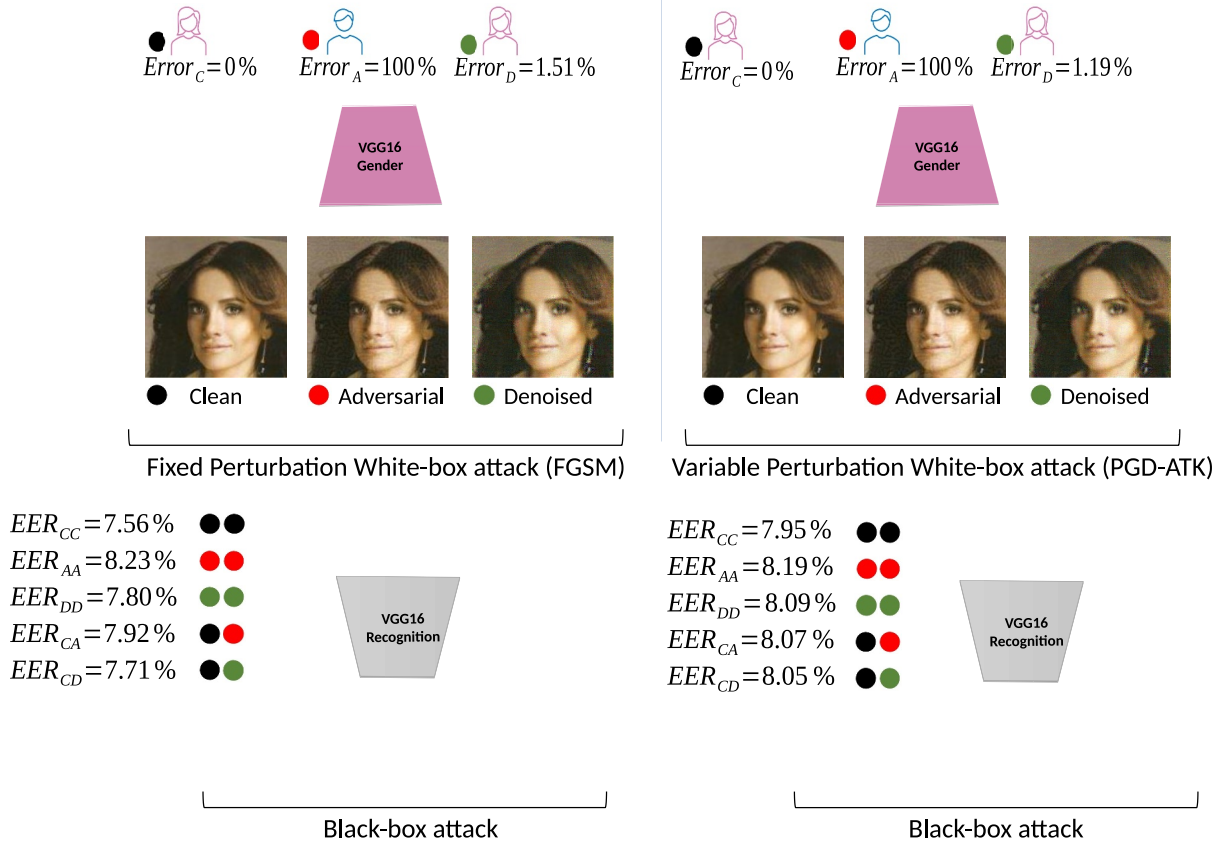
FIGURE 1 Overview of the transferability hypothesis from a white-box attack on a gender classifier to a black-box attack on a face recognition classifier with VGG16 architecture. We show the results from the Fast Gradient Sign Method (FGSM) fixed perturbation setting on the left and from the PGD attack (PGD-ATK) variable perturbation setting on the right. The black dots refer to the clean images denoted as C, the red dots refer to the adversarial images generated by each attack denoted as A and the green dots refer to the denoised images generated by the defence method denoted as D. $Error_C$, $Error_A$ and $Error_D$ correspond to the gender classification errors on the sets $B_{CleanTrans}$, $B_{AdvTrans}$ and $B_{DenTrans}$ before we use them for verification with the face recognition model. Once we use them for verification using different pair combinations, we derive the equal error rates EER_{CC}, EER_{AA}, EER_{DD}, EER_{CA}, and EER_{CD}. We see that the attacks make no significant difference on EER as we compare the EER among the different combinations, in particular the variable perturbation white-box attack (PGD-ATK)

iteration. Ref. [18] proposes a GAN-based approach to generate synthetic adversarial samples with imperceptible perturbations against FaceNet [19] and report effective results across different face recognition models. Similarly, Ref. [3] reports transferability of attacks from an open-source surrogate face recognition model to several commercial target face recognition models.

As approaches to privacy protection, authors in Ref. [20, 21] used GANs to edit input images by modifying their sensitive attribute category. This is often done via integrating pre-trained soft biometric attribute classifiers as privacy estimators and pre-trained face recognition systems as utility estimators into the framework of the GAN, using them as additional discriminator networks. This allows having realistic-looking images with the characteristics of the faces slightly changed to fool the pre-trained soft biometric attribute classifiers but not the face recognition models. While these methods are promising, they tend to be computationally heavy, unlike common adversarial attacks. In this regard, it is important to investigate whether fast and train-free adversarial attacks are transferable from soft biometric classifiers to face recognition systems. It turns out that when such attacks are

non-transferable, they can be used as a privacy protection mechanism to obfuscate a soft biometric attribute.

3 | METHODOLOGY

Let us denote X_F as the space of all facial images, X_C the space of clean images, X_{Adv} the space of adversarial images and X_{Den} the space of denoised adversarial images where $X_C \cup X_{Adv} \cup X_{Den} \subseteq X_F$. We denote $Y_G = \{0, 1\}$ the space of the gender labels and $Y_R = \{✓, ✗\}$ the space of recognition labels. We consider $G: X_F \to Y_G$ a gender classifier and $R: X_F \times X_F \to Y_R$ a facial recognition classifier.

Attack An adversarial attack $f_{Adv}: X_C \to X_{Adv}$ is considered successful if for $x \in X_F$ there is an adversarial sample $f_{Adv}(x) = x_{Adv} \in X_{Adv}$ such that:

$$G(x) = y_G \text{ and } G(x_{Adv}) = \overline{y_G}$$

Denoising Defence Let $f_{Den}: X_{Adv} \to X_{Den}$ denote a denoising function. Ideally, a denoised image $x_{Den} = f_{Den}$

$(x_{Adv}) \in X_{Den}$ and verifies $G(x_{Den}) = G(x)$ where $x \in X_C$ is the clean image such that $x_{adv} = f_{Adv}(x)$.

Gender-Recognition Attack Transferability We say that an attack f_{adv} is transferrable from the gender classifier G to a face recognition model R for $(x_1, x_2) \in X_C \times X_C$ if we have

$$R(x_1, x_2) \neq R(x_1, f_{Adv}(x_2))$$

Gender-Recognition Defence Transferability We say that a defence f_{Den} is transferrable from the gender classifier G to a face recognition model R for $(x_1, x_2) \in X_C \times X_C$ if we have

$$R(x_1, x_2) = R(x_1, f_{Den} \circ f_{Adv}(x_2))$$

Metrics we use the classification accuracy, that is, the number of correct predictions divided by the total number of predictions, to measure the performance of the gender classifier, and we derive different performance metrics for the face recognition classifier based on a similarity measure: the equal error rate (EER), the false non-match rate (FNMR) at a fixed false match rate (FMR) of 0.1% (FNMR at 0.1%FMR) as well as area under the detection error trade-off curve (AUC-DET). (See Figure 2).

Based on the above-mentioned definitions, we adopt the following procedure:

1. Train the gender classifier and measure its classification accuracy.
2. Attack the gender classifier to generate a set of adversarial samples.
3. Train a denoising defence on a subset of adversarial samples and their corresponding clean versions and evaluate it on a separate subset by comparing the classification accuracy of the gender classifier on the adversarial images and their denoised versions.
4. Run a face recognition model on a clean set, its adversarial, its denoised versions, and their combinations to assess the transferability of the attack and the defence methods in terms of the sensitivity of the recognition performance across the diverse sets of images as shown in Figure 2.

4 | BACKGROUND

4.1 | Fast Gradient Sign Method Attack

We use $J(\theta, x, y_G)$ to denote the loss function of the gender classifier G with respect to an input image $x \in X_C$ and its ground truth gender label $y_G \in Y_G$. The FGSM attack maximises the loss with respect to the input image [6] by adding to the image a step ϵ in the direction of the loss gradient. An FGSM adversarial attack $f_{Adv}: X_C \to X_{Adv}$, with perturbation magnitude $\epsilon \in \mathbb{R}$, results in adversarial images $x_{Adv} \in X_{Adv}$ such that: $x_{Adv} = x + \epsilon \cdot \text{sign}(\nabla J(\theta, x, y_G))$.

4.2 | Projected Gradient Descent Attack

PGD (Projected Gradient Descent) attack (PGD-ATK) [8] is an iterative attack where each iteration is an FGSM attack while it restricts the perturbation level inside an ℓ_∞ ball. This iterative process increases the chance of obtaining samples that would lie outside of the decision boundary of the correct class. Given S the set of possible perturbations, a projected gradient descent attack (PGD-ATK) $f_{AdvPGD}: X_C \to X_{Adv}$, with mini-step $\alpha \in \mathbb{R}$, results at step $t + 1$ in adversarial images $x_{Adv} \in X_{Adv}$ such that: $x^{t+1} = \prod_{x+S}(x^t + \alpha \, \text{sign}(\nabla_x J(\theta, x, y_G)))$ where the operator $\prod_{x+S}$ designates the projection on $x + S$.

4.3 | High-level representation and pixel guided denoisers

In this paper, we consider two types of denoisers: a pixel-guided denoiser (PGD) and high-level representation guided denoiser (HGD). A PGD learns to reconstruct a clean image x by reducing the loss defined as, $\mathscr{L}_{PGD} = \|x - x_{Adv}\|_1$, the pixel level difference between a clean image x and its adversarial version x_{Adv}. Whereas, a HGD [15] reduces the loss defined as, $\mathscr{L}_{HGD} = \|f_{emb}^i(x) - f_{emb}^i(x_{Adv})\|_1$, the difference between the deep features of a clean image x and the deep features of its adversarial version x_{Adv} where $f_{emb}^i: X_C \to \mathbb{R}^n$ denotes the function describing the attacked model until its ith layer outputs a feature vector of size n.

5 | EXPERIMENT AND EVALUATION

5.1 | Architectures

We used the VGG16 architecture as the gender classification network and restricted its last layer to two classes to suit our classification goal. The same architecture is used for the face recognition model VGGFace pre-trained on the VGGFace dataset [22]. VGG16 has a straightforward architecture that comprises 13 convolution layers and three fully connected layers. Additionally, we repeat the experiments using a ResNet50 architecture as the gender classification network, transforming its last layer to output two class probabilities corresponding to the gender categories. ResNet50 comprises a total of 48 convolution layers and two layers of max-pooling and average pooling, respectively, before its last fully connected layer for classification. We also use the same architecture for the face recognition network to which we transfer the attacks on the ResNet50 gender classifier. In this case, the face recognition network is pre-trained [23] with an arc loss on the MS1MV dataset.[1]

For the denoiser, similarly as Ref. [15], we use a U-Net based Denoising Convolution Neural Network,[2] a denoising model, which we will refer to in this work as UDnCNN. The
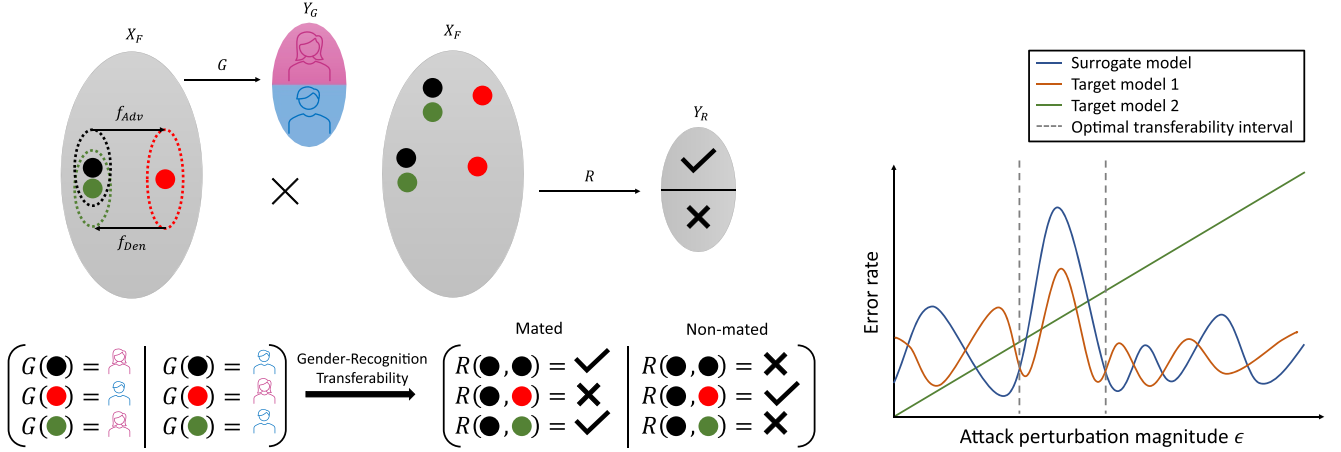

FIGURE 2 Methodology overview for analysing the transferability attack from a gender classifier (surrogate model) to a face recognition classifier (target model). A mated comparison is a comparison where two images are of the same subject. A non-mated comparison is a comparison where two images are of different subjects. The graph on the right is an illustration of the expected behaviour of a target model versus any target model where the attack is not transferable. For a target model that is not affected by the attack, we expect that the performance of such a model would only deteriorate due to higher distortion ϵ in the images (Target model 2). For a model to which an attack is transferable, its optimal interval where the error peaks would be similar to that of the surrogate model (Target model 1). Illustration from Ref. [1].
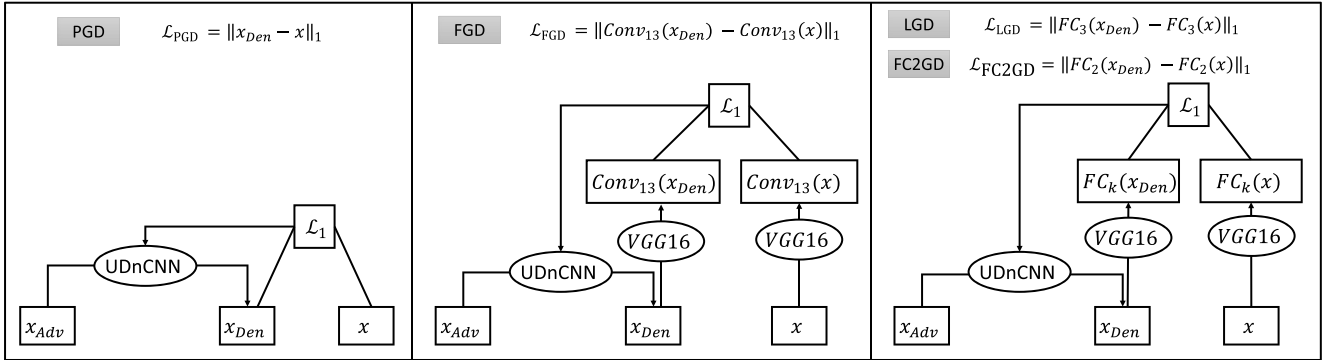

FIGURE 3 Training of UDnCNN denoiser when considering the pixel-guided denoiser (PGD) defence, the feature-guided denoiser (FGD) defence, and the logits guided denoiser (LGD) defence ($k = 3$) [15] and when considering the second fully connected layer-guided denoiser (FC2GD) defence ($k = 2$). Figure from Ref. [1].

structure of the UDnCNN denoiser has an encoding part, sharing skip connections with a decoding part. The skip connections allow the transfer of fine-grained information that could be lost in a regular auto-encoder, as shown in Figure 3.

5.2 | Dataset division

We use the CelebA dataset that comprises 202, 599 samples of 10, 177 different individuals. We divide this dataset into three sets: A (162, 770 samples), B (19, 962 samples), and C (19, 867 samples) with respect to the train-test-validation partition provided by the authors in Ref. [24] where identities do not overlap. For each adversarial attack against the gender classifier and the defence experiment, we use sets A and B to train and test the gender classifier and set C to generate adversarial images against the gender classifier. The resulting adversarial images and their corresponding clean versions are partitioned into four subsets: $C_{AdvTrain}$ and $C_{CleanTrain}$ of equal size as

well as $C_{AdvTest}$ and $C_{CleanTest}$. The subsets $C_{CleanTrain}$ and $C_{AdvTrain}$ are used for the training of the denoisers while $C_{CleanTest}$ and $C_{AdvTest}$ are used to evaluate them.

For the transferability experiments, we use set B to get the clean images from which we generate the adversarial images and their corresponding denoised images. Since not all the clean images from B are vulnerable to either of the attacks, we collect, for each adversarial image, the clean image it was derived from and its denoised image. As a result of each attack experiment, we have a set of clean images $B_{CleanTrans}$, a set of adversarial images $B_{AdvTrans}$, and another set of denoised images $B_{DenTrans}$ of the same size. Those three sets are used to analyse the transferability of the attacks on the face recognition classifier. We summarise in Tables 1 and 2 the number of samples in each of the sets for each of the attacks for the VGG16 and the ResNet50 architectures, respectively. We note that for the attacks with variable perturbations, there is a considerably smaller amount of adversarial images generated compared to the fixed-perturbation attack. That is explained

TABLE 1 Number of images in the different sets per attack for the VGG16 architecture. B_{Trans} (#ids) refers to the number of images and identities, respectively, present in each of the clean, adversarial and denoised sets used to assess the transferability of the attacks

Attack	FGSM-fixed	PGD-ATK	FGSM-variable
$C_{AdvTrain}/C_{CleanTrain}$	73,779	14,224	-
$C_{AdvTest}/C_{CleanTest}$	18,449	3,557	-
B_{Trans} (#ids)	94,965 (995)	17,473 (997)	19,703 (995)

TABLE 2 Number of images in the different sets per attack for the ResNet50 architecture. B_{Trans} (#ids) refers to the number of images and identities, respectively, present in each of the clean, adversarial and denoised sets used to assess the transferability of the attacks

Attack	FGSM-fixed	PGD-ATK	FGSM-variable
$C_{AdvTrain}/C_{CleanTrain}$	93,786	15,593	-
$C_{AdvTest}/C_{CleanTest}$	23,448	3,899	-
B_{Trans} (#ids)	117,025 (999)	19,470 (999)	19,506 (999)

by the fact that the projected gradient descent attack (PGD-ATK) iteratively performs a projected gradient descent until reaching one adversarial sample. When it comes to FGSM-variable, we select from the images generated by the FGSM-fixed, either randomly one perturbed version per image or the smallest perturbation per image; therefore, we do not keep various adversarial versions of an image.

5.3 | Performance metrics

To assess the gender classifier performance, either before the attack and the defence or after, we calculate the classification accuracy. To reason in terms of errors in the two models, we use the classification error rate (1−accuracy) for the gender classifier in Figure 1. For the face recognition performance, we use the cosine similarity formula in 5.3 to measure the similarity in terms of the cosine of the angle made between pairs of feature vectors: The more similar two features vectors are, the smaller the angle between them and thus the higher the cosine. The cosine similarity between vectors $\vec{a}$ and $\vec{b}$ is given by:

$$\cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \times \|\vec{b}\|}$$

This is further used to measure the FNMR at a fixed FMR of 0.1% as well as the AUC-DET and finally, the EER.

5.4 | Evaluation of the attacks and defence methods

5.4.1 | VGG16 architecture

Training the gender classifier on CelebA We trained our gender classifier from scratch for 20 epochs using batch normalisation after convolution layers to speed up the training of the baseline VGG16, achieving a validation accuracy of 98.62%.

Fast Gradient Sign Method Attack We run the FGSM attack on the VGG16 gender classifier using various values for the perturbation $\epsilon \in [0.005, 0.55]$. Figure 4 shows how the VGG16 classifier behaves for different values of ϵ. We observe that the accuracy decreases for ϵ between 0.01 and 0.035, and it starts to increase from 0.04.

Such behaviour is likely caused by a circular or crescent decision boundary where clean images belonging to an initial class get perturbed with a certain epsilon that pushes them into the opposite class, but with an epsilon too large, they are pushed back into their correct initial class.

As our goal is to study the effect of perturbations that are imperceptible to the human, we consider the following range of epsilons $\epsilon \in \{0.01, 0.015, 0.02, 0.025, 0.03, 0.035\}$ as it is where the VGG16 classifier is most vulnerable.

Projected Gradient Descent Attack We run the PGD attack (PGD-ATK) on the gender classifier using an ℓ_∞ ball of 0.035. This choice is based on the previous FGSM experiment showing that after that value the perturbations become increasingly less effective. The mini-step α of 0.001 is with 10 iterations. The accuracy of the gender classifier over the adversarial samples was 9.1%.

Denoising losses In addition to a PGD, we use three types of HGDs illustrated in Figure 3: FGD based on the last convolutional layer of the gender classifier, FC2GD based on the second fully connected layer and logits guided denoiser (LGD) based on the logits layer.
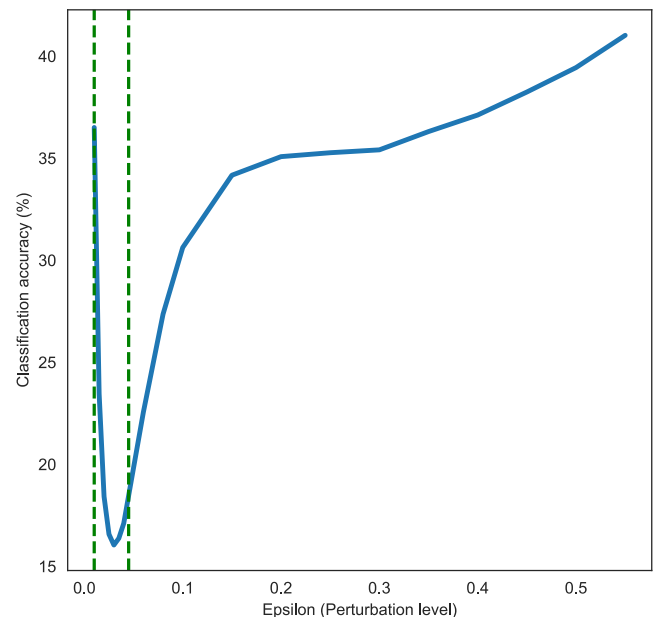

FIGURE 4 Sensitivity of the classification accuracy of the VGG16 gender classifier upon the choice of perturbation (epsilon) used in the Fast Gradient Sign Method (FGSM) attack. The green dashed lines represent the bounds of the range of epsilons selected for the transferability experiments. Figure from Ref. [1]

Comparison between the defence methods on clean and adversarial images with FGSM-fixed perturbation Figure 5 shows the performance of the defence methods over increasing values of epsilon. FC2GD seems to be the most robust against adversarial examples generated with values of epsilon outside of its training range, followed by LGD and FGD. Pixel-guided denoiser, on the other hand, is the most vulnerable to high epsilons. Nevertheless, we notice that the performance inevitably drops at a certain range for all three HGD methods before slowly increasing again.

Tables 3 and 4 compare the performance of the attacked VGG16 gender classifier when applying the different defence methods (columns 2–5) and without (first column), over clean images (row 2) and adversarial images (row 3). For PGD and FC2GD, both help considerably in defending the classifier against adversarial attacks as the accuracy reaches 84.34% on the adversarial test images with PGD denoising and 93.14% with FC2GD denoising. We also observe that there is a deterioration of the performance of the classifier on clean images after they are fed into the denoiser. This effect is particularly noticeable for the FC2GD. The latter seems to infer adversarial noise more effectively than PGD but with the expense of reduced discriminative power in clean images. For the HGD methods, we observe the higher the representation (i.e. the deeper the target layer) the better the defence method performs on clean images and that LGD seems to be the most convenient method for defence so far. Following this result from Ref. [1], we used LGD as a defence method for the remainder of the experiments.

5.4.2 | ResNet50 architecture

Training the gender classifier on CelebA We trained our baseline ResNet50 gender classifier from scratch for 20 epochs and achieved a validation accuracy of 98.13% at epoch 7.

Fast Gradient Sign Method Attack We run the FGSM attack on the ResNet50 gender classifier using the same perturbation values as an initial attack against VGG16 with $\epsilon \in [0.005, 0.55]$. Figure 6 shows that ResNet50 seems to be most sensitive to higher values of epsilons. We pick the most optimal range of epsilons as $\epsilon \in \{0.11, 0.115, 0.12, 0.125, 0.13, 0.135\}$ where the classifier makes the wrong prediction for all the perturbed samples. Given that the attack is effective on the totality of the attack generation set, the selection of the images for the FGSM in the variable setting was based on a random choice of perturbation for each image rather than choosing the smallest perturbation in order to avoid having only images perturbed with $\epsilon = 0.11$.

Projected Gradient Descent Attack We run the PGD attack (PGD-ATK) on the gender classifier using an ℓ_∞ ball of 0.135. This choice is based on the previous FGSM experiment showing that after that value the perturbations become increasingly less effective. The mini-step α of 0.001 is with 10 iterations. The accuracy of the gender classifier over the adversarial samples was 0.2%, making the PGD almost completely effective on the set of images.
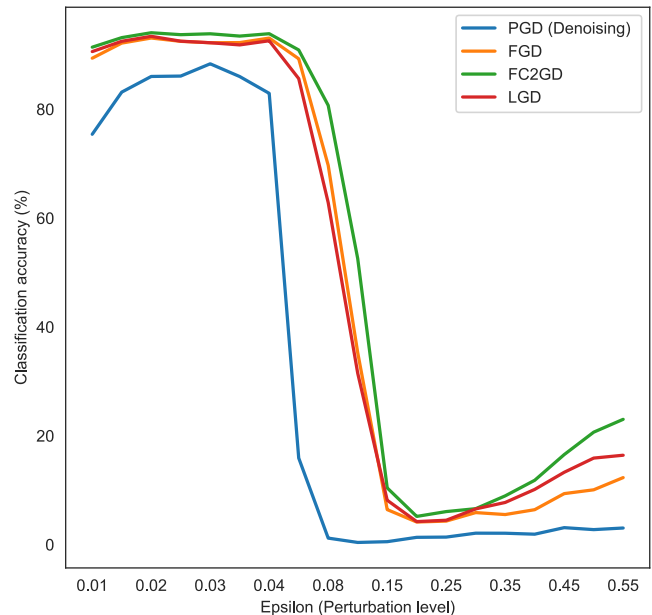

FIGURE 5 Classification accuracy of VGG16 gender classifier during Fast Gradient Sign Method (FGSM) attack and after applying the defence methods over various attack intensities. Effect of the defence methods on the classification accuracy of the VGG16 gender classifier over increasing values of ϵ. Figure from Ref. [1]

Denoising losses We maintain using a LGD for ResNet50 as well to allow for comparisons with results of the VGG16 architecture [1]. Similarly as the VGG16 architecture, we train the LGD denoiser such that the features that are input to the logits layer of the clean and the denoised images become progressively similar with regards to the L1 norm as shown in Figure 3.

5.5 | Transferability of fixed and variable perturbation settings of adversarial attacks for the VGG16 architecture

Instead of pushing each sample exactly on the decision boundary, these attacks are designed to push it to the opposite side of the decision boundary. In terms of privacy protection, the former would have been preferable, but for the sake of this study, we will evaluate the transferability of these attacks as done in the literature.

We study the transferability of the attack from the gender classifier (surrogate) to the face recognition model (target) by performing five comparison combinations of mated and non-mated comparisons depending on the type of the input images, either clean, adversarial or denoised. The totality of these combinations are illustrated in Figure 1 and an overview of the methodology is illustrated in Figure 2. We perform a verification entirely on the clean set (CC) to obtain a baseline performance of VGGFace before running the attacks. We then perform clean/adversarial (CA) and clean/denoised (CD) verifications to evaluate the transferability of both the attack

	Without denoising	PGD	FGD	FC2GD	LGD
B_{Clean}	98.19%	95.61%	57.50%	63.48%	83.05%
$C_{AdvTest}$ $\epsilon \in [0.01, 0.035]$	0%	84.34%	91.82%	93.14%	92.02%

T A B L E 3 Performance summary of the gender classifier attacked with FGSM-fixed perturbation in terms of accuracy with and without the defence methods. Table from Ref. [1]

	Without denoising	PGD	FGD	FC2GD	LGD
B_{Clean}	97.61%	94.58%	45.01%	66.76%	80.88%
$C_{AdvTest}$ $\epsilon \in [0.01, 0.035]$	0%	77.57%	87.47%	89.52%	87.84%

T A B L E 4 Performance summary of the gender classifier attacked with FGSM-fixed perturbation in terms of F1 score with and without the defence methods
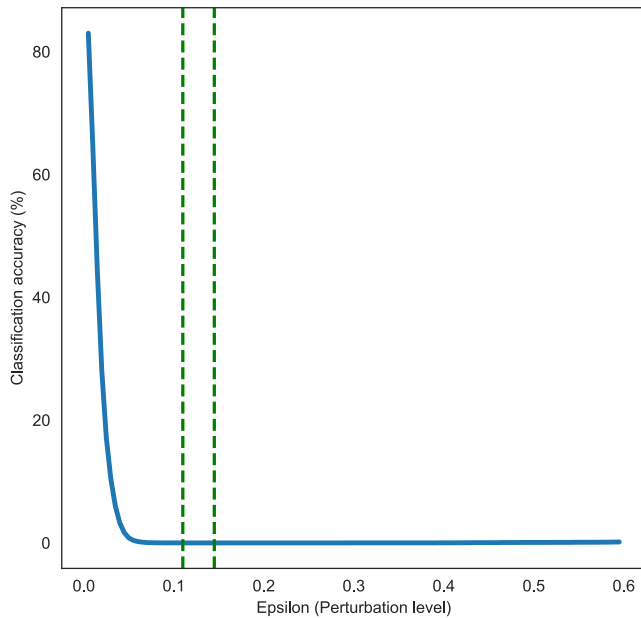

F I G U R E 6 Sensitivity of the classification accuracy of the ResNet50 gender classifier upon the choice of perturbation (epsilon) used in the Fast Gradient Sign Method (FGSM) attack. The green dashed lines represent the bounds of the range of epsilons selected for the transferability experiments

and the defence method. We also report the combinations, adversarial/adversarial (AA) and denoised/denoised (DD). To realise the comparisons, we select ~15 different images per subject where each image should be vulnerable to at least three values of ϵ out of 6. Table 5 summarises the resulting numbers of mated and non-mated comparisons per epsilon and in total (combining all the comparisons per epsilon). We provide in Table 6 the evaluation of the VGG16 gender classifier on the transferability sets before attacking the face recognition model.

We notice in the Figures, 7 and 8 that the presence of non-clean images (denoised and adversarial) regardless of the attack intensity decreases the recognition performance. The difference between the variation of the performance in the combinations CC, CD and DD, where there is 0% of adversarial samples, and the variation in combinations CA and AA, where there is 50% and 100% respectively, show that VGGFace is prone to degradation as more adversarial images are included in the comparisons. In case of the three comparison combinations CC, CD and CA, we observe that the recognition performance degrades from CC to CA and that the error

difference is larger than the error difference between CC and CD. This suggests that the defence partly compensates the performance degradation. Tables 7 and 8 show that for each combination involving adversarial or denoised images, the errors are the highest for the smallest perturbation 0.01; then, for the subsequent increasing perturbations, the errors decrease until perturbation 0.025 before they start to increase again. This implies a low transferability of the attack in the selected epsilon range. It is possible that if a more optimal range of epsilon values exists, that would result in a high transferability of the attack as shown in the illustrative graph in Figure 2. When it comes to the attacks with a variable perturbation setting, we design in the same manner as before for both FGSM and the PGD-ATK; the following are comparison combinations: CC, CA, CD, AA and DD. We also select 15 distinct images per subject to perform these comparisons. We give in Table 9 the number of mated and non-mated comparisons for each of the attacks. We observe from the Figure 7 that the PGD-ATK even though seems to have the same pattern in terms of EER as the FGSM in the fixed perturbation setting, the differences across the various combinations are not substantial. For the FGSM with variable perturbation, the EER is almost constant at around 7.9%. From Figure 8, we see that both attacks with a variable perturbation setting maintain a similar FNMR@0.1%FMR across combinations in contrast to the fixed perturbation setting. We also notice that for both EER and FNMR@0.1%FMR, the values are higher in most cases for the attacks with variable perturbation than with the FGSM with fixed perturbation. However, this is expected as in the fixed perturbation setting we compare images that are perturbed similarly. We understand from these observations that the transferability shown in the fixed perturbation setting does not apply in the variable perturbation setting.

5.6 | Transferability of fixed and variable perturbation settings of adversarial attacks for the ResNet50 architecture

We generate the transferability sets in a similar manner as we did for VGG16. We provide in Table 10 the evaluation of the ResNet50 gender classifier on the transferability sets before attacking the face recognition model and in Tables 11 and 12 the number of mated and non-mated comparisons for the fixed and variable settings, respectively. We observe in the

TABLE 5 Number of mated (M) and non-mated (U) comparisons in a fixed perturbation setting with adversarial images generated with a Fast Gradient Sign Method (FGSM) attack for the VGG16 architecture. Table from Ref. [1]

ϵ	0.01	0.015	0.02	0.025	0.03	0.035	All
CC	M = 4.7×10^3	M = 8.4×10^3	M = 1.1×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 6.7×10^4
	U = 1.2×10^6	U = 2.6×10^3	U = 3.8×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 2.1×10^7
DD	M = 4.7×10^3	M = 8.4×10^3	M = 1.1×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 6.7×10^4
	U = 1.2×10^6	U = 2.6×10^6	U = 3.8×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 2.1×10^7
AA	M = 4.7×10^3	M = 8.4×10^3	M = 1.1×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 1.4×10^4	M = 6.7×10^4
	U = 1.2×10^6	U = 2.6×10^6	U = 3.8×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 4.7×10^6	U = 2.1×10^7
CD	M = 9.4×10^3	M = 1.6×10^4	M = 2.3×10^4	M = 2.8×10^4	M = 2.8×10^4	M = 2.8×10^4	M = 1.3×10^5
	U = 2.4×10^6	U = 5.3×10^6	U = 7.7×10^6	U = 9.4×10^6	U = 9.4×10^6	U = 9.4×10^6	U = 4.3×10^7
CA	M = 9.4×10^3	M = 1.6×10^4	M = 2.3×10^4	M = 2.8×10^4	M = 2.8×10^4	M = 2.8×10^4	M = 1.3×10^5
	U = 2.4×10^6	U = 5.3×10^6	U = 7.7×10^6	U = 9.4×10^6	U = 9.4×10^6	U = 9.4×10^6	U = 4.3×10^7

TABLE 6 Evaluation of VGG16 gender classifier on the transferability sets

Attack	FGSM-fixed	PGD-ATK	FGSM-variable
$B_{\text{CleanTrans}}$	$accuracy = 100\%$ $F1 - score = 100\%$	$accuracy = 100\%$ $F1 - score = 100\%$	$accuracy = 100\%$ $F1 - score = 100\%$
B_{AdvTrans}	$accuracy = 0\%$ $F1 - score = 0\%$	$accuracy = 0\%$ $F1 - score = 0\%$	$accuracy = 0\%$ $F1 - score = 0\%$
B_{DenTrans}	$accuracy = 98.49\%$ $F1 - score = 97.82\%$	$accuracy = 99.15\%$ $F1 - score = 99.02\%$	$accuracy = 98.81\%$ $F1 - score = 98.34\%$
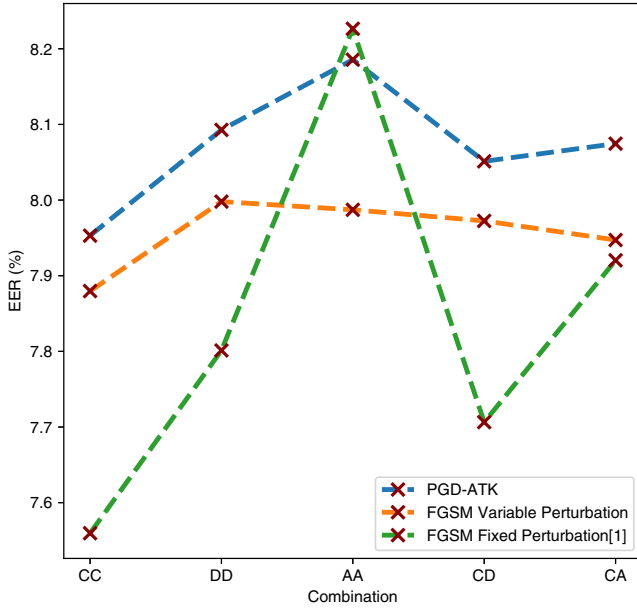
FIGURE 7 Performance in terms of equal error rate (EER) across the different comparison combinations on the VGG16 architecture: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised
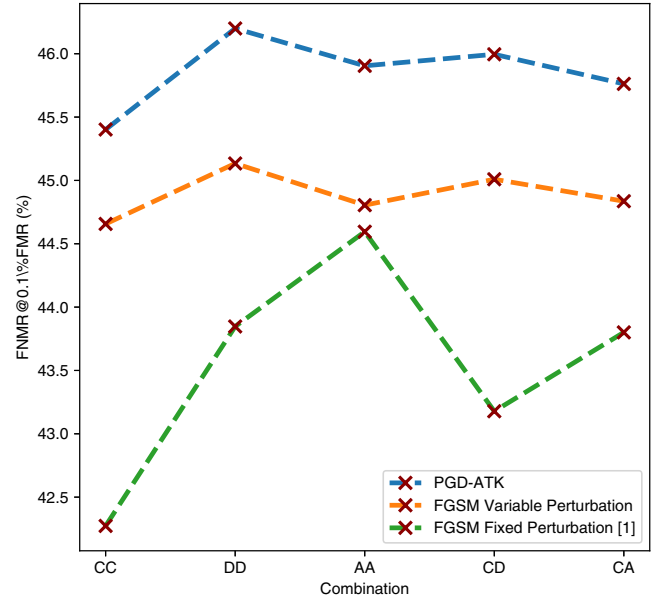
FIGURE 8 Performance in terms of FNMR@0.1%FMR across the different comparison combinations on the VGG16 architecture: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised

Figures 9 and 10 that unlike in the VGG16 case, the errors are the highest for the CD and DD combinations when it comes to the FGSM in fixed and variable settings. Additionally, we notice that these errors are substantially higher than in the VGG16 case, including for the CA and AA combinations that

reach an EER of 19.3% and 21.5%, respectively. When it comes to the variable setting attacks, we notice that while PGD follows a similar behaviour as in the VGG16 case, the variable FGSM is overlapping with the fixed FGSM. Table 13 and Table 14 show that for all combinations other than CC which

serves as a reference, the errors increase with the level of the perturbation.

To ensure that the high transferability error levels in combinations AA and CA are not simply due to the deterioration of the image quality, we performed an attack that we will refer to as random attack that adopts the same fixed perturbation levels as the fixed-FGSM setting but does not rely on

the gender classifier's gradient sign. To obtain the sign for each pixel of every image, we constructed a matrix S of the same size of the image with each element sampled from a Bernoulli distribution of probability $p = 0.5$. Then, zero elements were changed to -1. Results in Table 15 show the errors obtained are lower than those reported in the fixed-FGSM with the same levels of perturbation. This confirms that the sign of the gradient plays a role in the transferability of FGSM from the gender classifier to the face recognition classifier.

In the case of the ResNet50 experiments, we notice that the defence method further deteriorates verification results in the case of the FGSM attacks, both in the fixed and variable perturbation settings. This could be explained by two elements: the high magnitude of the perturbations and the nature of the defence, as it relies on the features extracted from the gender classifier and the depth of the architecture. In fact, given that the architecture is extensively deeper than VGG16 for instance, the feature space for the last layer of the gender classifier is likely to be more tailored for the gender classification task than it was the case for the VGG16 gender classifier. In this case, correcting images based on the distance of their gender features to those of the clean images would not necessarily mean that their recognition features would be pushed closer to each other as well and could be actually further pushed apart. We also notice that this behaviour does not occur for the PGD-ATK scenario, the defence does not correct the adversarial images for the recognition task and although it deteriorates them, the difference between CA and CD errors is not substantial. This could have to do with the fact that even though we set the ℓ_∞ ball to be bounded by 0.135, the PGD-ATK could be finding adversarial samples with substantially lower epsilon values and thus follows the behaviour we see in the experiment with the VGG16 architecture. Altogether, we can see that the FGSM attack, particularly in the fixed setting, can be highly transferable if we find the optimal perturbation levels. On the other hand, this comes at the expense of the transferability of the gender feature-guided denoising methods that can further deteriorate the adversarial images for the face recognition task.
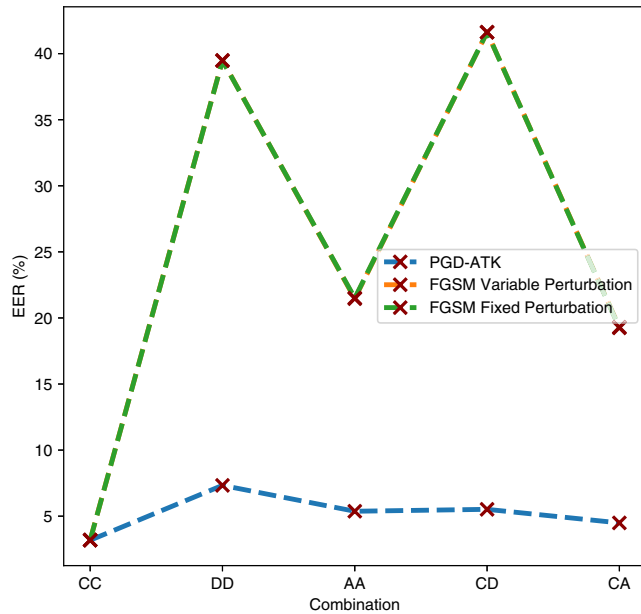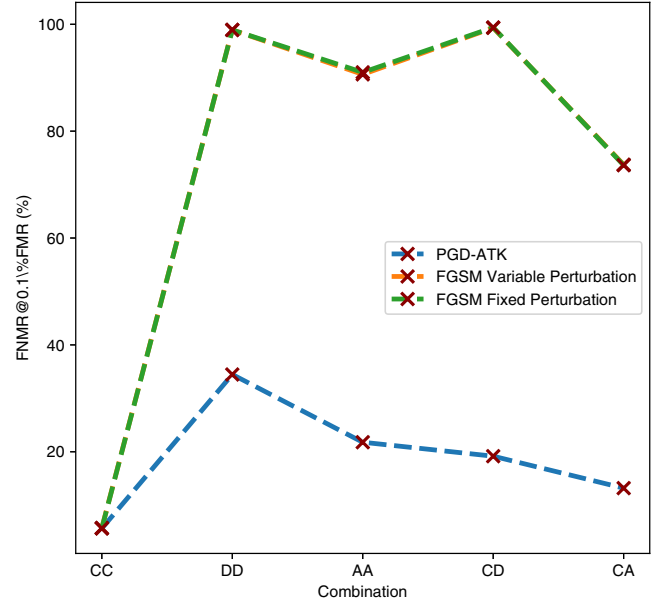
6 | CONCLUSION

In this paper, we studied the effect of two white-box attacks, the FGSM and PGD-ATK attacks, on gender classifiers with VGG16 and ResNet50 architectures, respectively and then assessed their robustness to defence methods by applying a

TABLE 7 Comparison performance of different combinations per epsilon in terms of FNMR@0.1%FMR in percentage (%) for the VGG16 architecture where the first row serves as a reference with only clean images

ϵ	0.01	0.015	0.02	0.025	0.03	0.035
CC	46.96	44.14	42.05	41.41	41.41	41.39
DD	47.70	45.66	43.62	43.08	43.12	43.30
AA	47.28	45.16	43.96	43.38	44.78	44.76
CD	47.39	45.08	42.95	42.33	42.42	42.52
CA	47.18	44.86	43.41	42.78	43.70	43.68

TABLE 8 Comparison performance of different combinations per epsilon in terms of area under the DET curve (area under the detection error trade-off curve (AUC-DET)) in percentage (%) for the VGG16 architecture. The first row serves as a reference with only clean images. Table from Ref. [1]

ϵ	0.01	0.015	0.02	0.025	0.03	0.035
CC	2.98	2.68	2.51	2.41	2.41	2.41
DD	3.08	2.78	2.61	2.50	2.52	2.55
AA	3.08	2.87	2.83	2.70	2.86	2.86
CD	3.04	2.73	2.57	2.46	2.46	2.48
CA	3.04	2.79	2.70	2.58	2.69	2.69

TABLE 9 Number of mated (M) and non-mated (U) comparisons in a variable perturbation setting with adversarial images generated with a PGD attack (PGD-ATK) and a Fast Gradient Sign Method (FGSM) attack for the VGG16 architecture

Attack	PGD-ATK	FGSM
CC	M = 8.0×10^4; U = 7.8×10^7	M = 8.2×10^4; U = 7.9×10^7
DD	M = 8.0×10^4; U = 7.8×10^7	M = 8.2×10^4; U = 7.9×10^7
AA	M = 8.0×10^4; U = 7.8×10^7	M = 8.2×10^4; U = 7.9×10^7
CD	M = 1.6×10^5; U = 1.6×10^8	M = 1.6×10^5; U = 1.6×10^8
CA	M = 1.6×10^5; U = 1.6×10^8	M = 1.6×10^5; U = 1.6×10^8

TABLE 10 Evaluation of ResNet50 gender classifier on the transferability sets

Attack	FGSM-fixed	PGD-ATK	FGSM-variable
$B_{\text{CleanTrans}}$	$accuracy = 100\%$ $F1 - score = 100\%$	$accuracy = 100\%$ $F1 - score = 100\%$	$accuracy = 100\%$ $F1 - score = 100\%$
B_{AdvTrans}	$accuracy = 0\%$ $F1 - score = 0\%$	$accuracy = 0\%$ $F1 - score = 0\%$	$accuracy = 0\%$ $F1 - score = 0\%$
B_{DenTrans}	$accuracy = 87.83\%$ $F1 - score = 86.25\%$	$accuracy = 97.58\%$ $F1 - score = 96.88\%$	$accuracy = 87.89\%$ $F1 - score = 86.33\%$

TABLE 11 Number of mated (M) and non-mated (U) comparisons in a fixed perturbation setting with adversarial images generated with a Fast Gradient Sign Method (FGSM) attack for the ResNet50 architecture

ϵ	0.11	0.115	0.12	0.125	0.13	0.135	All
CC	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 104$	$M = 5 \times 10^5$
	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 107$	$U = 4.1 \times 10^8$
DD	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 5 \times 10^5$
	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 4.1 \times 10^8$
AA	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 8.4 \times 10^4$	$M = 5 \times 10^5$
	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 8.2 \times 10^7$	$U = 4.1 \times 10^8$
CD	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 8.5 \times 10^5$
	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 8 \times 10^8$
CA	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 1.7 \times 10^5$	$M = 8.5 \times 10^5$
	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 1.6 \times 10^8$	$U = 8 \times 10^8$

TABLE 12 Number of mated (M) and non-mated (U) comparisons in a variable perturbation setting with adversarial images generated with a PGD attack (PGD-ATK) and a Fast Gradient Sign Method (FGSM) attack for the ResNet50 architecture

Attack	PGD-ATK	FGSM
CC	$M = 8.1 \times 10^4; U = 7.9 \times 10^7$	$M = 8.2 \times 10^4; U = 7.9 \times 10^7$
DD	$M = 8.1 \times 10^4; U = 7.9 \times 10^7$	$M = 8.2 \times 10^4; U = 7.9 \times 10^7$
AA	$M = 8.1 \times 10^4; U = 7.9 \times 10^7$	$M = 8.2 \times 10^4; U = 7.9 \times 10^7$
CD	$M = 1.7 \times 10^5; U = 1.6 \times 10^8$	$M = 1.7 \times 10^5; U = 1.6 \times 10^8$
CA	$M = 1.7 \times 10^5; U = 1.6 \times 10^8$	$M = 1.7 \times 10^5; U = 1.6 \times 10^8$

FIGURE 10 Performance in terms of FNMR@0.1%FMR across the different comparison combinations on the ResNet50 architecture: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised. FGSM-Fixed and FGSM-Variable perturbation line plots are overlapping

FIGURE 9 Performance in terms of equal error rate (EER) across the different comparison combinations on the ResNet50 architecture: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised. FGSM-Fixed and FGSM-Variable perturbation line plots are overlapping

TABLE 13 Comparison performance of different combinations per epsilon in terms of area under the DET curve (area under the detection error trade-off curve (AUC-DET)) in percentage (%) for the ResNet50 architecture. The first row serves as a reference with only clean images

ϵ	0.11	0.115	0.12	0.125	0.13	0.135
CC	1.24	1.24	1.24	1.24	1.24	1.24
DD	34.80	34.95	35.06	35.19	35.35	35.50
AA	9.98	10.89	12.81	13.80	14.80	15.80
CD	37.24	37.55	37.84	38.19	38.57	38.93
CA	8.51	9.25	10.87	11.74	12.65	13.61

TABLE 14 Comparison performance of different combinations per epsilon in terms of FNMR@0.1%FMR in percentage (%) for the ResNet50 architecture where the first rows serves as a reference with only clean images

ϵ	0.11	0.115	0.12	0.125	0.13	0.135
CC	5.67	5.67	5.67	5.67	5.67	5.67
DD	98.86	98.88	98.95	98.96	98.97	98.99
AA	85.77	87.36	89.94	90.94	91.81	92.59
CD	99.27	99.32	99.39	99.43	99.50	99.51
CA	65.48	68.20	73.49	75.97	78.31	80.47

TABLE 15 Comparison performance of combinations adversarial/adversarial (AA) and clean/adversarial (CA) per epsilon in the fixed Fast Gradient Sign Method (FGSM) setting and in the random attack (RAND-ATK) setting in terms of area under the DET curve (area under the detection error trade-off curve (AUC-DET)) in percentage (%) for the ResNet50 architecture

ϵ	0.11	0.115	0.12	0.125	0.13	0.135
AA	9.98	10.89	12.81	13.80	14.80	15.80
AA (RAND-ATK)	2.49	2.68	2.84	3.08	3.28	3.54
CA	8.51	9.25	10.87	11.74	12.65	13.61
CA (RAND-ATK)	2.07	2.20	2.32	2.50	2.66	2.81

feature-guided denoising method. Once the effectiveness of these attacks was established in fooling the gender classifier, we tested their transferability from the gender classification task to the facial recognition task with similar architectures in a black-box manner. To assess the performance of the target facial recognition classifier, we used two different settings for the verification comparisons: a fixed perturbation setting, in which we only compare images perturbed with the same level of perturbation, and a variable perturbation setting, in which we compare images perturbed with different perturbations. Altogether, we can see that the FGSM attack, particularly in the fixed setting, can be highly transferable if we find the optimal perturbation levels. On the other hand, this comes at the expense of the transferability of the gender feature-guided denoising methods that can further deteriorate the adversarial images for the face recognition task. Further work should be done to assess the transferability of variable perturbation settings such as the PGD-ATK. The non-transferability of such an attack targeting a soft biometric classifier (e.g. a gender classifier) to a face recognition classifier makes this attack a privacy protection mechanism that prevents an accurate inference of the targeted soft biometric attribute while preserving information relevant to identity in facial images.

ACKNOWLEDGEMENTS

CONFLICT OF INTEREST STATEMENT

Authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

ORCID

Zohra Rezgui https://orcid.org/0000-0002-4363-114X
Amina Bassit https://orcid.org/0000-0002-1331-9702
Raymond Veldhuis https://orcid.org/0000-0002-0381-5235

REFERENCES

1. Rezgui, Z., Bassit, A.: Transferability analysis of an adversarial attack on gender classification to face recognition. In: 2021 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–7. IEEE (2021)
2. Demontis, A., et al.: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th {USENIX} Security Symposium (2019)
3. Zhong, Y., Deng, W.: Towards Transferable Adversarial Attack against deep Face Recognition. IEEE Transactions on Information Forensics and Security (2020)
4. Gonzalez-Sosa, E., et al.: Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and Cots Evaluation. IEEE Transactions on Information Forensics and Security (2018)
5. Ozbulak, G., Aytar, Y., Ekenel, H.K.: How transferable are cnn-based features for age and gender classification? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE (2016)
6. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (2014). *arXiv preprint arXiv:1412.6572*
7. Kurakin, A., et al.: Adversarial Examples in the Physical World (2016)
8. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial Machine Learning at Scale (2016). *arXiv preprint arXiv:1611.01236*
9. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
10. Goswami, G., et al.: Unravelling robustness of deep learning based face recognition against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
11. Mirjalili, V., et al.: Semi-adversarial networks: convolutional autoencoders for imparting privacy to face images. In: 2018 International Conference on Biometrics (ICB). IEEE (2018)
12. Huang, R., et al.: Learning with a Strong Adversary (2015). *arXiv preprint arXiv:1511.03034*
13. Abbasi, M., Gagné, C.: Robustness to Adversarial Examples through an Ensemble of Specialists (2017). *arXiv preprint arXiv:1702.06856*
14. Gu, S., Rigazio, L.: Towards deep Neural Network Architectures Robust to Adversarial Examples (2014). *arXiv preprint arXiv:1412.5068*
15. Liao, F., et al.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
16. Papernot, N., McDaniel, P., Goodfellow, I.: Transferability in Machine Learning: From Phenomena to Black-Box Attacks Using Adversarial Samples (2016). *arXiv preprint arXiv:1605.07277*
17. Xie, C., et al.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
18. Deb, D., Zhang, J., Jain, A.K. In: 2020 IEEE International Joint Conference on Biometrics (IJCB) (2019). AdvFaces: Adversarial Face Synthesis
19. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

20. Mirjalili, V., Raschka, S., Ross, A.: FlowSAN: privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. IEEE Access. 7, 99735–99745 (2019). https://doi.org/10.1109/access.2019.2924619

21. Mirjalili, V., Raschka, S., Ross, A.: PrivacyNet: semi-adversarial networks for multi-attribute face privacy. IEEE Trans. Image Process. 29, 9400–9412 (2020). https://doi.org/10.1109/tip.2020.3024026

22. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference (2015)

23. Deng, J., et al.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)

24. Liu, Z., et al.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision. ICCV) (2015)