A. Brömme, C. Busch, N. Damer, A. Dantcheva, M. Gomez-Barrero, K. Raja,
C. Rathgeb, A. Sequeira, and A. Uhl (Eds.): BIOSIG 2021,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2021

Transferability Analysis of an Adversarial Attack on Gender Classification to Face Recognition

Zohra Rezgui [1], Amina Bassit[2]

Abstract: Modern biometric systems establish their decision based on the outcome of machine learning (ML) classifiers trained to make accurate predictions. Such classifiers are vulnerable to diverse adversarial attacks, altering the classifiers' predictions by adding a crafted perturbation. According to ML literature, those attacks are transferable among models that perform the same task. However, models performing different tasks, but sharing the same input space and the same model architecture, were never included in transferability scenarios. In this paper, we analyze this phenomenon for the special case of VGG16-based biometric classifiers. Concretely, we study the effect of the white-box FGSM attack, on a gender classifier and compare several defense methods as countermeasures. Then, in a black-box manner, we attack a pre-trained face recognition classifier using adversarial images generated by the FGSM. Our experiments show that this attack is transferable from a gender classifier to a face recognition classifier where both were independently trained.

Keywords: Transferability, adversarial attacks, gender classification, face recognition.

1 Introduction

The cutting edge advances in deep learning (DL) have made computer vision problems more approachable. However, the black box nature of neural networks has made 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 a 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, [De19, ZD20] 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 works showing that different face recognition systems can be

[1] University of Twente, DMB Group, Enschede, The Netherlands, z.rezgui@utwente.nl
[2] University of Twente, DMB Group and SCS Group, Enschede, The Netherlands, a.bassit@utwente.nl

enhanced with a soft biometric classifier such as a gender classifier [Go18a]. Similarly, deep face recognition features are known to be discriminative for soft biometric classification [OAE16] via transfer learning. This association incites us to further investigate the transferability potential of adversarial attacks on models sharing the same input space but trained independently to perform different tasks. However, such hypothesis was not included in previous studies on the transferability of adversarial attacks.

In this paper, we investigate the transferability of an adversarial attack 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 an existing gradient-based attack and deep features-based defense on the gender classifier. Subsequently, we use the generated adversarial images along with those resulting from the defense against a pre-trained face recognition model to analyze the tranferability of both the attack and the defense. Our results, illustrated in Figure 1, support the transferability hypothesis of the chosen attack and defense.

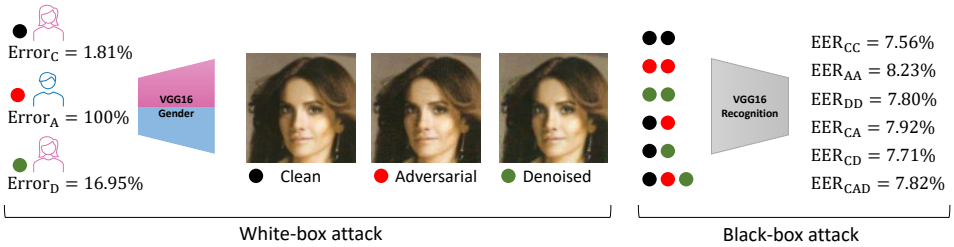


Fig. 1: Overview of the transferability hypothesis from a white-box attack on a gender classifier to a black-box attack on a face recognition classifier. The black dots refer to the clean images, the red dots refer to the adversarial images generated by the FGSM attack and the green dots refer to the denoised images generated by the defense method. *Error* refers to the binary classification error rate of the gender classifier while *EER* refers to the equal error rate of the recognition classifier.

2 Related Work

Adversarial attacks have become an active area of research as they expose the design vulnerabilities of deep learning-based models. Several white-box attacks are gradient-based such as the Fast Gradient Sign Method (FGSM) [GSS14] its iterative version (IFGSM) [Ku16], and Projected Gradient Descent [KGB16]. 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, [MDFF16] finds the minimal perturbation possible to an image that would make it misclassified, via projecting inputs on the closest classification hyperplane. Results in [Go18b] show that deep learning-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, [Mi18] uses 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 deep learning models, many defense approaches have been proposed to withstand these attacks. [GSS14] and [Hu15] 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. [AG17] 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 [GR14] by using a denoising auto-encoder on the adversarial images in order to remove the perturbations. This method has been improved in [Li18] 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 [PMG16] 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*. [PMG16] 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 SVM). [De19] analyzes 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, [Xi19] modifies the IFGSM attack by randomly resizing the images at each iteration. [DZJ19] proposes a GAN-based approach to generate synthetic adversarial samples with imperceptible perturbations against FaceNet [SKP15] and report effective results across different face recognition models. Similarly, [ZD20] reports transferability of attacks from an open-source surrogate face recognition model to several commercial target face recognition models.

3 Methodology

Let us denote X_F 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 $Y_R = \{\checkmark, \times\}$ the space of recognition labels. We consider $G : X_F \rightarrow Y_G$ a gender classifier and $R : X_F \times X_F \rightarrow Y_R$ a facial recognition classifier.

- **Attack:** An adversarial attack $f_{Adv} : X_C \rightarrow X_{Adv}$ is considered successful if for $x \in X_F$ there is an adversarial sample $f_{Adv}(x) = x_{Adv} \in X_{Adv}$ such that: $G(x) = y_G$ and $G(x_{Adv}) = \overline{y_G}$.

- **Denoising Defense:** Let $f_{Den} : X_{Adv} \rightarrow X_{Den}$ 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 Transferability:** We say that an attack f_{adv} and a defense f_{Den} are 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 $R(x_1, x_2) \neq R(x_1, f_{Adv}(x_2))$ and $R(x_1, x_2) = R(x_1, f_{Den} \circ f_{Adv}(x_2))$
- **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.

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 defense 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 defense methods in terms of the sensitivity of the recognition performance across the diverse sets of images.

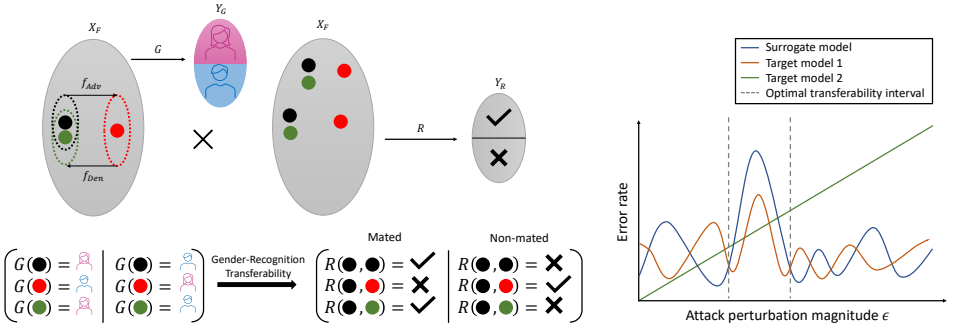


Fig. 2: Methodology overview for analyzing the transferability attack from a gender classifier (surrogate model) to a face recognition classifier (target model). The graph on the right illustrates the expected behaviour of the error variation of surrogate and target models as a function of the applied perturbation. Target model 1 is sensitive to the attack on the surrogate model unlike target model 2 that deteriorates when an image is reduced in quality, considering ε similarly as blurring or random noise.

4 Background

FGSM Attack on Gender Classifier: 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 maximizes the loss with respect to the input image [GSS14] by adding to the image a step ε in the direction of the loss gradient. An FGSM adversarial attack

$f_{Adv}: X_C \rightarrow X_{Adv}$, with perturbation magnitude $\varepsilon \in \mathbb{R}$, results in adversarial images $x_{Adv} \in X_{Adv}$ such that: $x_{Adv} = x + \varepsilon \cdot sign(\nabla J(\theta, x, y_G))$. Note that FGSM attack does not rely on equalizing the probabilities of the different input classes and thus we cannot expect an equal probability between the classes after the attack. Instead, it relies on changing the prediction by simulating a gradient ascent behaviour on the sample images.

High-level representation and pixel guided denoisers: In this paper, we consider two types of denoisers: 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, $\mathscr{L}_{\text{PGD}} = \|x - x_{Adv}\|_1$, the pixel level difference between a clean image x and its adversarial version x_{Adv}. Whereas, a HGD [Li18] reduces the loss defined as, $\mathscr{L}_{\text{HGD}} = \|f^i_{emb}(x) - f^i_{emb}(x_{Adv})\|_1$, the difference between the deep features of a clean image x and the deep features of its adversarial version x_{Adv} where $f^i_{emb}: X_C \rightarrow \mathbb{R}^n$ denotes the function describing the attacked model until its i^{th} layer that outputs a feature vector of size n.

5 Experiment and Evaluation

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 [PVZ15]. VGG16 has a straightforward architecture that comprises 13 convolution layers and 3 fully connected layers. For the denoiser, similarly as [Li18], we use a U-Net based Denoising Convolution Neural Network (DnCNN) [3] a denoising model that we will refer to in this work as UDnCNN. The 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.

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 [Li15] where identities do not overlap. For the FGSM attack against the gender classifier and the defenses experiment, we use sets A and B to train and test the gender classifier and set C to generate FGSM 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 (73,779 samples each) as well as $C_{AdvTest}$ and $C_{CleanTest}$ (18,449 samples each). 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 experiment, 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 FGSM, we collect for each adversarial image, the clean image it was derived from and its denoised image. As a result, we have a set of clean images, a set of adversarial images, and another set of denoised images of the same size (94,965 samples and 995 identities each). Those three sets are used to analyze the transferability of the FGSM attack on the face recognition classifier.

[3] https://github.com/lychengr3x/Image-Denoising-with-Deep-CNNs

Performance metrics: To assess the gender classifier performance, either before the attack and the defense 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 cosine similarity to measure the False Non-Match Rate (FNMR) at a fixed False Match Rate (FMR) of 0.1% as well as the Area Under the Detection Error Trade-off Curve (AUC-DET) and finally, the Equal Error Rate (EER).

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

FGSM attack: We run the FGSM attack on the VGG16 gender classifier using various values for the perturbation $\varepsilon \in [0.005, 0.55]$. Figure 4a shows how the 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. As our goal is to study the effect of perturbations that are imperceptible to the human, we consider the following range of epsilons $\varepsilon \in \{0.01, 0.015, 0.02, 0.025, 0.03, 0.035\}$ as it is where the classifier is most vulnerable.

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 LGD based on the logits layer.

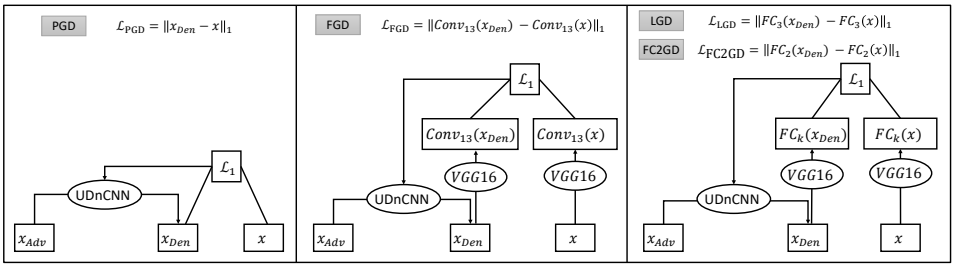
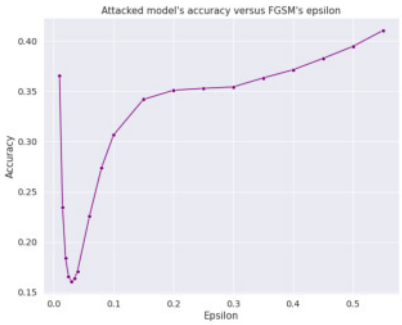


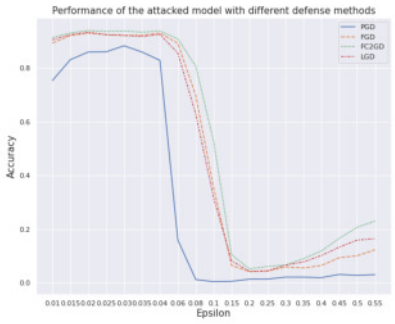
Fig. 3: Training of UDnCNN denoiser when considering the PGD defense, the FGD defense, the LGD defense (k = 3) [Li18] and when considering the FC2GD defense (k = 2).

Figure 4b shows the performance of the defense 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. PGD 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.

Comparison between the defense methods: Table 1 compares the performance of the attacked VGG16 gender classifier when applying the different defense methods (columns 2 to 5) and without (first column), over clean images (row 2) and adversarial images (row 3). For PGD and FC2GD, both considerably help 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



(a) Sensitivity of the classification accuracy of the VGG16 gender classifier upon the choice of perturbation (epsilon) used in the FGSM attack.



(b) Effect of the defense methods on the classification accuracy of the VGG16 gender classifier over increasing values of ε.

Fig. 4: Classification accuracy of VGG16 gender classifier during FGSM attack and after applying the defense methods over various attack intensities.

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 defense method performs on clean images and that LGD seems to be the most convenient method for defense so far.

	Without denoising	PGD	FGD	FC2GD	LGD
Clean Test	98.19%	95.61%	57.50%	63.48%	83.05%
Adversarial Test $\varepsilon \in [0.01,0.035]$	0%	84.34%	91.82%	93.14%	92.02%

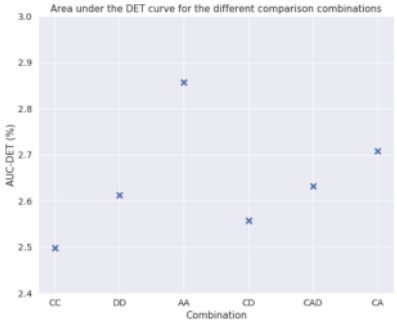
Tab. 1: Performance summary of the attacked VGG16 gender classifier in terms of accuracy with and without the defense methods

Transferability of FGSM on Face Recognition Classifier: We study the transferability of the attack from the gender classifier (surrogate) to the face recognition model (target) by performing six 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. We perform a verification entirely on the clean set (CC) to obtain a base-line performance of VGGFace before running the FGSM attack. We then perform clean/adversarial (CA) and clean/denoised (CD) verifications to evaluate the transferability of both the attack and the defense method. We also report the combinations adversarial/adversarial (AA), denoised/denoised (DD) and a blind verification on the three sets combined (CAD) to further assess the robustness of VGGFace. To realize the comparisons, we select $\sim$15 different images per subject where each image should be vulnerable to at least 3 values of ε out of 6. Table 3 summarizes the resulting numbers of mated and non-mated comparisons per epsilon and in total.

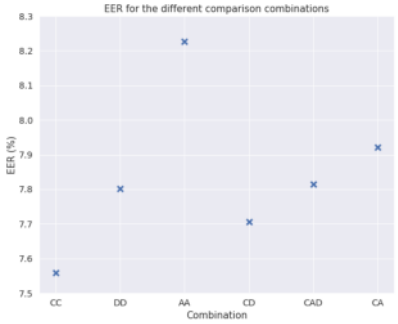
We notice in the Figures 5a, 5b and 5c that the presence of non-clean images (denoised and adversarial) regardless of the attack intensity, decreases the recognition performance. The dif-

ference 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 CAD, CA and AA, where there is 33%, 50% and 100% respectively, shows 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 defense partly compensates the performance degradation.

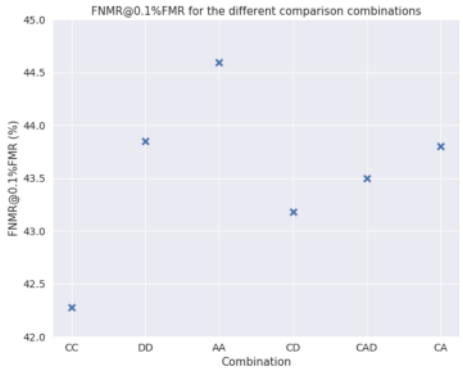
Table 2 (a) and Table 2 (b) show that for each combination involving adversarial or denoised images, the errors are the highest for the smallest perturbation 0.01 then for the subsequent increasing perturbations, the errors decrease until perturbation 0.025 before they start to increase again. This implies a low transferability of the attack in the selected epsilon range. It is possible that a more optimal range of epsilon values exist, that would result in a high transferability of the attack as shown in the illustrative graph in Figure 2.



(a) Performance in terms of AUC-DET.



(b) Performance in terms of EER.



(c) Performance in terms of FNMR@0.1%FMR.

Fig. 5: Performance measures across the different comparison combinations: **C** designates Clean, **A** designates Adversarial and **D** refers to denoised.

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
CAD	**47.28**	44.98	43.20	42.57	43.03	43.08
CA	**47.18**	44.86	43.41	42.78	43.70	43.68

(a) FNMR@0.1%FMR in percentage (%)

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
CAD	**3.04**	2.76	2.63	2.52	2.58	2.58
CA	**3.04**	2.79	2.70	2.58	2.69	2.69

(b) AUC-DET in percentage (%)

Tab. 2: Comparison performance of different combinations per epsilon in terms of FNMR@0.1%FMR in (a) and area under the DET curve (AUC-DET) in (b) where the first rows serve as a reference with only clean images.

ε	0.01	0.015	0.02	0.025	0.03	0.035	All
CC	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E3	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
DD	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E6	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
AA	M = 4.7E3 U = 1.2E6	M = 8.4E3 U = 2.6E6	M = 1.1E4 U = 3.8E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 1.4E4 U = 4.7E6	M = 6.7E4 U = 2.1E7
CD	M = 9.4E3 U = 2.4E6	M = 1.6E4 U = 5.3E6	M = 2.3E4 U = 7.7E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 1.3E5 U = 4.3E7
CAD	M = 1.8E4 U = 4.9E6	M = 3.3E4 U = 1.0E7	M = 4.7E4 U = 1.5E7	M = 5.7E4 U = 1.8E7	M = 5.7E4 U = 1.8E7	M = 5.7E4 U = 1.8E7	M = 2.7E5 U = 8.7E7
CA	M = 9.4E3 U = 2.4E6	M = 1.6E4 U = 5.3E6	M = 2.3E4 U = 7.7E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 2.8E4 U = 9.4E6	M = 1.3E5 U = 4.3E7

Tab. 3: Number of mated (M) and non-mated (U) comparisons

6 Conclusion

In this work, we studied the effect of the FGSM attack on the VGG16 gender classifier over a variety of perturbations. We also applied defense methods from the literature such as a pixel guided denoiser PGD and variants of high-level representation guided denoisers. We studied the transferability of the FGSM attack with a selected range of epsilons and the LGD defense on a pre-trained face recognition model. Our experiments confirmed that the attack and the defense of the gender classifier impact the performance of the face recognition model. This result consolidates the existing literature reporting an association between face recognition and gender classification, except that this time, this association is demonstrated through an adversarial attack and defense. We hope this work opens grounds to think about transferability of adversarial attacks between models built for different tasks while maintaining the same input space domain.

Acknowledgment

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

[AG17] Abbasi, Mahdieh; Gagné, Christian: Robustness to adversarial examples through an ensemble of specialists. arXiv preprint arXiv:1702.06856, 2017.

[De19] Demontis, Ambra; Melis, Marco; Pintor, Maura; Jagielski, Matthew; Biggio, Battista; Oprea, Alina; Nita-Rotaru, Cristina; Roli, Fabio: Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In: 28th {USENIX} Security Symposium. 2019.

[DZJ19] Deb, Debayan; Zhang, Jianbang; Jain, Anil K: Advfaces: Adversarial face synthesis. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). 2019.

[Go18a] Gonzalez-Sosa, Ester; Fierrez, Julian; Vera-Rodriguez, Ruben; Alonso-Fernandez, Fernando: Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation. IEEE Transactions on Information Forensics and Security, 2018.

[Go18b] Goswami, Gaurav; Ratha, Nalini; Agarwal, Akshay; Singh, Richa; Vatsa, Mayank: Unravelling robustness of deep learning based face recognition against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[GR14] Gu, Shixiang; Rigazio, Luca: Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068, 2014.

[GSS14] Goodfellow, Ian J; Shlens, Jonathon; Szegedy, Christian: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[Hu15] Huang, Ruitong; Xu, Bing; Schuurmans, Dale; Szepesvári, Csaba: Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015.

[KGB16] Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236, 2016.

[Ku16] Kurakin, Alexey; Goodfellow, Ian; Bengio, Samy et al.: , Adversarial examples in the physical world, 2016.

[Li15] Liu, Ziwei; Luo, Ping; Wang, Xiaogang; Tang, Xiaoou: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV). 2015.

[Li18] Liao, Fangzhou; Liang, Ming; Dong, Yinpeng; Pang, Tianyu; Hu, Xiaolin; Zhu, Jun: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[MDFF16] Moosavi-Dezfooli, Seyed-Mohsen; Fawzi, Alhussein; Frossard, Pascal: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[Mi18] Mirjalili, Vahid; Raschka, Sebastian; Namboodiri, Anoop; Ross, Arun: Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In: 2018 International Conference on Biometrics (ICB). IEEE, 2018.

[OAE16] Ozbulak, Gokhan; Aytar, Yusuf; Ekenel, Hazim Kemal: How transferable are CNN-based features for age and gender classification? In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2016.

[PMG16] Papernot, Nicolas; McDaniel, Patrick; Goodfellow, Ian: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277, 2016.

[PVZ15] Parkhi, Omkar M.; Vedaldi, Andrea; Zisserman, Andrew: Deep Face Recognition. In: British Machine Vision Conference. 2015.

[SKP15] Schroff, Florian; Kalenichenko, Dmitry; Philbin, James: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2015.

[Xi19] Xie, Cihang; Zhang, Zhishuai; Zhou, Yuyin; Bai, Song; Wang, Jianyu; Ren, Zhou; Yuille, Alan L: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[ZD20] Zhong, Yaoyao; Deng, Weihong: Towards Transferable Adversarial Attack Against Deep Face Recognition. IEEE Transactions on Information Forensics and Security, 2020.