

SEMI-ADVERSARIAL NETWORKS FOR IMPARTING DEMOGRAPHIC PRIVACY TO FACE IMAGES

By

Vahid Mirjalili

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Doctor of Philosophy

2020

ABSTRACT

SEMI-ADVERSARIAL NETWORKS FOR IMPARTING DEMOGRAPHIC PRIVACY TO FACE IMAGES

By

Vahid Mirjalili

Face recognition systems are being widely used in a number of applications ranging from user authentication in hand-held devices to identifying people of interest from surveillance videos. In several such applications, face images are stored in a central database. In such cases, it is necessary to ensure that the stored face images are used for the stated purpose and not for any other purposes. For example, advanced machine learning methods can be used to automatically extract age, gender, race and so on from the stored face images. These cues are often referred to as demographic attributes. When such attributes are extracted without the consent of individuals, it can lead to potential violation of privacy. Indeed, the European Union's General Data Protection and Regulation (GDPR) requires the primary purpose of data collection to be declared to individuals prior to data collection. GDPR strictly prohibits the use of this data for any purpose beyond what was stated.

In this thesis, we consider this type of regulation and develop methods for enhancing the privacy accorded to face images with respect to the automatic extraction of demographic attributes. In particular, we design algorithms that modify input face images such that certain specified demographic attributes cannot be reliably extracted from them. At the same time, the biometric utility of the images is retained, i.e., the modified face images can still be used for matching purposes. The primary objective of this research is not necessarily to fool human observers, but rather to prevent machine learning methods from automatically extracting such information.

The following are the contributions of this thesis. First, we design a convolutional autoencoder known as a semi-adversarial neural network, or SAN, that perturbs input face images such that they are adversarial with respect to an attribute classifier (e.g., gender classifier) while still retaining their utility with respect to a face matcher. Second, we develop techniques to ensure that the adversarial outputs produced by the SAN are generalizable across multiple attribute classifiers, including those that may not have been used during the training phase. Third, we extend the SAN architecture and develop a neural network known as PrivacyNet, that can be used for imparting multi-attribute privacy to face images. Fourth, we conduct extensive experimental analysis using several face image datasets to evaluate the performance of the proposed methods as well as visualize the perturbations induced by the methods. Results suggest the benefits of using semi-adversarial networks to impart privacy to face images while still retaining the biometric utility of the ensuing face images.

ACKNOWLEDGMENTS

I would like to acknowledge the funding source from National Science Foundation to carry out this research (NSF Grant 1618518). Furthermore, I would like to thank the support from my committee members, Dr. Xiaoming Liu, Dr. Vishnu Boddetti, and Dr. Selin Aviyente. I like to specially thank my PhD advisor Dr. Arun Ross for all his support throughout my PhD career. Lastly, special thanks to Dr. Sebastian Raschka who also helped me a lot in conducting this research.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	x
Chapter 1 Introduction	1
1.1 Privacy Motivation	1
1.1.1 Re-identification via data sharing	2
1.1.2 Targeted advertisement	4
1.1.3 Consent-free inference and ethics	5
1.2 European Union General Data Protection and Regulation (EU GDPR)	5
1.3 Biometric Recognition	9
1.3.1 Deep learning-based face matchers	10
1.3.2 Demographic attributes	11
1.3.3 Identifying different notions of demographic attributes	14
1.4 Privacy in Biometrics	15
1.4.1 Face de-identification	16
1.4.2 Demographic privacy	19
1.5 Goals and Objectives	24
1.6 Summary	26
Chapter 2 Imparting Demographic Privacy to Face Images	28
2.1 Introduction	28
2.1.1 Adversarial images	29
2.2 Proposed Method	31
2.2.1 Problem formulation	31
2.2.2 Finding perturbation direction	34
2.3 Experiments and Results	36
2.3.1 Gender perturbation	37
2.3.2 Match scores	42
2.4 Summary and Future Work	46
Chapter 3 Semi Adversarial Networks for Gender Privacy to Face Images	47
3.1 Introduction	47
3.2 Proposed Method	49
3.2.1 Problem formulation	49
3.2.2 Semi-adversarial network architecture	50
3.2.3 Auxiliary CNN-based gender classifier	55
3.2.4 Auxiliary CNN-based face matcher	55
3.2.5 Loss function	57
3.2.6 Datasets	58
3.2.7 Implementation details and software	59

3.3	Experiments and Results	59
3.3.1	Evaluation and verification	61
3.3.2	Perturbing gender	61
3.3.3	Retaining matching accuracy	63
3.4	Summary and Future Work	64
Chapter 4	On the Generalization Ability of Gender Privacy using Semi-Adversarial Networks	67
4.1	Introduction	67
4.2	Proposed Method	69
4.2.1	Ensemble SAN Formulation	70
4.2.2	Diversity in Autoencoder Ensembles	70
4.2.3	Ensemble SAN Architecture	71
4.2.4	Ensemble of SANs: Training Approach	74
4.2.5	Datasets	76
4.2.6	Obtaining Race Labels	77
4.3	Experiments and Results	78
4.3.1	Unseen Gender Classifiers	78
4.3.2	Unseen Face Matchers	82
4.4	Summary and Future Work	85
Chapter 5	FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers	86
5.1	Introduction	86
5.2	Proposed Method	89
5.2.1	Training and Evaluation of an Ensemble SAN model	92
5.2.2	FlowSAN: Connecting Multiple SAN Models	94
5.2.3	Training Procedure for Stacking SAN Models	96
5.2.4	Evaluating the FlowSAN Model	98
5.3	Experiments and Results	98
5.3.1	Performance in Confounding Unseen Gender Classifiers	100
5.3.2	Retaining the Performance of Unseen Face Matchers	104
5.3.3	Preserving Privacy	105
5.3.4	Computational Efficiency	106
5.4	Summary and Future Work	107
Chapter 6	PrivacyNet: Semi-Adversarial Networks for Multiattribute Face Privacy 110	
6.1	Introduction	110
6.2	Proposed method	113
6.2.1	Problem Formulation	113
6.2.2	PrivacyNet	116
6.2.3	Neural Network Architecture of PrivacyNet	118
6.2.4	Datasets	119
6.3	Experimental Results	122
6.3.1	Perturbing Facial Attributes	123

6.3.2	Retaining the Matching Utility of Face Images	131
6.4	Ablation study on cycle-consistency term	132
6.5	Debiasing face recognition	138
6.6	Summary and Future Work	140
Chapter 7	A deeper study on perturbations	141
7.1	Motivation	141
7.2	An Overview of Semi-Adversarial Networks	144
7.3	Understanding SAN perturbations	146
7.3.1	Studying the interpretability of SAN perturbations using CNN-fixations . .	147
7.3.2	Studying the interpretability of SAN perturbations using Grad-CAM	149
7.3.3	Summary of interpretability studies on SAN perturbations	151
7.4	Studying the effect of perturbations from the human perspective	151
7.4.1	Preliminary experiment with volunteer assessors	151
7.4.2	Human perception study with Amazon MTurk	154
7.4.3	Controlling the trade-off between the degree of privacy vs. matching per- formance	161
7.5	Summary	162
Chapter 8	Summary and Conclusions	171
8.1	Research contributions and main findings of this work	171
8.2	Limitations of this study	173
8.3	Recommendations and future work	174

LIST OF TABLES

Table 1.1:	Overview of face de-identification approaches.	19
Table 1.2:	Overview of attribute conversion methods.	20
Table 2.1:	Summary of designed experiments.	37
Table 2.2:	Gender prediction errors (%) computed using IntraFace and GCOTS on the MUCT and LFW datasets.	40
Table 2.3:	Confusion matrices for gender prediction using IntraFace, on the original MUCT dataset (top) and after perturbations guided by IntraFace (bottom).	40
Table 3.1:	Sizes of the datasets used in this study for training and evaluation. CelebA-train was used for training only, while the other four datasets were used to evaluate the final performance of the trained model.	59
Table 3.2:	Error rates in gender prediction using IntraFace and G-COTS gender classification softwares on the original datasets before and after perturbation. Note the substantial increase in the prediction error upon perturbation via the convolutional autoencoder model using opposite-gender prototypes.	62
Table 3.3:	True (TMR) and false (FMR) matching rates (measured at values of 1%) of the independent, commercial M-COTS matcher after perturbing face images via the convolution autoencoder using same (SM), neutral (NT), and opposite (OP) gender prototypes, indicating that the biometric matching accuracy is not substantially affected by confounding gender predictions.	64
Table 4.1:	Overview of datasets used in this study. The letters in the “Usage” column indicate the tasks for which the datasets were used. A: training auxiliary gender classifiers, B: SAN training, C: SAN evaluation, D: constructing unseen gender classifiers used for evaluating SAN models.	77
Table 4.2:	Error rates of the auxiliary gender classifiers on the CelebA / MORPH-test datasets. E3 (95% confidence interval: 5.46%–5.63%) performs significantly better ($p \ll 0.01$) on the MORPH dataset compared to E1 (CI95: 6.24%–6.42%) and E2 (CI95: 6.25%–6.43%). At the end, ensemble diversities are reported [80].	79
Table 4.3:	List of the nine unseen gender classifiers used for evaluating the outputs of the proposed ensemble SAN models.	81

Table 5.1:	Overview of datasets used in this study. The letters in the “Usage” column indicate the tasks for which the datasets were used. a: training auxiliary gender classifiers, b: SAN training, c: SAN evaluation, d: constructing unseen gender classifiers used for evaluating SAN models.	100
Table 5.2:	Comparing the overall average performance of six unseen gender classifiers and four unseen face matchers over the four evaluation datasets using $n = 3$ or $n = 5$ SAN models. This shows that stacking 3 SAN models results in gender anonymization $EER \approx 0.5$, while the the average matching performance is still comparable to the unmodified images as well as the matching performance on the outputs form other existing methods.	106
Table 5.3:	Comparing the overall average Equal Error Rate (EER) of six unseen gender classifiers averaged over all four evaluation datasets (CelebA-test, MORPH-test, MUCT, and RaFD), higher is better. Note that the Ens-Best method is the result of “oracle best” selected classifier from an ensemble of multiple SANs, which assumes knowledge of the gender classifier. While this is impractical in a real-world privacy application, we show the results for comparison purposes. .	109
Table 6.1:	Overview of datasets used in this study, with the number of face images corresponding to each attribute. Samples which belong to a race other than the two categories shown below, as well as those whose age-group could not be determined, are omitted.	120
Table 6.2:	Summary of the datasets used in this study, with the number of subjects and samples in the train-test partitions. The “Excluded Experiments” column indicate datasets that were removed from an experiment for the reasons given in the text.	122

LIST OF FIGURES

Figure 1.1:	Re-identification of the records in an anonymous database by linking it to a database of identifiable records via their common quasi-identifiers.	3
Figure 1.2:	EU GDPR regulates how personal data of EU citizens are processed by a company or an organization.	6
Figure 1.3:	Five criterion based on GDPR for lawful personal data processing.	7
Figure 1.4:	Matching a pair of face images: (A) a genuine pair results in match score M close to 1, (B) an impostor pair results in match score M close to 0.	11
Figure 1.5:	Examples of various demographic attributes that can be extracted from different biometric modalities.	12
Figure 1.6:	Ad-hoc face de-identification techniques used traditionally: blocking some facial components, pixelation, and blurring. These techniques damage the data utility significantly.	18
Figure 1.7:	Imparting demographic privacy to face images: confounding gender/age/race classifiers while retaining the biometric utility of a face image.	21
Figure 2.1:	Objective of our work: perturb a face image such that the gender attribute is flipped as assessed by an automated gender classifier.	29
Figure 2.2:	Workflow of the proposed method for finding per image perturbations in order to flip the gender attribute of a face image.	32
Figure 2.3:	Example of Delaunay triangulation on landmark points extracted from an input face image.	35
Figure 2.4:	Two examples showing the progress of incremental gender perturbation based on IntraFace gender classifier. (a) Input image initially classified as female (gender score= -0.4), gradually perturbed until classified as male (gender score= 0.1). (b) Input image initially classified as male (gender score= 1.7), gradually perturbed until classified as female (gender score= -0.1).	38

Figure 2.5:	Histogram of gender scores obtained by IntraFace [142] ((a),(c), and (e)) and GCOTS ((b),(d), and (f)) on the MUCT dataset. Top row shows the histogram of gender scores in the original data set before perturbations, and middle shows histograms after perturbation. For comparison, the histograms of gender scores for the method proposed by Othman and Ross [107] is shown in (e) and (f). Note that the proposed algorithm is successfully flipping the gender attribute as assessed by both gender classifiers.	41
Figure 2.6:	Distribution of genuine and impostor match scores obtained via MCOTS software (left column) and using the c VGG face descriptor [109] (right column) on the MUCT dataset; results from original dataset ((a),(b)); after gender perturbations as guided by the IntraFace gender classifier ((c),(d)); cross comparison between original and perturbed images, where perturbation was guided by IntraFace ((e),(f)); after gender perturbations as guided by the GCOTS gender classifier ((g),(h)); and cross comparison between original and perturbed images, where perturbation is guided by GCOTS ((i),(j)).	43
Figure 2.7:	ROC curves for face matching obtained using the MCOTS software ((a), (c)) and the VGG face descriptor [109] ((b), (d)). Top row shows the results obtained on the MUCT dataset, and the bottom row on the LFW dataset. Note that the recognition performance is not significantly impacted in most cases. . .	44
Figure 2.8:	Two examples of unsuccessful cases where our method fails to completely flip the gender attribute as assessed by IntraFace [142].	45
Figure 3.1:	Schematic representation of the semi-adversarial neural network architecture designed to derive perturbations that are able to confound gender classifiers while still allowing biometric matchers to perform well. The overall network consists of three sub-components: a convolutional autoencoder (subnetwork I), an auxiliary gender classifier (subnetwork II), and an auxiliary matcher (subnetwork III).	51
Figure 3.2:	Architecture of the autoencoder augmented with gender-prototype images. The encoder receives a one-channel gray-scale image as input, which is concatenated with the RGB channels of the same-gender prototype image. After the compressed representation is passed through the decoder part of the autoencoder for reconstruction (128 channels), the proto-combiner concatenates it with the RGB channels of a same-, neutral-, or opposite-gender prototype resulting in 131 channels that are then passed to a final convolutional layer. . .	53
Figure 3.3:	Gender prototypes used to confound gender classifiers while maintaining biometric matching during the semi-adversarial training of the convolutional autoencoder.	54

Figure 3.4:	Architecture of the CNN-based auxiliary gender classifier that was used during the training of the convolutional autoencoder. This classifier was used as an auxiliary (fixed) component in the final model to derive the image perturbations according to the objective function described in Section 3.2.1.	56
Figure 3.5:	Example input images with their reconstructions using same, neutral, and opposite gender prototypes from the CelebA-test (first two rows) and MUCT (last two rows) datasets.	60
Figure 3.6:	ROC curves comparing the performance of IntraFace (a-d) and G-COTS (e-h) gender classification software on original images (“Before”) as well as images perturbed via the convolutional autoencoder model (“After”) on four different datasets: CelebA-test, MUCT, LFW, and AR-face.	63
Figure 3.7:	ROC curves showing the performance (true and false matching rates) of M-COTS biometric matching software on the original images (“Before”) compared to the perturbed images (“After”) generated by the convolutional autoencoder model using same-, neutral-, or opposite-gender prototypes for three different datasets: (a) MUCT, (b) LFW, and (c) AR-face.	65
Figure 4.1:	Diversity in an ensemble SAN can be enhanced through its auxiliary gender classifiers (see Figure 4.2). When the auxiliary gender classifiers lack diversity, ensemble SAN cannot generalize well to arbitrary gender classifiers. . . .	71
Figure 4.2:	Architecture of the original SAN model [101].	72
Figure 4.3:	Schematic of the proposed ensemble of t SAN models. During the training phase, each SAN model, S^i , is associated with an auxiliary gender classifier G^i and an auxiliary face matcher M^i (common across all SANs). During the evaluation phase, the trained SAN models are used to generate t outputs $\{Y^1, Y^2, \dots, Y^t\}$	73
Figure 4.4:	Face prototypes computed for each group of attribute labels. The abbreviations at the bottom of each image refer to the prototype attribute-classes, where Y=young, O=old, M=male, F=female, W=white, B=black.	74
Figure 4.5:	An example illustrating the oversampling technique used for enforcing diversity among SAN models in an ensemble. A: A random subset of samples are duplicated. B: Different Ensemble SANs (E1, E2, and E3) are trained on the CelebA-train dataset. SANs of the E1 ensemble are trained on the same dataset with different random seeds. In addition to using different random seeds, E2 SAN models are trained on datasets created by resampling the original dataset (duplicating a random subset of the images). Finally, for E3, a random subset of black subjects was duplicated for training the different SANs in the ensemble.	75

Figure 4.6:	Four example images with their perturbed outputs using the original SAN model from Ref. [101] and the outputs of five individual SAN models. Note that the ensemble SAN generates diverse outputs that is necessary for generalizing to arbitrary gender classifiers.	79
Figure 4.7:	Data augmentation at the evaluation phase using random illumination and contrast adjustments. The left column shows the perturbed images before augmentation, and the next seven columns show the samples used for augmentation along with their gender prediction scores. Finally, average prediction scores obtained using the CNN-Merged model on these seven augmented samples are computed and denoted as CNN-Aug-Eval in the text.	81
Figure 4.8:	ROC curves of the nine unseen gender classifiers (each row corresponds to one classifier) on the perturbed images generated by each SAN model of the E3 ensemble on four evaluation datasets: CelebA-test, MORPH-test, MUCT, and LFW. Note that the gender classification performance shows a wide degree of change on perturbed samples, but in all cases, there is always one output from each ensemble degrading the performance.	83
Figure 4.9:	ROC curves of the four unseen face matchers (each row corresponds to one matcher) on the perturbed images generated by each SAN model of the E3 ensemble on four evaluation datasets: CelebA-test, MORPH-test, MUCT, and RaFD. Note that the matching performance is mostly retained except for some small degradations in the case of FaceNet and OpenFace.	84
Figure 5.1:	Illustration of the FlowSAN model, which sequentially combines individual SAN models in order to sequentially perturb a previously unseen gender classifier, while the performance of an unseen face matcher is preserved. A: An input gray-scale face image I_{orig} is passed to the first SAN model (SAN_1) in the ensemble. The output image of SAN_1 , I'_1 , is then passed to the second SAN model in the ensemble, SAN_2 , and so forth. B: An unmodified face image from the CelebA [86] dataset (I_{orig}) and the perturbed variants I'_i after passing it through the different SAN models sequentially. The gender prediction results measured as probability of being male ($P(\text{Male})$) as well as the face match score between the original (I_{orig}) and the perturbed images (I'_i) are shown.	88
Figure 5.2:	Architecture of the original SAN model [101] composed of three subnetworks: I: a convolutional autoencoder [16], II: an auxiliary face matcher (M), and III: an auxiliary gender classifier (G). In addition, the unit D computes the pixel-wise dissimilarity between input and perturbed images during model training. .	89

Figure 5.3:	Illustration of an ensemble SAN, where individual SAN models are trained <i>independent</i> of each other using n diverse, pre-trained, auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, \dots, G_n\}$), and a face matcher M that computes face representation vectors for both input face image I_{orig} and the output of the SAN model. D refers to a module that computes pixelwise dissimilarity between an input and output face image.	93
Figure 5.4:	Two approaches for evaluating an ensemble of SAN models: Combining a set of n SAN models trained in the ensemble by (A) averaging n output images, and (B) randomly selecting an output (Gibbs).	95
Figure 5.5:	An illustration of a FlowSAN model: n SAN models are trained sequentially using n auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, \dots, G_n\}$), and a face matcher M that computes face representation vectors for both input image I and the output of SAN model. Both auxiliary face matcher and the dissimilarity unit (D) use the original image along with the output of their corresponding SAN.	98
Figure 5.6:	Area under the ROC curve (AUC) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images ("Orig.") is shown (blue dashed line) as well as the perturbed samples using the three ensemble-based models (Ens-Avg, Ens-Gibbs, Ens-Best) the proposed FlowSAN model, and the face mixing approach [107] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, \dots, I'_5 \rangle$ obtained by varying the ensemble size, n . In almost all cases, stacking three SAN models results in an AUC of approximately 0.5 (a perfectly random gender prediction).	101
Figure 5.7:	A randomly selected set of examples showing input face images and their outputs from I'_1 to I'_5 using (A) the ensemble model, Ens-Avg, and (F) using the FlowSAN model.	103
Figure 5.8:	True Match Rate (TMR) values at False Match Rate (FMR) of 0.1% obtained from four unseen face matchers, M-COTS, DR-GAN, FaceNet, and OpenFace on the original images as well as perturbed outputs after applying stacking SAN models and the ensemble models (Ens-Avg and Ens-Gibbs). Note that the matchers' performance obtained after applying the first three SANs in the FlowSAN model is close to the original performance, but it further diminishes when the sequence is extended.	105

Figure 5.9:	Equal Error Rate (EER) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images ("Orig.") is shown (blue dashed line) as well as the perturbed samples using the three ensemble models (Ens-Avg, Ens-Gibbs, Ens-Best), the proposed FlowSAN model, and the face mixing approach [107] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, \dots, I'_5 \rangle$ obtained by varying the ensemble size, n	108
Figure 6.1:	The overall idea of this work: transforming an input face image across three dimensions for imparting multi-attribute privacy selectively while retaining recognition utility. The abbreviated letters are M: Matching, G: Gender, A: Age and R:Race.	111
Figure 6.2:	Schematic representation of the architecture of PrivacyNet for deriving perturbations to obfuscate three attribute classifiers – gender, age and race – while allowing biometric face matchers to perform well. (A) Different components of the PrivacyNet: generator, source discriminator, attribute classifier, and auxiliary face matcher. (B) Cycle-consistency constraint applied to the generator by transforming an input face image to a target label and reconstructing the original version.	113
Figure 6.3:	Schematic representation of the architecture of PrivacyNet for deriving perturbations to obfuscate three attribute classifiers – gender, age and race – while allowing biometric face matchers to perform well. (A) Different components of the PrivacyNet: generator, source discriminator, attribute classifier, and auxiliary face matcher. (B) Cycle-consistency constraint applied to the generator by transforming an input face image to a target label and reconstructing the original version.	114
Figure 6.4:	The detailed neural network architecture of the four sub-networks of PrivacyNet: the generator G , the discriminators D_{src} and D_{attr} , and the pre-trained auxiliary face matcher \mathcal{M} . Note that D_{src} and D_{attr} share the same convolutional layers and only differ in their respective output layers.	115
Figure 6.5:	Five example face images from the CelebA dataset along with their transformed versions using PrivacyNet and baseline-GAN models. The rows are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is specified as A0 (young), A1 (middle-aged), or A2 (old).	124

Figure 6.6:	Performance of three gender classifiers – G-COTS, AFFACT, and IntraFace – on original images as well as different outputs of the proposed model (the larger the difference the better). The results of a face mixing approach, as described in [107], are also shown. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is abbreviated as A0 (young), A1 (middle-aged), and A2 (old). The outputs of PrivacyNet, where the gender attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.	125
Figure 6.7:	Change in distribution of gender prediction scores as assessed by G-COTS on the original images as well as the outputs of PrivacyNet.	127
Figure 6.8:	Change in distribution of gender prediction scores as assessed by AFFACT on the original images as well as the outputs of PrivacyNet.	127
Figure 6.9:	Performance of the race classifier, R-COTS, on original images as well as different outputs of the proposed model. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1, and A2 (the larger the difference the better). The outputs of PrivacyNet, where the race attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.	128
Figure 6.10:	Change in distribution of gender prediction scores as assessed by G-COTS on the original images as well as the outputs of PrivacyNet.	129
Figure 6.11:	Change in age prediction of A-COTS on different outputs of the proposed model. This is with respect to the age predicted on original images for CelebA, MUCT and RaFD, and the ground-truth age values for MORPH and UTK-face. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1 and A2. The outputs of PrivacyNet, where the age attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.	130
Figure 6.12:	ROC curves showing the performance of unseen face matchers on the original images for PrivacyNet, the baseline-GAN model, face mixing [107] approach and the controllable face privacy [131] method. The results show that ROC curves of PrivacyNet have the smallest deviation from the ROC curve of original images suggesting that the performance of face matching is minimally impacted, which is desired.	133

Figure 6.13:	CMC curves showing the identification accuracy of unseen face matchers on the original images. Also shown is the range of CMC curves for the PrivacyNet model and the baseline-GAN model, along with that of the face mixing [107] and controllable face privacy [131] approaches. It must be noted that in cases where the results of PrivacyNet or GAN are not visible, the curves overlapped with the CMC curve of the original images: this means that there was no change in matching performance at all (which is the optimal case). The results confirm that transformations made by PrivacyNet preserve the matching utility of face images.	134
Figure 6.14:	Some input and output example images, based on the PrivacyNet, as well as two ablation models (PrivacyNet without cycle-consistency loss, PrivacyNet without the matching loss term).	135
Figure 6.15:	Comparison of gender classification on the original images as well as the outputs of PrivacyNet and two ablation experiments: Baseline1 is the model without the matching term, and Baseline2 is the model without the cycle-consistency loss.	136
Figure 6.16:	Comparison of matching accuracy on the original images, as well as the outputs of the PrivacyNet model along with two ablation experiments: Baseline1 is the model without the matching term, and Baseline2 is the model without the cycle-consistency loss (Top in normal scale, bottom in log-scale).	137
Figure 6.17:	Comparison of verification accuracy for different demographic groups before (original) and after applying PrivacyNet perturbations, to investigate algorithmic bias.	139
Figure 7.1:	Example perturbations added automatically to input face images by the PrivacyNet [97] model for confounding gender information with respect to automated gender classifiers. The perturbations target gender-related features, with minimal adverse effects on recognition performance.	142
Figure 7.2:	The general idea of semi-adversarial networks originally proposed by Mirjalili <i>et al.</i> [101] for imparting demographic privacy to face images. The SAN model adds perturbations to an input face image such that demographic attribute(s) cannot be reliably extracted from the output, while the output can still be used for face recognition.	145
Figure 7.3:	Heatmaps of average perturbations for Female and Male samples obtained using a privacy preserving method (PrivacyNet [97]). The results indicate that in most datasets, PrivacyNet focuses on hair and eyes for input female face images, while for input male face images. Furthermore, the area related to facial hair (beard and moustache) have opposite effects for male and female face images.	146

Figure 7.4:	Detecting important features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet [97] outputs).	148
Figure 7.5:	Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs).	150
Figure 7.6:	Cropped face images vs. full-face portrait images. Humans can easily detect gender from full-face portraits based on the peripheral information such as hair and clothing, paying minimal attention to the details of the facial texture. Therefore, predicting gender from the non-cropped original, cropped original, and non-cropped face images is trivial for humans while detecting gender from cropped face images after applying perturbations (right-most column) is more challenging.	153
Figure 7.7:	Our preliminary experiments on measuring the performance of human observers in gender classification of perturbed images using PrivacyNet. The responses from 11 participants are acquired for each image. The cases where the majority of participants have correctly predicted the gender (matches with the original label) are highlighted with green, and those in which the majority of predictions do not match with the original label are highlighted with orange.	155
Figure 7.8:	Our preliminary experiments on measuring the performance of human observers in gender classification of perturbed images using PrivacyNet. The responses from 11 participants are acquired for each image. The cases where the majority of participants have correctly predicted the gender (matches with the original label) are highlighted with green, and those in which the majority of predictions do not match with the original label are highlighted with orange.	156
Figure 7.9:	Snapshots of two example queries as were displayed to Amazon MTurk participants: Given the displayed image to the participants, they were asked to choose the gender based on their best judgment.	158
Figure 7.10:	Accuracy of gender prediction on original images, SAN outputs without cropping, and cropped SAN outputs: performance of ML-based gender predictor, G-COTS (top), and human performance using Amazon MTurk (bottom).	159
Figure 7.11:	Confusion matrices of human performance in gender classification on samples from the original datasets, as well as outputs of SAN without cropping, and cropped SAN outputs: (a) CelebA, (b) MORPH, (c) MUCT, (d) RaFD, (e) UTK-face.	160
Figure 7.12:	Operating curve showing trade-off between matching accuracy and the degree of privacy that can be obtained using PrivacyNet model. This shows the feasibility of controlling the trade-off by carefully selecting the operating point for an application.	162

Figure 7.13: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MORPH dataset.	163
Figure 7.14: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MUCT dataset.	164
Figure 7.15: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in RaFD dataset.	165
Figure 7.16: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in UTK-face dataset.	166
Figure 7.17: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MORPH dataset.	167
Figure 7.18: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MUCT dataset.	168
Figure 7.19: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in RaFD dataset.	169
Figure 7.20: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in UTK-face dataset.	170

Chapter 1

Introduction

“If this is the age of information, then privacy is the issue of our times.” [3]

1.1 Privacy Motivation

In the past decade, the growth of social media coupled with low-cost mobile devices has resulted in assimilation of massive data in different forms [3, 6]. Examples of such forms include *field-structured data* as well as *multimedia* such as images and videos, all of which contain valuable information including some *person-specific* information [5, 105, 36]. Examples of such data include a patient’s medical records in a hospital such as their medical history, a customer’s online shopping history, a user’s social media activities, data about users on a dating website/application, data obtained from mobile phones and wearable devices, questions and answers on online forums, browsing and online reading activities, *etc.* [6, 149, 79]. The concurrent developments in data-mining techniques and the exponential growth of user-generated data has enabled large-scale data analysis, thus, benefiting both the data holders and the individuals [6, 131]. On the other hand, this may lead to extracting some sensitive information about individuals (such as behaviour patterns), while they may or may not be aware of such activities [100]. This, therefore, raises privacy concerns as well as concerns regarding violations of social fairness [22, 156, 28, 75], and other unethical or security-related issues [5, 107].

Regarding the individuals' privacy, traditionally, the users were given the option to choose their desired level of privacy in form of choice and consent. For example, users could choose their gender, race, or age from a list of possible options including an option for "do not wish to say", which if selected, indicates that the particular user is not willing to provide such information [98]. However, this traditional choice and consent scheme is not adequate for today's modern digital age, since, with the advances in data mining, such information can still be inferred from other parts of the collected data. For example, a person's demographic information such as gender, age, and race (are also called demographic attributes) could be inferred from their face images stored in the databases [36, 70]. As a result, developing privacy-protecting schemes beyond traditional methods is crucial.

In the following sub-sections, we will review some implications that can result from disclosing or inferring the demographic attributes.

1.1.1 Re-identification via data sharing

While typically these databases are held privately, sharing such data across different third-party organizations or releasing certain types of data to the public is important for research and business purposes [45]. For example, researchers would need to get access to patients' data from a hospital to study patterns in disease symptoms, or sharing customers' purchasing history to business and marketing researchers. Given that such data contain sensitive information about individuals, an obvious requirement is to remove the person-specific fields that are directly associated with an individual, for example, name, address, phone number, IDs, SSN, *etc.* However, removing these identifiers is not enough to fully protect against privacy attacks [42, 130], and can lead to re-identification (associating each record to the individual through inference). In particular, as shown in Figure 1.1, the presence of a combination of *quasi-identifiers* can still lead to person

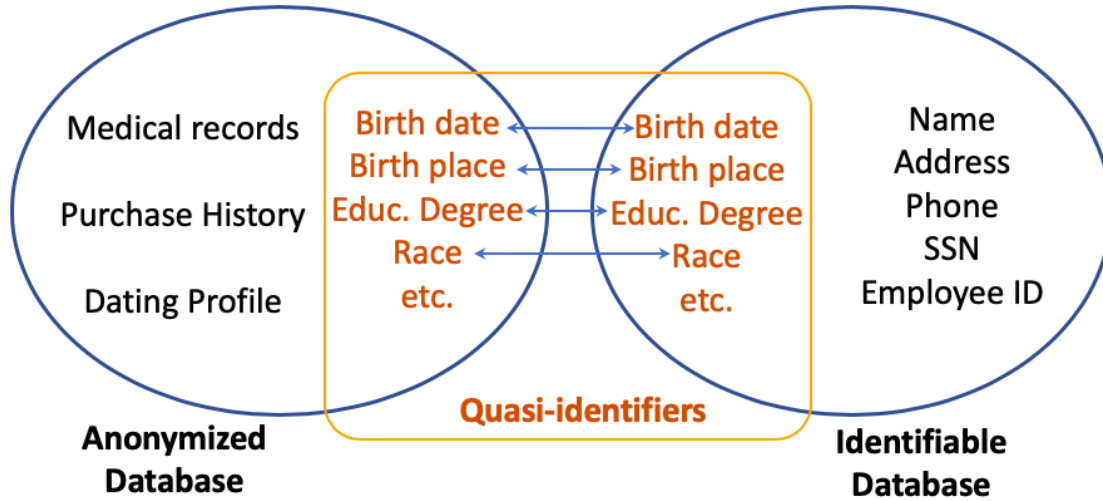


Figure 1.1: Re-identification of the records in an anonymous database by linking it to a database of identifiable records via their common quasi-identifiers.

re-identification. [4]. Such quasi-identifiers are attributes that describe a group of individuals, for example, demographic attributes such as gender, age, race, place of birth, as well as other social status information such as marital status, education degree, *etc.* Even though such quasi-identifiers do not correspond to a unique individual but rather a group of individuals, the existence of such information in the shared data can lead to person re-identification through linking attacks, and consequently, identity theft, predicting Social Security Numbers (SSN) [4], violating users privacy, and other unethical issues.

While sharing the quasi-identifiers such as age, gender, and race may not appear as a critically privacy-related issue, but we should note that different levels of information can be inferred from such shared attributes. The resulting consequences can be tangible for example, identity theft, or intangible, for example a stranger knowing a patient's medical records [3]. Acquisti and Gross [4] investigated the possibility of inferring Social Security numbers using the birth information of individuals combined with publicly available data such as Death Master File (DMF). Sweeny [134] was able to re-identify the subjects in anonymous medical records by linking the database of med-

ical records with a purchased voter registration database, while the person-specific identifiable fields were removed but the quasi-identifiers were still present. The serious privacy-related consequences such as identity theft and person re-identification solely by the use of demographic attributes further show the importance of carefully protecting such quasi-identifier attributes.

1.1.2 Targeted advertisement

Personalizing some online services and products can be beneficial for marketing purposes as well as benefiting the end-user or customer [6, 131]. For example, personalizing the news-feed on social media could make it more appealing for the user to utilize the service. Search engines and online shopping services could personalize the search results which can benefit the end-user by helping them find what they are looking for, in less amount of time. However, these benefits and advantages come with the risks pertaining to misuse and abuse of information [5]. In particular, it raises fairness issues as well as privacy issues. For example, decisions about jobs should not be based on race and gender [71]. According to [115], some job advertisers on the Facebook platform were only targeting men in certain areas, which could be unethical. Online shopping services may utilize certain demographic attributes such as age, gender, and race for targeted advertisement (which is to suggest certain products to their customers). However, this also leads to several fairness issues, as is the case that such personalization activities may narrow down the choices for the users and customers. Search engines may systematically suggest local news to the users, thereby, negatively affecting the global awareness of their users. Besides, it is possible that the vendors in online shopping may use that information to personalize the offered prices for such products.

1.1.3 Consent-free inference and ethics

In addition to the issues mentioned in the previous sub-section, inferring information about the individuals without their consent is considered a violation of their privacy [98]. Addressing such ethical issues requires new regulations to be implemented. For instance, the European Union has already compiled General Data Protection and Regulation (GDPR) [120] which is a set of regulations in order to protect the personal data of individuals. In this regard, GDPR has specific terms and conditions, according to which, any type of processing applied to individuals' data requires their consent. In the next section, we will see an overview of some terms in GDPR that are directly related to this work.

1.2 European Union General Data Protection and Regulation (EU GDPR)

Regulation (EU) 2016/679 of the European Parliament, also known as General Data Protection and Regulation (GDPR) [120], was passed in April 2016 after four years of preparation, and enforced in May 2018.¹² The overall goal of GDPR is to protect and empower the data privacy of all EU citizens, and regulates how the personal data of EU residents are processed by an individual, company, or an organization (see Figure 1.2).

Personal Data: According to GDPR, *personal data* is defined as any information related to an identified or identifiable individual, in other words, any data that can ultimately lead to the identification of a person. Some examples of personal data include name, address, phone number, IP address, cookie ID, data collected from an individual in a hospital, as well as biometric data

¹<https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679>

²<https://gdpr-info.eu/>

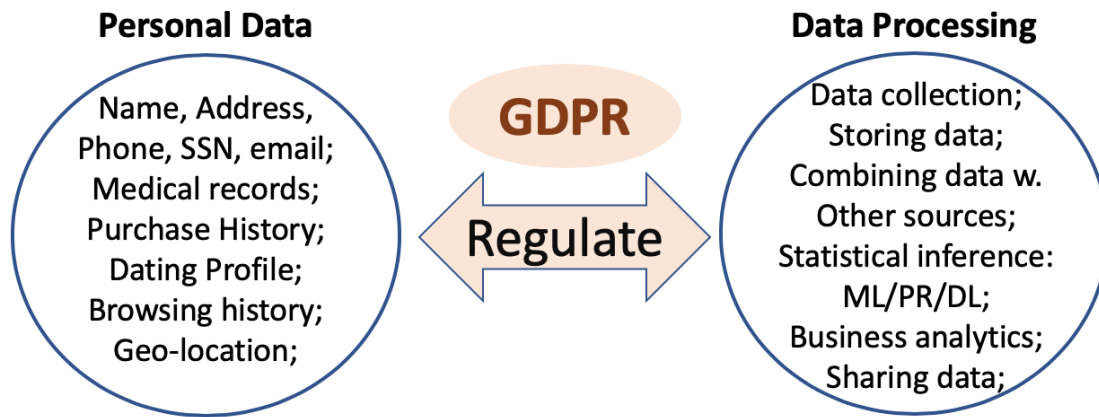


Figure 1.2: EU GDPR regulates how personal data of EU citizens are processed by a company or an organization.

such as face images, fingerprints, iris images, *etc.*

Data Processing: Any form of manual or automated operations performed on personal data constitutes data processing. It includes various operations such as data collection, structuring the data, applying some alterations and combining data with other sources. Application of machine learning and pattern recognition based sophisticated data analysis tools is also considered data processing.

Based on GDPR requirements, when applying any data processing, one needs to consider the following five criteria (see Figure 1.3):

- **Specific Purpose:** At the time of any data collection, the specific purpose for collecting and storing the data must be explicitly stated to the user/customer, and their consent must be acquired for the stated purpose. Regulations also specify how the consent is acquired from the users. For instance, users should not be deprived of certain services if they do not wish to give their consent for their participation in some data collection (unless if the collected data is essential for that service). In other words, accessing the services provided by the organization should not be conditioned on users' consent. Some examples of the specific

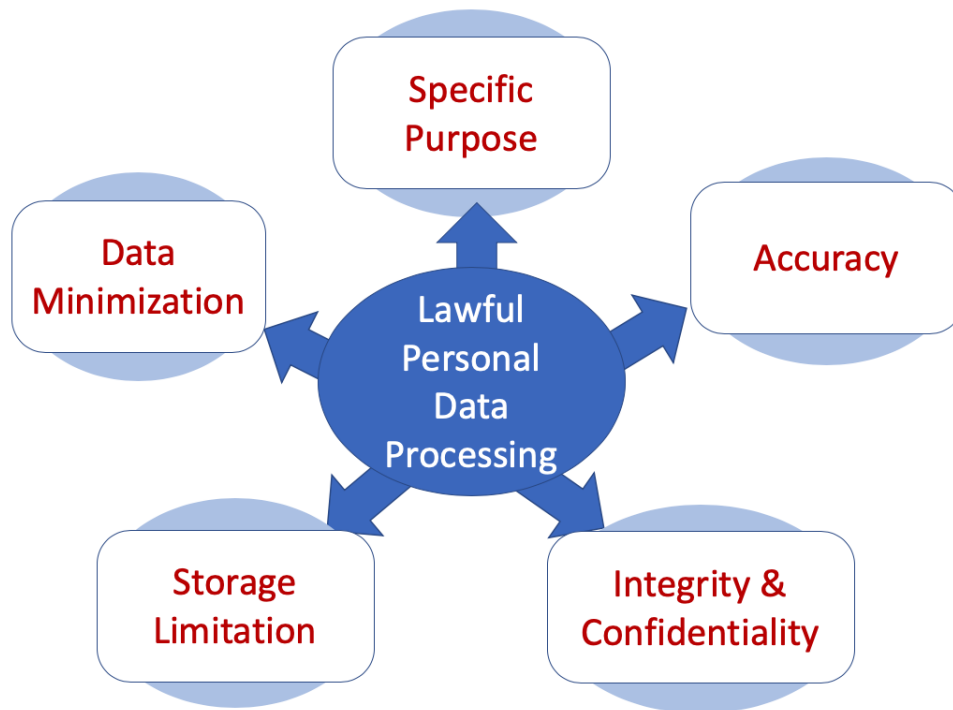


Figure 1.3: Five criterion based on GDPR for lawful personal data processing.

purpose of data collection are as follows

Example: Collecting biometric data of individuals for recognition purposes.

Example: Recording a phone conversation for quality purposes.

Among the criteria specified for legal data processing, this is probably the most important one, since this criterion (specific purpose) is used as the basis or conditions of multiple other criteria.

- **Data Minimization:** *Restricting the amount of data collection to what is essential for the stated specific purpose.* Based on the specific purpose which is stated to the user/customer and their consent for that purpose is acquired, the specified data can be lawfully collected and the data processing steps that towards fulfilling that purpose can be applied on the data. However, collecting data and applying processing beyond the stated purpose is prohibited.

Example: Based on the previous example, we can lawfully collect biometric data of the users for recognition provided that the user is informed of this purpose and his/her consent is acquired. However, collecting information that are not related to recognition, or even, inferring other demographic attributes from their biometric data is not allowed.

Example: For ride-sharing apps, accommodation, or food delivery services, collecting name, address, and credit card information are allowed while collecting race information is beyond the purpose of such services, thereby, violating the data minimization criterion.

- **Accuracy:** The data holder must ensure that the collected data is accurate and up-to-date.

Example: Career agencies, recruiters, and websites that collect resumes of candidates must ensure that the stored resumes are up-to-date.

- **Storage Limitation:** The data holder must not store the data longer than what is necessary for the specified purpose. This is also related to the *right to be forgotten*.

Example: In the previous example with collecting resumes from job seekers when the purpose is accomplished, *i.e.* the job seeker has found a job, or the client is no longer available in the job market, the data holder must remove the data of this individual.

- **Integrity & Confidentiality:** The data holder must ensure to implement appropriate means in order to *safeguard* the personal data against unauthorized access, accidental loss, or damage.

Example: Encrypting the data to avoid data leakage.

Example: Anonymizing the personal data of individuals to avoid re-identification.

Example: While carefully protecting the person-specific information (*e.g.*, name, address, phone, SSN) is necessary, the other demographic information (*e.g.* age, gender, race, place of birth, *etc.*) could still be used for re-identification. Therefore, safeguarding those

attributes is also necessary.

These five criteria summarize the requirements for lawfully performing data processing on individuals' data. However, note that simply having these requirements does not guarantee privacy, as is the case that simply having laws that prohibit crimes does not guarantee to stop crimes. For example, to prevent shop-lifting, even though the law already exists, stores still have to implement security cameras and control exit-gates to prevent shop-lifting. Similarly, companies and organizations that deal with personal data of individuals are still required to implement certain measures to prevent data breaches and privacy violation.

We also note that besides GDPR, other states and governments has also provided regulations to protect and support customers' data. California Customer Privacy ACT(CCPA) [1] is a notable example, which gives the residents of California the right to know what personal data is being collected by the organizations, accessing and requesting to delete their personal data, as well as allowing users to say no to the sale of their personal data without being discriminated against due to exercising their privacy rights.

1.3 Biometric Recognition

Biometrics is the science of recognizing individuals based on their physical or behavioral characteristics such as face images, fingerprints, iris, gait, *etc.* [70]. A typical biometric system acquires biometric data from a subject (*e.g.*, a 2D face image), extracts a feature set, and compares the feature set to templates in a database (*e.g.*, face images labeled with an identifier) in order to verify a person's claimed identity (verification) or to determine the person's identity (identification) [70, 148].

Among the different biometric modalities, face biometrics has been the most widely used

modality since it requires the least active user participation, as well as having high accessibility and public acceptance [139]. In addition, with the emergence of deep learning technology, face recognition accuracy has improved drastically [127, 148], addressing challenges posed by the variations in illumination, age, pose and expression [143, 27]. Typically, a face recognition algorithm digitizes an input face image and extracts an *embedding representation* (also known as face representation vector). The representations obtained from face images must contain salient features that are unique to each individual, and therefore, relevant for recognizing the individual in the image. However, in practice, they may contain background information, variations regarding illumination, the face pose and facial expression, as well as other information about the demographics of the person in the image, such as gender, age, ethnicity, *etc.* Face recognition models that are robust to pose, facial expression, and other background variations have been built [127, 148]. Demographic information embedded in the face representation vectors could potentially help the recognition performance, however, such information is considered *sensitive* information, and recently, new attempts have been made towards removing such sensitive information from face representations [103, 139].

1.3.1 Deep learning-based face matchers

Modern face recognition systems rely on using a *deep face matcher*, which leverages deep learning for internally extracting and comparing the face representation vectors. Given a pair of face images, a face matcher computes the embedding face representation vectors for each face image, and then a *match score* (in the range $[0, 1]$) is defined as the similarity between the two face representation vectors. A high match score (*e.g.*, close to 1) indicates a *genuine* pair of face images, meaning they belong to the same subject (Fig. 1.4-A), while, a low match score (*e.g.*, close to 0) implies the given pair is an *impostor* pair, meaning the two face images belong to two different persons

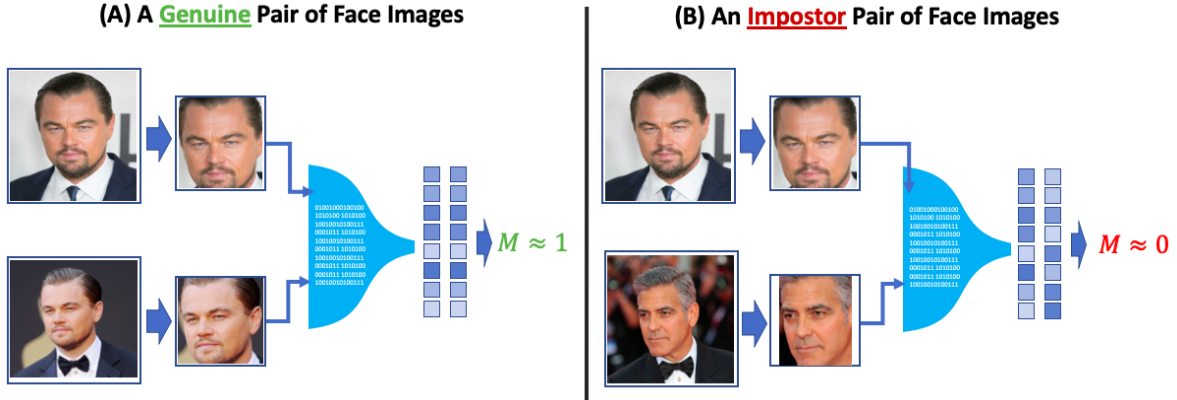


Figure 1.4: Matching a pair of face images: (A) a genuine pair results in match score M close to 1, (B) an impostor pair results in match score M close to 0.

(Fig. 1.4-B).

1.3.2 Demographic attributes

Besides recognizing individuals, biometric data including face images, iris, fingerprint, gait, and voice contain ancillary information. Such information include gender (sex), age, race/ethnicity, health characteristics such as body-mass-index (BMI), height, accessories, *etc.* [36]. These ancillary information are often called *demographic attributes* since they are not necessarily unique to an individual due to their lack of distinctiveness, and rather shared among a group [37]. Demographic attributes can be used in various applications, such as narrowing down the search space for identification [37], age-based access control, *etc.* [66, 37, 153].

While extracting certain attributes from face images such as gender [62, 89], age [29, 47], race/ethnicity [62, 87], as well as body-mass-index (BMI) [150] has been well studied in the literature, extracting certain attributes from other biometric modalities such as iris and fingerprint are still challenging [123]. Below, we will review some of these attributes.

- **Gender from face:** From an evolutionary standpoint, humans are naturally trained to detect

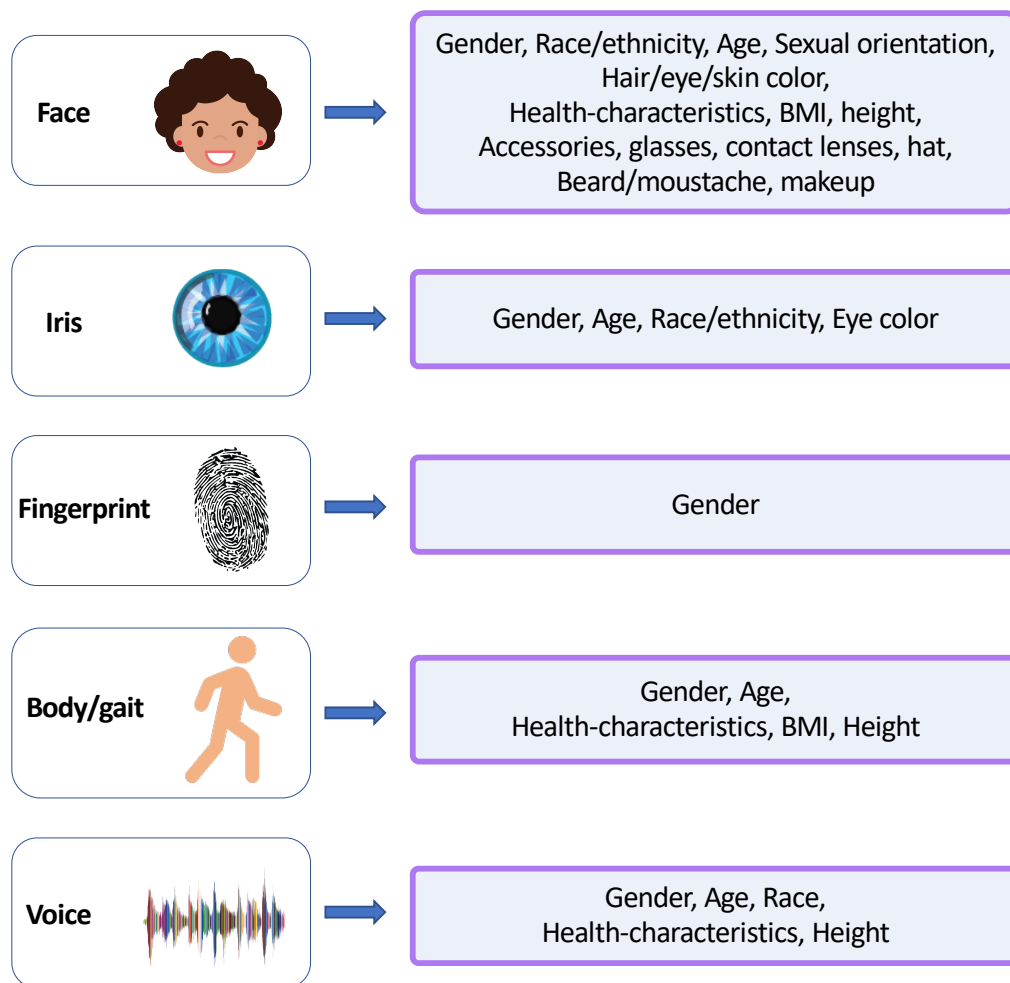


Figure 1.5: Examples of various demographic attributes that can be extracted from different biometric modalities.

the gender³. Automatic detection of gender from face images using machine learning and artificial neural networks has been vastly studied in the literature [84, 121, 89, 142, 60]. One of the early methods for this purpose used a small neural network model, coined SEXNET, which was composed of two fully connected layers that received input face images of size 30×30 , and achieved 91.2% average accuracy [50]. Most notably, Perez *et al.* [113] fused features from intensity, shape, and texture with different scales and achieved 99.13% accuracy on FERET dataset. Deep Convolutional Neural Networks (DCNN) have been able to reach human-like performance on CelebA dataset [86] without aligning face images [60].

- **Age from face:** Predicting age from face images has potential applications for demographic analyses, age-specific access control, *etc.* Age prediction from face images has been studied widely in the literature [84, 124, 15, 30, 47, 48]. Niu *et al.* [106] have used ordinal regression with extended binary classification as binary rankings [85], which was able to outperform metric regression for age estimation. Cao *et al.* [24] further extended the ordinal regression formulation to enforce consistency among the individual binary classifiers, and obtained state-of-the-art performance.
- **Gender and race from iris:** Thomas *et al.* [140] were the first to investigate gender prediction from near-infrared iris images. Demographic prediction from iris texture has been vastly studied, including gender and race prediction [82, 20], as well as predicting eye-color from NIR iris images [19]. Recently, Tapia and Aravena [136] have explored gender prediction using deep learning methods.
- **Gender from fingerprint:** Gender estimation from fingerprints has gained considerable attention in the biometrics community [129, 119]. In this regard, discrete-wavelet trans-

³Past literature has used the terms “gender” and “sex” interchangeably. In this work, we use gender as a binary class (“male” and “female”), but wish to point out that many categories of gender have been identified

formation has been used as a viable technique for feature extraction. [61, 55] Gnanasivam and Muttan [49] combined wavelet-based features and singular value decomposition and achieved 87.52% prediction accuracy.

- **Attributes from voice:** Gender prediction from voice signal is motivated by natural differences between male and female voices, especially, given the difference in the range of their voice frequencies [151]. Although gender prediction from voice signals obtained in real-world conditions with background noise is very challenging [114, 5], it is shown that 100% classification accuracy can be obtained using clean data [32]. Walavalkar *et al.* [147] have compared the performance of different classifier models in gender recognition for voice signals. Pronobis and Magimai-Doss [114] have evaluated the gender prediction performance using fundamental frequency (F_0) and cepstral features under clean and noisy conditions and concluded that under noisy conditions cepstral-based features work better than F_0 features. In addition to gender, biological studies show the relationship between the vocal and the body size, which is justified with evolutionary perspectives [43], leading to predicting height from voice.

1.3.3 Identifying different notions of demographic attributes

As we consider three demographic attributes in this work, viz., gender, age and race, it is worth noting that there could be different definitions of these attributes in different societies and communities. Therefore, in this sub-section, we will clarify our usage of these attributes, while admitting that our usage can be widely different from society’s evolving understanding of these attributes.

Gender versus sex From a societal perspective, gender can have a number of categories. For example, Facebook currently provides a list of 58 possible options for gender identity. This brings

a point of distinction between the biological sex of an individual and their gender. In this work, we consider gender as a binary label, *i.e.* Male or Female.

Chronological age versus Biological age: The chronological age of a person is the number of years since his/her birth, while the biological age can be same, lower or higher than the chronological age, due to the genetic effects and environmental conditions. In this work, we consider chronological age, which can be accurately measured according to the birth-year of the person.

Race versus Ethnicity: Similar to gender and sex, there is a major distinction between race and ethnicity. While race of a person is based on their genetic ancestry, ethnicity is often defined on the basis of cultural, religious, nationality and language factors. A person can choose their ethnicity according to how they identify themselves. In this work, however, we consider race as determined by the physical traits (although, the labels are not drawn from the genetic data.)

1.4 Privacy in Biometrics

Based on what we have seen so far, biometric data can reveal a lot of information about the individuals, ranging from person-specific information to demographic attributes. Leveraging the person-specific information in biometric data and the ability to use them for recognizing individuals makes them a suitable choice for authentication systems. Using biometrics for authentication is more advantageous than the traditional password-based authentication, due to ease-of-use, omitting the need to memorize long passwords, higher security since they cannot be shared or stolen, *etc.* [104]. On the other hand, if the biometric data stored for verification are exposed in a data breach, such data can be misused by adversaries for undesired purposes with serious consequences. Furthermore, in contrast to traditional authentications such as passwords and pins, the biometric traits of an individual cannot be altered. Therefore, if they are exposed in a data-breach, they are exposed

permanently. This results in serious security and privacy consequences, and therefore, necessitates the need to properly secure the biometric data using privacy-preserving schemes.

Privacy in biometrics literature covers a broad range of topics, from security and cryptography to face de-identification and demographic privacy. Regarding the security aspects, extensive work regarding protecting the privacy of biometric data has been done [118, 104]. Natgunanathan *et al.* [104] conducted a comprehensive study on the performances and shortcomings of existing of privacy-preserving biometric schemes (PPBS), including encryption-based schemes [74], cancelable-based schemes [118], and multimodal and hybrid-based schemes [111, 146, 14, 112]. These approaches are applicable for protecting the biometric data in data breach scenarios, aiming to prevent an adversary or a cracker from gaining access using stolen biometric data. However, in this work, we mainly deal with confounding the extraction of certain information from biometric data without compromising the recognition utility of the data, which has applications different from those of PPBS. Given our primary modality is face images, here, we review the existing approaches under two scenarios: 1) face de-identification for preserving the anonymity of subjects in an image or video 2) confounding demographic attributes.

1.4.1 Face de-identification

Advances in image and video acquisition technology accompanied by computing hardware storage have made it possible to capture a massive amount of visual data from surveillance cameras, hand-held devices, and cameras installed on autonomous cars, *etc.* The captured images and videos have various applications in person re-identification, behavioral analysis, crowd-activity detection, *etc.* However, as society embraces this technology, the privacy concerns grow as well [7]. The captured images and videos in public places contain subjects from whom we may not have or will not be able to acquire their consent prior to sharing or publicizing the images and videos. This raises

several privacy issues, identity theft, and security concerns. Therefore, it is necessary to be able to automatically remove the person-specific features from such data. Face de-identification is an attempt to protect the privacy of subjects in an image or video without hurting the data utility, by removing as much identity information as possible while other non-identity information are preserved.

For example, in surveillance videos, detecting the activities in a public place suffices in most cases, without needing the identity of subjects in the scene [144]. As another example, when taking a video in a public place, where the scene contains subjects, publishing this video on TV, social media, or other public domains may require consent from all subjects in the scene which can be infeasible in many cases. An alternative way to preserve the privacy of those subjects is face de-identification, which can be used to remove the facial features that could potentially lead to the identification of those subjects in the video.

Naïve face de-identification

The ad-hoc solution for face de-identification is to naïvely distort images by blurring or pixelation [21] (see Fig. 1.6). While these ad-hoc methods are able to prevent humans from recognizing the subjects, it was shown by [59, 57] that these techniques often provide very poor privacy-protection. Furthermore, they distort the images and obscure the facial details, which makes the output images lose their utility in various classification tasks. As a result, various face de-identification techniques are proposed in the literature that can overcome the limitations of these ad-hoc methods.

Face de-identification based on k-same family of methods

First, Newton *et al.* [105] proposed k-same technique, which is based on the k-anonymity method, and computes the average color and texture of k faces. Given a set of face images

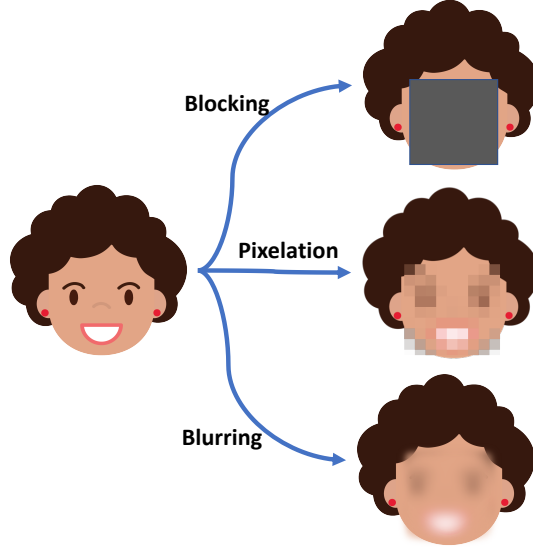


Figure 1.6: Ad-hoc face de-identification techniques used traditionally: blocking some facial components, pixelation, and blurring. These techniques damage the data utility significantly.

$S_0 = \{I_1, I_2, \dots, I_n\}$, k-same generates a new de-identified set $S_d = \{I_1^{(d)}, I_2^{(d)}, \dots, I_n^{(d)}\}$, where each de-identified image $I_i^{(d)}$ can relate to k other images in S , and therefore, it can be shown that the best possible recognition rate is reduced to $1/k$. However, this algorithm has some limitations such as introducing undesirable artifacts, ghosting effects, losing the demographic utility, and disturbing the facial expression, *etc.* Gross *et al.* [58] proposed k-Same-Select in order to preserve the demographic utility, in which they proposed selecting k images from the same cohort. As a result, they showed that their method was able to preserve the facial expression.

Later, Gross *et al.* [59] improved this technique and proposed k-same-M by incorporating Active Appearance Model (AAM) [34]. In 2008, face-swapping was introduced by Bitouk *et al.* [17] where a source face is seamlessly blended with candidate images which were similar to the query image in terms of pose and appearance. A generative multi-factor model was proposed in [57] that separates identity and non-identity components of a face image prior to the de-identification process. Their results improved the expression classification performance on the de-identified im-

ages. Jourabloo *et al.* [72] adopted k-Same algorithm and proposed Attribute-Preserved Face De-identification (APFD) using an optimization scheme in order to find the optimal set of weights such that face attributes are preserved. Recently, deep-learning-based schemes for face de-identification have also emerged [152, 93, 92]. Most notably, Meden *et al.* [93] proposed k-Same-Net, which leverages a generative deep neural network for face de-identification.

Table 1.1: Overview of face de-identification approaches.

Authors	Year	Model	Highlighted Features
Newton <i>et al.</i> [105]	2005	k-Same	Provable privacy Undesirable artifacts
Gross <i>et al.</i> [58]	2005	k-Same-Select	Selecting k input images based on attributes Preserve facial expression
Gross <i>et al.</i> [59]	2006	k-Same-M	Active-Appearance Model (AAM) Reducing the artifacts
Gross [56]	2008	Multi-Factor (MF)	Separating identity & non-identity components
Du <i>et al.</i> [41]	2014	GARP-face	Preserving Gender, Age and Race
Jourabloo <i>et al.</i> [72]	2017	APFD	Optimizing the weights to preserve attributes
Medn <i>et al.</i> [93]	2018	k-Same-Net	Neural-Network-based de-identification Preserves demographic utility

1.4.2 Demographic privacy

As we mentioned previously (Section 1.3.2), demographic attributes such as age, gender, ethnicity, *etc.* can be automatically extracted from biometric data. Another aspect of privacy is to suppress the automatic extraction of such attributes from biometric data while preserving their recognition capability [123]. This has several privacy applications, such as preventing the misuse of demographic information, preventing targeted advertisement without users' consent, preventing profiling users based on their demographic attributes, *etc.* Table 1.2 shows an overview of existing

techniques for attribute conversion from face images.

Table 1.2: Overview of attribute conversion methods.

Authors	Year	Proposed Model	Highlighted Features
Rowland and Perrett [125]	1995	Prototypes of males/females	Ghosting artifacts Too much changes
Tiddeman <i>et al.</i> [141]	2006	Prototypes in the wavelet domain	Improved ghosting effects
Suo <i>et al.</i> [133]	2011	Component-based	Seamless; No ghosting effect Too much changes
Othman and Ross [107]	2014	Mixing faces of opposite gender	Generating multiple outputs Ghosting artifacts
Sim and Zhang [131]	2015	MMDA	Considering multiple-attributes Loosing recognition utility
Mirjalili and Ross [100]	2017	Adversarial examples	Non-transferable
Choi <i>et al.</i> [33]	2017	StarGAN	Realistic-looking outputs Loosing recognition utility
Mirjalili <i>et al.</i> [101]	2018	SAN	Transferable Non-generalizable
Chhabra <i>et al.</i> [31]	2018	Adversarial examples	Multiple attributes Non-transferable
Mirjalili <i>et al.</i> [98]	2019	FlowSAN	Generalizable
Mirjalili <i>et al.</i> [97]	2020	PrivacyNet	GAN + an auxiliary matcher Multi-attribute privacy

One of the earliest attempts in this regard was done by Rowland and Perrett [125] for gender conversion of facial images. In their work, they proposed finding the prototypes of male and female faces and use these prototypes as the gender conversion axis. In their work, they utilized Active Appearance Model (AAM) [34] for aligning face images together and finding the prototypes of face shapes, as well as the color/texture prototypes. Later, Suo *et al.* [133] proposed a component-based approach in which face images are decomposed into several facial components. Then, for a given face image, they replaced each facial component with that of the closest match from the opposite gender group. They showed that the identity of face images are preserved since the closest

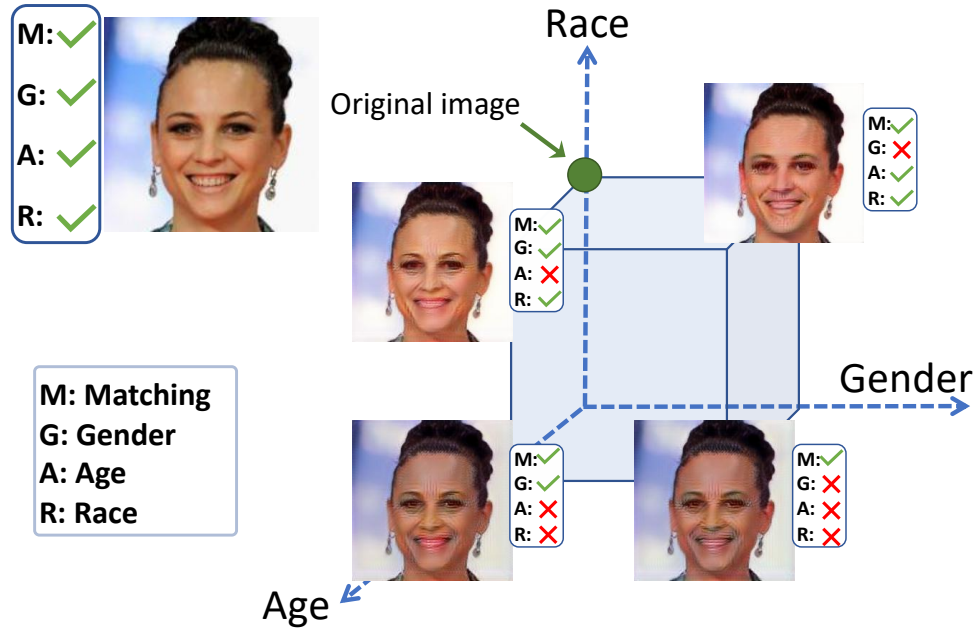


Figure 1.7: Imparting demographic privacy to face images: confounding gender/age/race classifiers while retaining the biometric utility of a face image.

matching components were used to build the new face.

While these existing techniques focus mainly on converting the gender of a face image and ignore the recognition capability, Othman and Ross [107] developed the notion of demographic privacy in which gender information is confounded while recognition as the primary biometric utility is retained. They proposed a face-mixing approach where a source face image is combined with a candidate image from the opposite gender. Therefore, based on k -anonymity principle, the best possible gender classification rate is at most $1/k$ (with $k = 2$). Later, Sim and Zhang [131] extended this notion of face privacy to include multiple attributes, age, gender, and ethnicity. They proposed a technique to map an input face image into orthogonal axes corresponding to age, gender, and ethnicity using Multi-Modal Discriminant Analysis (MMDA). Then a new image can be reconstructed using linear interpolation based on the desired degree of preserving or flipping each attribute.

With the discovery of adversarial examples [135, 54, 126], Mirjalili and Ross [100] investigated

the possibility of generating adversarial examples in order to confound gender classifiers while the recognition utility of face images are retained. Later, Chhabra *et al.* [31] extended this approach to multi-attributes including gender, age, and ethnicity. While both these works were able to confound specific attribute classifiers that were used to derive the perturbations, they failed in transferring the perturbations to unseen attribute classifiers. In order to overcome this limitation, Mirjalili *et al.* [101] proposed a convolutional autoencoder, coined Semi-Adversarial Networks (SAN), which generates perturbed samples that are transferable to unseen gender classifiers. In their technique, they used two auxiliary networks to derive the cost function, (1) a CNN-based gender classifier, which its output is used to derive gender-based perturbations, and (2) a CNN-based auxiliary face matcher, to ensure that the perturbed face image still matches with its original version.

Furthermore, to address the generalizability issue of SAN models, Mirjalili *et al.* [99] proposed an ensemble of SAN models, which were trained on a set of diverse gender classifiers. Generalizability to unseen gender classifiers was further improved in [98] by sequentially training SAN models, and stacking them for evaluation.

Generative Adversarial Networks (GANs) [53] have also been successfully used for converting certain facial attributes. Choi *et al.* [33] proposed a model called StarGAN [33], in which they used a cycle-consistent conditional-GAN to convert any selective combination of five facial attributes. However, conventional GANs are not able to preserve the recognition utility of face images. Mirjalili *et al.* [97] proposed a GAN model called PrivacyNet, which applies a constraint on recognition via a pre-trained auxiliary face matcher so that the output face images still match with their original version. They showed that this technique can overcome the limitations of conventional GANs in preserving the recognition capability.

The summary of contributions of this research is as follows:

- Formulating the problem of demographic privacy that is to avoid the extraction of demo-

graphic attributes while retaining the recognition capability.

- Investigating an *efficient* method for generating adversarial examples to fool a particular gender classifier in a complete *blackbox* scenario while retaining face matching utility.
- Designing a neural network model, called Semi-Adversarial Network (SAN), for generating adversarial examples that are *transferable* to unseen gender classifiers.
- Designing two algorithms for making the SAN model *generalizable to arbitrary unseen gender classifiers*;
- Combining the SAN idea with Generative Adversarial Networks (GANs) for generating perturbations for *multiple attributes selectively* that are generalizable to unseen attribute classifiers.

Next, we will describe some practical application of this work. Potential applications of this work are as follows

- **Imparting demographic privacy to face images for users of online shopping and social media websites:**

When users are signing up for a social media or online shopping website, they are typically asked for their demographic information, but with an option to not disclose such information. However, in such cases, the organization that stores users information has the ability to extract the demographic information using their face images.

- **Protecting the information related to demographic distribution of users or patients while sharing data across different organizations:**

Let us consider an example for an online shopping application where users have provided their personal information to use the services. The service provider uses face recognition

services of a third party. From the perspective of a business competitive strategy, it is desired for the online shopping service to protect the demographic distribution of their user-base from the competitors. Therefore, while the shopping service needs to share the data with the third party for face recognition purposes, they also need to protect the information regarding the distribution of their users.

- **Prevent user-profiling and targeted advertisement in compliance with GDPR:**

Let us consider services that store and use personal information of users for specific purposes that are already stated to the users. Note that GDPR allows users to freely accept or deny terms of use, which includes whether or not disclose their demographic information. When users opt out of targeted advertisement, service provider can no longer profile users based on their demographic attributes for targeted advertisement or other business purpose, and removing the demographic attributes would further prevent the service provider from such activities.

1.5 Goals and Objectives

As we reviewed the importance of preserving the privacy of biometric data, our objective in this work is to design and implement methods to preserve the privacy of face images. In particular, we focus on developing methods for modifying face images such that the biometric utility of the face images are preserved, while extracting demographic attributes is confounded. The detailed description of the proposed algorithms are provided in the following chapters:

- **Chapter 2:** We investigate the possibility of imparting gender privacy to face images using additive perturbations. Our objective is to derive perturbations for a specific gender classifier, such that the performance of the gender classifier is negatively impacted. Given a face image

and its ground-truth gender, this objective is accomplished by mixing patches of the face image with a candidate image from an opposite gender using Delaunay triangulation.

- **Chapter 3:** The adversarial perturbations derived using the proposed method in Chapter 2 are not transferable to an unseen gender classifier. With “unseen classifier”, we refer to a classifier that was not used for generating the perturbation or used for training a model that derives such perturbations. Therefore, here we focus on developing a CNN-based model, coined SAN, for generating perturbations that are transferable to such unseen gender classifiers. The SAN model consists of a convolutional autoencoder, which is trained using an auxiliary gender classifier and an auxiliary face matcher.
- **Chapter 4:** While the SAN model developed in Chapter 3 is shown to generate perturbations that are transferable to unseen gender classifiers, the issue of generalizability to *arbitrary* unseen gender classifier still holds. The SAN model relies heavily on an auxiliary gender classifier during its training to derive the perturbations, which means the generalizability of perturbations might be affected. In this regard, we propose using an ensemble of SAN models in order to generate multiple perturbed outputs for each input face image, to address the generalizability of the SAN model.
- **Chapter 5:** To address the generalizability issues of the SAN model, we further extended the ensemble SAN model proposed in Chapter 4, and investigated the possibility of stacking multiple SAN models to form a stronger model. We designed a new model, coined FlowSAN, in which we train SAN models sequentially and stack them such that the output of SAN_{i-1} is given as input to SAN_i . As a result, FlowSAN can successively degrade the performance of unseen face-based gender classifiers.
- **Chapter 6:** While the methods proposed in the previous chapters mainly consider the gender

attribute, in this chapter, we extend the SAN model to include three demographic attributes: gender, age, and ethnicity. To further improve the performance and the visual quality of the output face images, we use the Cycle-GAN algorithm, trained on CelebA and MORPH datasets.

- **Chapter 7:** In this chapter, we introduce two open problems with regards to imparting privacy using the SAN model and describe our approach to investigate these problems. First, we investigate the interpretability of SAN perturbations, and try to understand why SAN has resulted in the specific perturbations for a particular image. The second problem we introduce is to study how humans perceive the perturbed images and investigate whether human observers could reveal the attributes which the SAN model is trying to confound. We have used Amazon MTurk for this study and investigated the performance of MTurk participants in gender classification on original images (before perturbation), as well as outputs of our model. In addition, we hypothesize that human performance in gender classification on face images strongly relies on the presence of peripheral information. In order to verify our hypothesis, we cropped face images to exclude the peripheral information in face images, and studied the performance of human observers.

1.6 Summary

In this chapter, we first reviewed the importance of data privacy and motivations for preserving the privacy of individuals. We first introduced the consequences of sharing data that contain sensitive personal information about individuals. Targeted advertisement, inferring sensitive information without consent from users, and other unethical issues were discussed. Such issues demand proper regulations by the legislative body, as well as scientific research to investigate and develop tools

for protecting the privacy of individuals. In this regard, the European Union passed the regulations known as EU GDPR, for protecting the privacy of individuals, which define what constitutes a lawful processing of individuals' data.

Then, we reviewed existing techniques for data-anonymization and privacy-preserving schemes for field-structured data. Shifting our focus to biometric data, and in particular, face images, we considered two aspects of face privacy in the literature: 1) face de-identification which tries to remove the identity information from a face image while preserving its utility, 2) demographic privacy which is to suppress the automatic extraction of demographic attributes from a face image, while retaining its recognition utility. The existing techniques and their drawbacks on both aspects of face privacy are described. Then, the objective of our work for imparting demographic privacy to face images is provided.

Chapter 2

Imparting Demographic Privacy to Face Images

Portions of this chapter have been published in:

- V. Mirjalili, A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender", International Joint Conference on Biometrics (IJCB 2017).

2.1 Introduction

In Chapter 1, we have discussed the privacy issues related with extracting demographic attributes such as gender, age and ethnicity from biometric data of individuals. In this regard, automatic extraction of such attributes without users' consent is considered a privacy violation. In this chapter, we investigate the possibility of applying some perturbations to face images in order to confound machine learning models from extracting these attributes, while at same time, biometric utility of such data is still retained. In particular, we focus on transforming a face image such that it can be used for recognition purposes by a biometric matcher, but information such as gender and race cannot be reliably estimated by a demographic attribute classifier. Specifically, we consider flipping the gender attribute of a given face image, as shown in Fig. 2.1. Given the original input image (left column), applying the perturbations in the middle column results in an output image

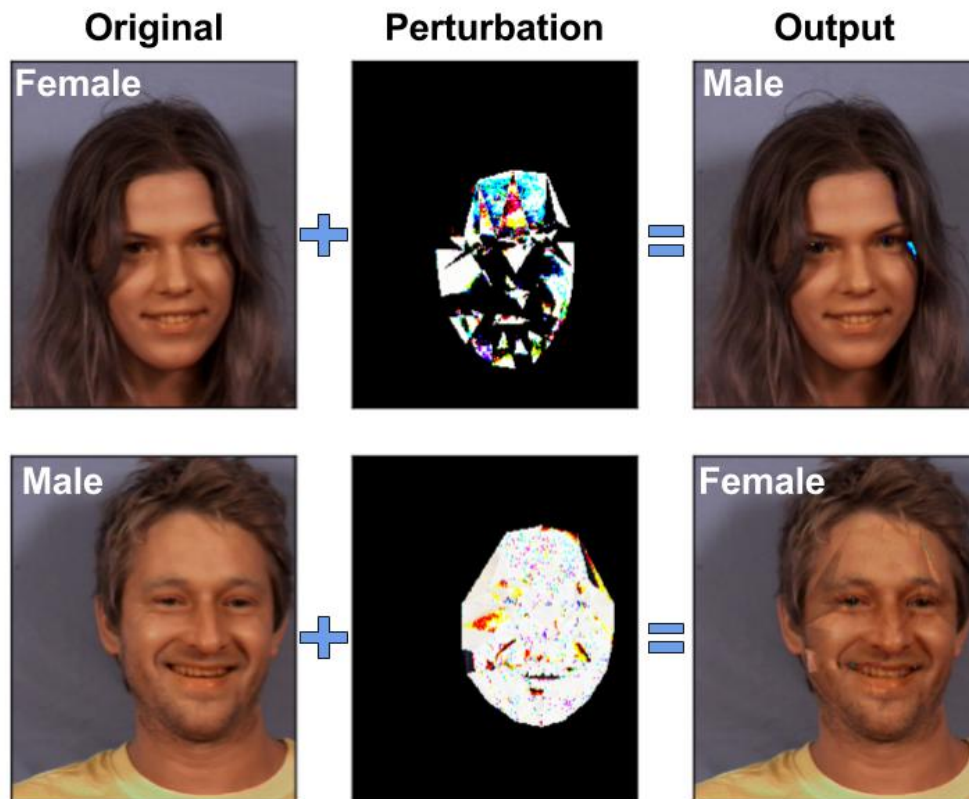


Figure 2.1: Objective of our work: perturb a face image such that the gender attribute is flipped as assessed by an automated gender classifier.

(right column) whose gender, as assessed by an automated gender classifier, is flipped to male (respectively, female) in the top (respectively, bottom) row. Our goal is to find these perturbations for an input face image.

2.1.1 Adversarial images

Szegedy *et al.* [135] discovered that neural networks are vulnerable to adversarial examples. They defined adversarial examples as input data that are perturbed slightly in such a way that the network will misclassify them. They proposed a box-constrained optimization problem to find the smallest perturbations required to modify the input such that the output target label is changed. The perturbations applied to input images are barely perceptible to the human eye. Further, they showed

that these perturbations are robust in that the same perturbations can cause misclassification on different networks with the same topology but trained on different subsets of data. Later, Goodfellow *et al.* [54] proposed a fast-gradient sign method for generating adversarial perturbations, and observed that adversarial examples generalize well to different neural network models with different architectures or trained using disjoint training sets. Rozsa *et al.* [126] proposed a fast flipping attribute (FFA) algorithm for generating adversarial examples, which leverages backpropagation and the negative gradients of the decision layer of a neural network to perturb input examples. They found that input examples that were misclassified naturally (referred to as natural adversarial examples) could be correctly classified after perturbations using the fast flipping algorithm.

Focusing on the gender attribute of a face image, besides the adversarial technique that leverages neural networks, the previously mentioned attribute conversion methods relied on either using prototypes for different classes or fusion of facial components. Considering the fact that the primary objective of these methods was to modify the apparent gender of an input face image as assessed by a human observer, the output contains perceptual changes compared to the input face image. As a result, a human observer is potentially fooled into assigning an incorrect gender label to the modified image.¹ These methods modify the face and texture of the input face image, without explicitly determining the specific features of the face that is being exploited by the attribute classifier. Therefore, these methods induce unnecessary changes to the input face image which may not be directly affecting attribute conversion. As opposed to previous methods, in this work, our goal is to apply changes that *specifically* target a particular attribute (in this case, flipping the gender attribute). Given a specific face matcher and a specific face attribute classifier, we propose an attribute flipping algorithm to iteratively perturb face images and show that it is possible to

¹It is worthwhile to note that the focus of our work is on fooling an *automated classifier* as opposed to a *human observer*.

generate adversarial images which are misclassified by a robust attribute classifier (*e.g.*, gender classifier). We show that, in most cases, slightly perturbing a few pixels in the input can confuse the gender classifier, while retaining the utility of a biometric matcher. Note that we target a specific gender classifier; in other words, we do not intend to flip the gender attribute as assessed by a human observer. As a result, if a human is monitoring the images, they may be able to detect the correct gender. In summary, the contributions of the proposed work compared to previous methods are as follows:

- The proposed method is generalizable to work with any biometric matcher and facial attribute classifier;
- The proposed method finds perturbations that specifically target flipping an attribute, resulting in imperceptible changes (in most cases).

2.2 Proposed Method

2.2.1 Problem formulation

We assume that we are given a binary ² attribute classifier, f , that outputs a classification score for an input image $X \in R^n$, *i.e.*, $f : R^n \rightarrow R$. The class label is computed as $\text{sign}(f(X))$. Furthermore, we denote $M(X_i, X_j)$ as a biometric matcher that computes the match score between face images X_i and X_j . Our goal is to efficiently find a minimally perturbed image $X' = \phi(X)$ such that $g(X)f(X') < 0$, where $g(X) \in \{-1, 1\}$ is the ground-truth attribute label (*e.g.*, female=-1; male=1). Here, ϕ is the transformation function that modifies the image. Therefore, we define the

²As explained in Chapter 1, in this work, we assume that gender has 2 labels; however, it must be noted that societal and personal interpretation of gender can result in many more classes. Facebook, for example, suggests over 58 gender classes: <https://goo.gl/lwTJhr>

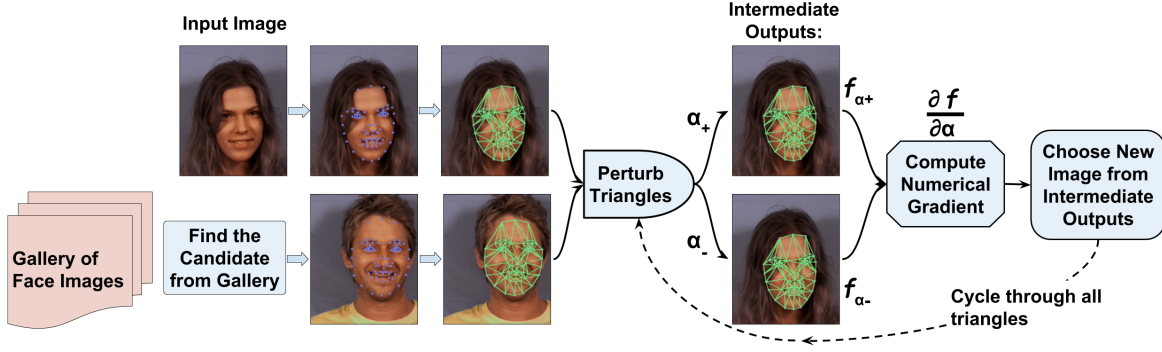


Figure 2.2: Workflow of the proposed method for finding per image perturbations in order to flip the gender attribute of a face image.

following cost function for optimization:

$$J(X', X) = f(X') \text{ sign}(g(X)). \quad (2.1)$$

This cost function is designed to give a positive value if the estimated attribute score of the perturbed image X' has the same sign (positive for male, negative for female) as its ground-truth label.

Optimizing Attribute Perturbation: Based on the objective function given above, the optimization problem for attribute perturbation can be stated as follows:

$$\min_{\phi} J(X', X) \text{ where } X' = \phi(X). \quad (2.2)$$

For this optimization, we apply small incremental perturbations to an input image as described next.

Algorithm for attribute perturbations The inputs and outputs to this algorithm are as follows:

- **Inputs:** a 2D face image X_0^S , a gallery of face images, a face attribute classifier f , threshold

η

- **Output:** perturbed image X' whose attribute is flipped

The steps of the algorithm are as follows:

- Find landmark points L^S on X_0^S
- Find a candidate in gallery, X^C , that has highest correlation of landmark points, L^C , with those of X_0^S
- Apply Delaunay triangulation to L^S and find the corresponding triangles in L^C
- Initialize $X^S \leftarrow X_0^S$
- Repeat the following steps until the cost function $J(X^S, X_0^S)$ goes below threshold η . For each triangle T in L^S , perform the following steps:
 - Create matrix $Mask_T$ with ones for pixels inside triangle T and zeros everywhere else
 - Estimate the affine transformation matrix A_t : $T^S = A_t T^C$
 - Define $\alpha = (\epsilon^+, \epsilon^-)$
 - Apply perturbations in two directions:

$$X_{T,\alpha}^{S'} = (1 - Mask_T)X^S + Mask_T \left((1 - \alpha)X^S + \alpha A_t X^C \right) \quad \forall \alpha$$

- Calculate the cost function J associated with perturbed images:

$$J(X_{T,\epsilon^+}^{S'}, X_0^S), J(X_{T,\epsilon^-}^{S'}, X_0^S)$$

- Compute numerical gradients for cost function:

$$\nabla_T J = \frac{J(X_{T,\epsilon^+}^{S'}, X_0^S) - J(X_{T,\epsilon^-}^{S'}, X_0^S)}{2 \times \epsilon^+}$$

- Choose the perturbed new image for the next step:

$$X^{S^{new}} = \begin{cases} X_{T,\epsilon^+}^{S'}, & \text{if } \nabla_T J < 0, \\ X_{T,\epsilon^-}^{S'}, & \text{otherwise.} \end{cases}$$

2.2.2 Finding perturbation direction

One way of perturbing an input image is through modifying one pixel at a time and computing the cost function. However, this method is not efficient given the large search space; also, the attribute classification output may not be useful due to low sensitivity after changing only one pixel. Therefore, we use a warping technique to simultaneously modify a group of pixels. The group of pixels to be modified are determined via Delaunay triangulation based on facial landmark points [96]. Also, in order to find the “direction” of perturbing a group of pixels in one triangle, we first select a candidate face image from a gallery set that has the highest correlation of facial landmark points with the input face image. The proposed method for finding the perturbations that would flip the gender attribute of a face image is illustrated in Fig. 2.2.

Given a source face image X_0^S , a set of 77 facial landmark points L^S are extracted using the Stasm software (see Fig. 2.3 for an example). Then, a candidate image X^C that has the highest correlation of landmark points L^C with those of X_0^S is chosen from a gallery set of faces. The correlations are calculated by averaging over the correlations of x and y coordinates of corresponding landmark points. Next, Delaunay triangulation is performed on points in L^S . For each triangle T^S ,

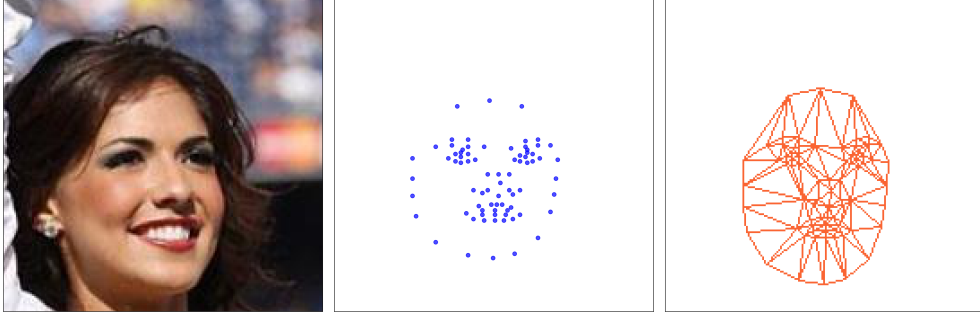


Figure 2.3: Example of Delaunay triangulation on landmark points extracted from an input face image.

the corresponding triangle points T^C are found from the candidate image X^C . The iterative perturbation procedure starts by initializing $X^S \leftarrow X_0^S$. For each triangle T^S and its corresponding T^C , the affine transformation matrix A_t is estimated that maps T^C onto T^S (i.e., $T^S = A_t T^C$). Next, a binary mask $Mask_T$ is defined, that has a value of 1 corresponding to image pixels inside triangle T^S . Finally, the source image is perturbed as,

$$X_{T,\alpha}^{S'} = (1 - Mask_T)X^S + Mask_T \left((1 - \alpha)X^S + \alpha A_t X^C \right), \quad (2.3)$$

where, coefficient α determines perturbing pixels either towards the candidate image (when $\alpha = \epsilon^+ > 0$) or away from the candidate image (when $\alpha = \epsilon^- < 0$), and $|\alpha|$ determines the magnitude of the perturbations. $Mask_T$ ensures that the perturbations are only applied to the triangle T^S , while face pixels outside T^S stay unmodified.

Since we do not have the closed mathematical form of attribute classifier $f(X)$, we use the central-difference to compute the gradient of cost function $\nabla_{T,\alpha} J$ numerically:

$$\nabla_T J(X_T^{S'}, X_0) = \frac{J(X_{T,\epsilon^+}^{S'}, X_0) - J(X_{T,\epsilon^-}^{S'}, X_0)}{2\epsilon^+}. \quad (2.4)$$

Based on the numerical gradient computed above, the perturbation which results in decreasing the cost function J is accepted according to the following rule:

$$X^{S^{new}} = \begin{cases} X_{T,\epsilon^+}^{S'}, & \text{if } \nabla_T J < 0, \\ X_{T,\epsilon^-}^{S'}, & \text{otherwise.} \end{cases} \quad (2.5)$$

The next iteration starts with $X^S = X^{S^{new}}$. This procedure is repeatedly performed for all triangles until the cost function goes below a predefined threshold η .

The entire process is summarized in Algorithm 1.

2.3 Experiments and Results

We used two face datasets for evaluating the proposed method of perturbing face attributes: the MUCT dataset [94] which has 276 subjects – 131 male subjects, 145 female subjects – with 10 or 15 samples per subject and the LFW dataset [68] which has 5740 subjects – 1461 female and 4274 male subjects – and a total of 13227 face samples. In this section, we present the results of our approach to perturb the gender attribute of input images. For this purpose, we designed two experiments as shown in Table 2.1, where perturbations are generated based on two gender classifiers, IntraFace [142] and a Commercial-of-The-Shelf (GCOTS) software. For computing the numerical gradient as mentioned in the previous section, we used $\epsilon^+ = 0.05$ and $\epsilon^- = -0.05$. Our algorithm is run iteratively until the cost function reaches $\eta = -0.1$ or less. A secondary stopping criterion is invoked when the number of iterations exceeds a maximum user set value; in our experiments, this value is set to 40.

Table 2.1: Summary of designed experiments.

Experiments	Perturbations guided by
Exp1	IntraFace
Exp2	GCOTS
Exp3	None (Ref. [107])

Analysis 1: Assessing how gender prediction is affected.

Dataset	Gender classifier	
Original (before)	IntraFace	GCOTS
Exp1 output	IntraFace	GCOTS
Exp2 output	IntraFace	GCOTS
Exp3 output	IntraFace	GCOTS

Analysis 2: Assessing how identity matching is affected through computing genuine/impostor match scores.

Dataset	Match score estimator	
Original (before)	VGG	MCOTS
Exp1 output	VGG	MCOTS
Exp2 output	VGG	MCOTS

2.3.1 Gender perturbation

Two examples from the MUCT dataset are shown in Fig. 2.4 where the gender score is progressively suppressed as assessed by the IntraFace gender classifier. Figure 2.4(a) shows a face image whose gender score is initially negative (*i.e.*, female), which after 31 triangle update steps becomes positive (*i.e.*, male). A similar trend is observed in Fig. 2.4(b), where the initial positive gender score (suggesting a male face) is successfully flipped to a negative value (suggesting a female face).

While some of the previous gender perturbation methods rely on utilizing a candidate face image of the opposite gender [107, 125, 133], our method does not stipulate this condition. Given an input source image, our method was tested using candidates from the same gender as well as candidates from the opposite gender; it was observed that our method successfully works with both types of candidates. However, in some cases, utilizing candidates from opposite gender required fewer iterations, although the difference was not statistically significant.

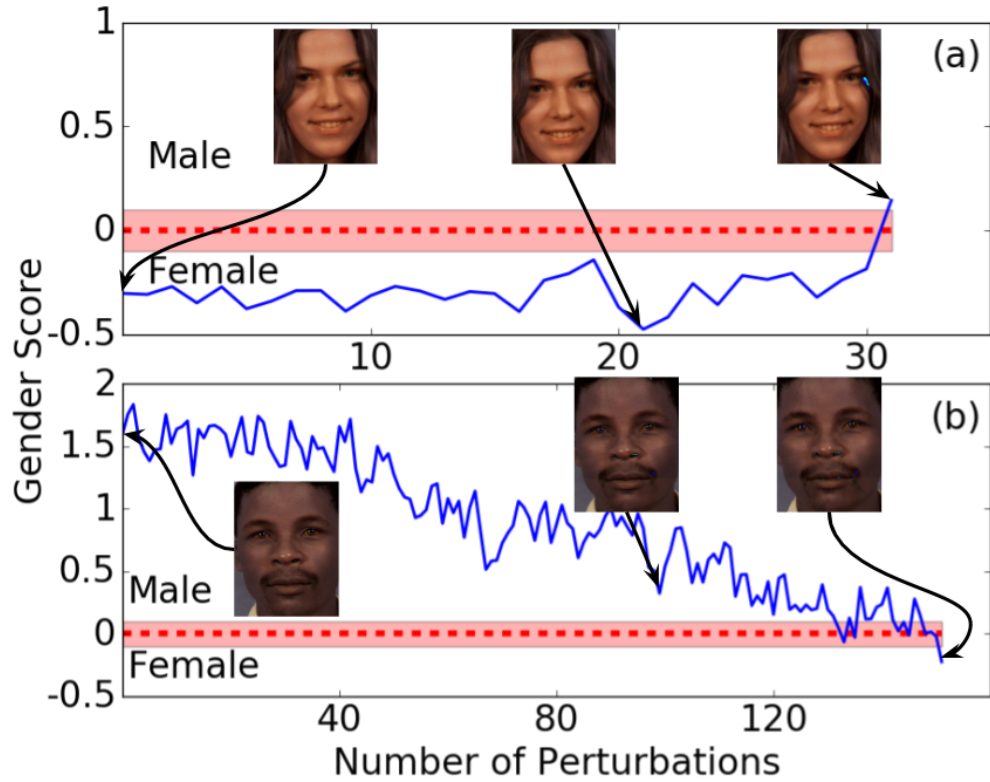


Figure 2.4: Two examples showing the progress of incremental gender perturbation based on IntraFace gender classifier. (a) Input image initially classified as female (gender score=-0.4), gradually perturbed until classified as male (gender score=0.1). (b) Input image initially classified as male (gender score=1.7), gradually perturbed until classified as female (gender score=-0.1).

The histograms of male and female scores of images in the MUCT dataset before and after gender perturbation are displayed in Fig. 2.5. In Fig. 2.5, panels (a) and (b) show the distribution of gender scores of images in the original dataset as computed using IntraFace and GCOTS, respectively. Panels (c) and (d) show the distribution of gender scores on the images after they have been perturbed using the proposed method, and panels (e) and (f) shows the scores for output images generated using the algorithm in [107]. In the original dataset (Fig. 2.5(a)), the distribution of gender score for male subjects is shown in white, and that of female subjects is shown in green. The histogram of gender scores after applying our gender-perturbation method and using the IntraFace gender classifier to guide the process, is shown in Fig. 2.5(c). This analysis indicates that gender scores are flipped, *i.e.*, those which were originally classified as male now have negative scores, and vice versa. Similar results are obtained when GCOTS is used to guide the perturbation process (see Fig. 2.5(b) and (d)). Note that in this case, although the distribution of gender scores for ground-truth males and females are very well separated in the original dataset, yet our method can successfully perturb the gender attribute. The histograms of gender scores computed for output images from [107] are completely overlapped. This is expected according to the K-anonymity [134] principle, since two subjects from opposite genders are mapped to a single mixed face. Quantitatively, the confusion matrices before and after gender perturbation (see Table 2.3) indicate that a 14.9% misclassification rate in the original dataset has increased to 76.6% after gender perturbation based on the proposed method and guided by IntraFace.

While the objectives of our current work are similar to that of [107], there are some important differences. In our work, we intend to flip the gender attribute as assessed by a specific gender classifier, while in [107], two face images from opposite genders are mixed without taking into account any specific gender classifier. Furthermore, in their work, both shape and texture are modified, while in our work, only the texture has changed and the shape of the source face image

Table 2.2: Gender prediction errors (%) computed using IntraFace and GCOTS on the MUCT and LFW datasets.

Dataset		IntraFace	GCOTS
MUCT	Original	14.9	5.1
	Perturbed by IntraFace	76.6	5.4
	Perturbed by GCOTS	24.1	90.1
	Ref. [107]	50.2	51.9
LFW	Original	10.5	2.7
	Perturbed by IntraFace	90.0	2.5
	Perturbed by GCOTS	24.3	68.6
	Ref. [107]	55.5	45.0

Table 2.3: Confusion matrices for gender prediction using IntraFace, on the original MUCT dataset (top) and after perturbations guided by IntraFace (bottom).

		Predictions	
		Male	Female
Ground Truth	Male	1762	17
	Female	521	1300

		Predictions	
		Male	Female
Ground Truth	Male	276	1503
	Female	1255	566

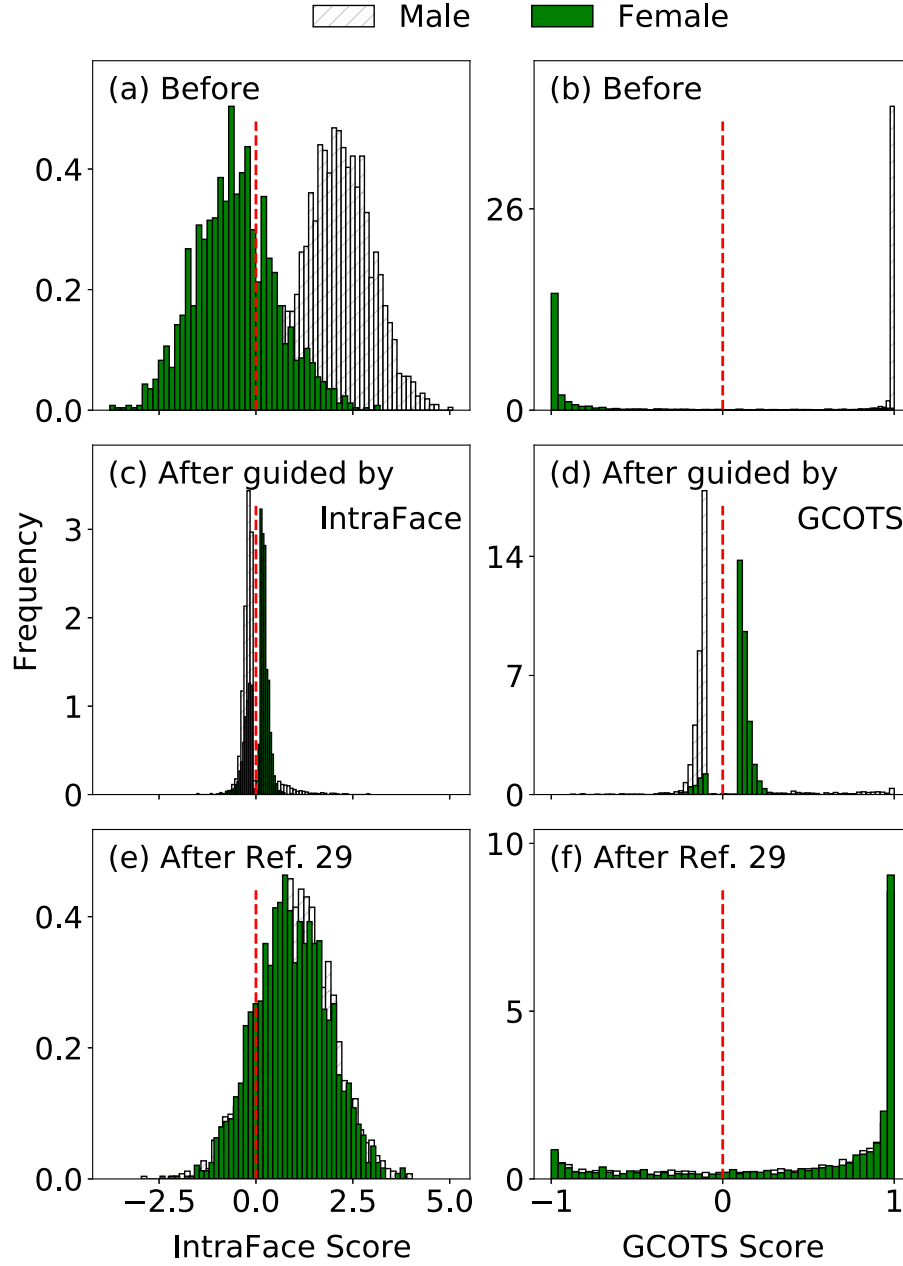


Figure 2.5: Histogram of gender scores obtained by IntraFace [142] ((a),(c), and (e)) and GCOTS ((b),(d), and (f)) on the MUCT dataset. Top row shows the histogram of gender scores in the original data set before perturbations, and middle shows histograms after perturbation. For comparison, the histograms of gender scores for the method proposed by Othman and Ross [107] is shown in (e) and (f). Note that the proposed algorithm is successfully flipping the gender attribute as assessed by both gender classifiers.

stays unchanged.

2.3.2 Match scores

In order to determine if the match scores are affected by the proposed gender perturbation method, we computed genuine and impostor match scores on the original MUCT and LFW datasets (before perturbation), as well as genuine and impostor scores on the perturbed datasets (guided by IntraFace gender classifier and GCOTS gender classifier). Furthermore, we also computed cross-genuine and cross-impostor scores, where face images in the original dataset are compared against those in the perturbed datasets. These experiments are conducted using two face matchers: MCOTS³ and VGG face descriptor [109]. To obtain match scores using the VGG face descriptor, we used the cosine similarity to compare feature descriptors corresponding to a pair of images. Comparing the genuine and impostor histograms for the original dataset and the perturbed datasets (Fig. 2.6) shows that the distributions of genuine and impostor match scores are still well separated.

Furthermore, Receiver Operating Characteristic (ROC) curves for all three cases (before perturbation, after perturbation, and cross-comparison (before/after)) is shown in Fig. 2.7. The ROC curves for all these three cases show little divergence from each other, which provides further evidence that the matching accuracy is not adversely affected by the perturbations.

Two unsuccessful cases are shown in Fig. 2.8, where the gender scores of the original images and perturbed images are both in the positive region thereby indicating the male class. We observed that the average number of perturbations in successful cases was 1084.5 (± 20) for male faces, and 667.2 (± 18) for female faces, when IntraFace is used to guide the perturbation process. These numbers were found to be slightly higher when the GCOTS software is used to guide the

³The face matcher in this case is a state-of-the-art COTS software that demonstrates excellent performance in challenging face datasets.

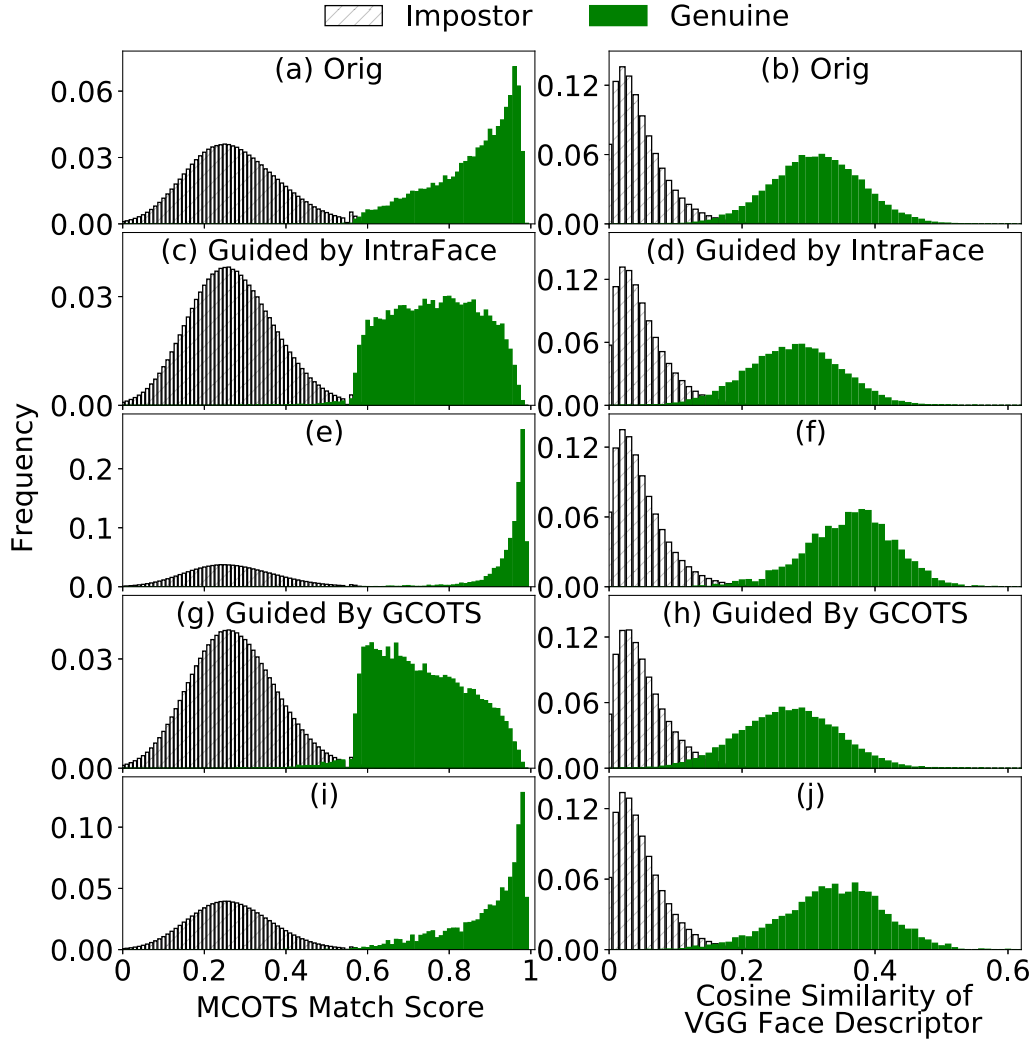


Figure 2.6: Distribution of genuine and impostor match scores obtained via MCOTS software (left column) and using the c VGG face descriptor [109] (right column) on the MUCT dataset; results from original dataset ((a),(b)); after gender perturbations as guided by the IntraFace gender classifier ((c),(d)); cross comparison between original and perturbed images, where perturbation was guided by IntraFace ((e),(f)); after gender perturbations as guided by the GCOTS gender classifier ((g),(h)); and cross comparison between original and perturbed images, where perturbation is guided by GCOTS ((i),(j)).

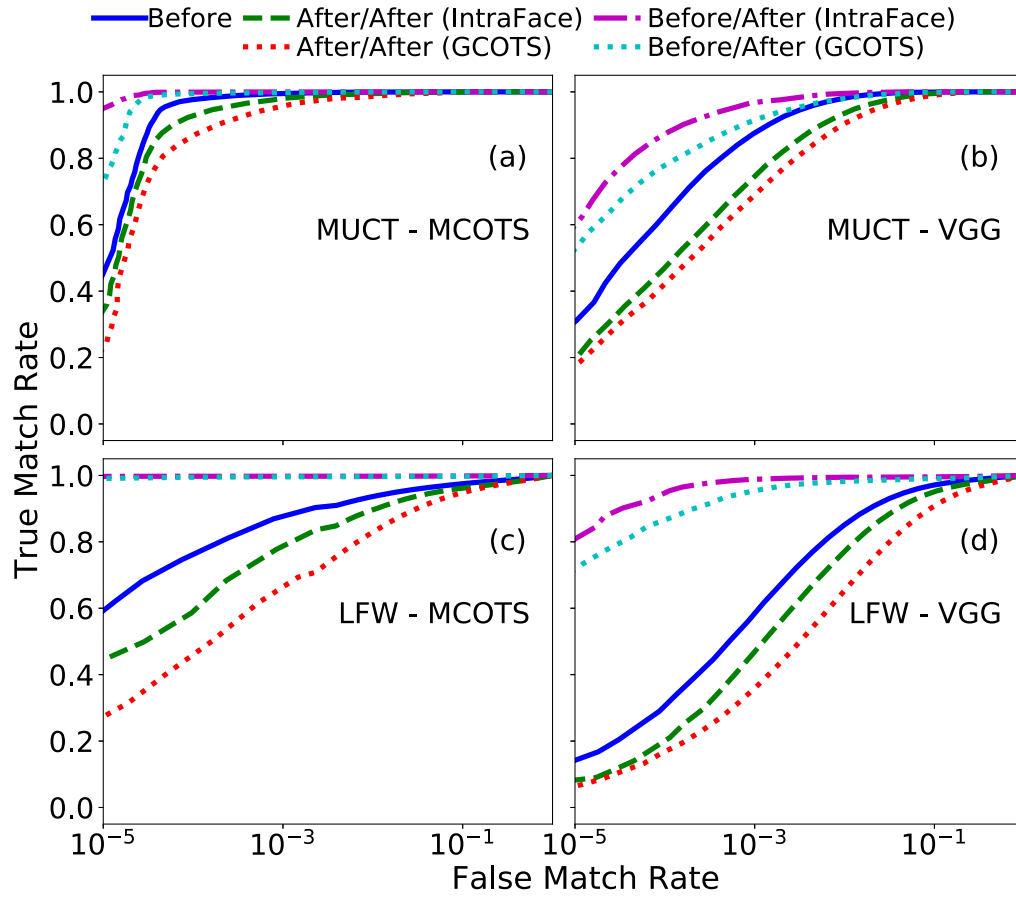


Figure 2.7: ROC curves for face matching obtained using the MCOTS software ((a), (c)) and the VGG face descriptor [109] ((b), (d)). Top row shows the results obtained on the MUCT dataset, and the bottom row on the LFW dataset. Note that the recognition performance is not significantly impacted in most cases.

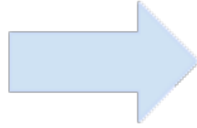
Gender Score: 1.93



Gender Score: 0.10



Gender Score: 3.25



Gender Score: 1.09



Figure 2.8: Two examples of unsuccessful cases where our method fails to completely flip the gender attribute as assessed by IntraFace [142].

perturbation process: $1592 (\pm 32)$ for male faces, and $782 (\pm 22)$ for female faces.

In a practical application, perturbing the gender attribute of *all* stored face images would not be prudent since the output of the attribute classifier can be trivially flipped by the user in order to obtain the true attribute value. In order to avoid this, we can apply the perturbation randomly on a certain proportion of the stored images and leave the rest unchanged. As a result, the certainty of the correct gender label will be reduced.

2.4 Summary and Future Work

While biometric data is solely expected to be used for recognizing an individual, advances in machine learning has made it possible to extract additional information such as age, gender, ethnicity, and health indicators from biometric data. These auxiliary attributes are referred to as demographic attributes. Extracting such attributes from the biometric data of an individual, without their knowledge, has raised several privacy concerns. In this work, we focused on extending privacy to face images. In particular, we designed a technique to modify a face image such that gender information cannot be easily extracted from it, while the image can still be used for biometric recognition purposes. The proposed method entails iteratively perturbing a given face image such that the performance of the face matcher is not adversely affected, but that of the demographic attribute classifier is confounded. The perturbation is accomplished using a gradient descent technique. Experiments involving 2 face matchers and 2 gender classifiers convey the efficacy of the proposed method.

While we showed that the perturbations introduced in this chapter are able to confound the particular gender classifier that was used to derive such perturbations, such perturbations are not transferable to unseen gender classifiers. In the next chapter, we will address this issue by introducing a CNN-based algorithm that can derive perturbations that can be transferable to unseen classifiers.

Chapter 3

Semi Adversarial Networks for Gender

Privacy to Face Images

Portions of this chapter have been published in:

- V. Mirjalili, S. Raschka, A. Namboodiri, A. Ross, "Semi-Adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images", 11th IAPR International Conference on Biometrics (ICB 2018).

3.1 Introduction

In Chapter 1, we investigated the possibility of generating adversarial examples and deriving perturbations to confound a machine learning based gender classifier on face images, while the performance of face matchers were still retained. While the adversarial perturbations derived based on a gender classifier could successfully confound that particular gender classifier, an unseen gender classifier could still correctly predict the gender of the perturbed face images. This shows that the perturbations are not transferable to an unseen classifier.

In this chapter, we develop a convolutional autoencoder (CAE) that generates a perturbed face image that can be successfully used by a *face matcher* but not by an *unseen gender classifier*. The proposed CAE is referred to as a **semi-adversarial network** since its output is adversarial to the

gender classifier but not to the face matcher. The proposed network can be easily appropriated for use with other attributes (such as age or race). In principle, the design of the semi-adversarial network can be utilized in other problem domains where there is a need to confound some classifiers while retaining the utility of other classifiers.

In particular, we provide an alternative solution by designing a convolutional autoencoder that transforms input images such that the performance of an *arbitrary* gender classifier is impacted, while that of an *arbitrary* face matcher is retained. The contributions of this chapter, in this regard, are the following: (a) formulating the privacy-preserving problem in terms of a convolutional autoencoder that does *not* require prior knowledge about the gender classifier nor the face matcher being used; (b) incorporating an explicit term related to the matching accuracy in the objective function which ensures that the *utility* of the perturbed images is not negatively impacted; (c) developing a *generalizable* solution that can be trained on one dataset and applied to other previously unseen datasets.

To the best of our knowledge, this is the first work where adversarial training is used to design a generator component that is able to maximize the performance with respect to one classifier while minimizing the performance with respect to another. Experimental results show that the proposed method of semi-adversarial learning for multi-objective functions is efficient for deriving perturbations that are generalizable to other classifiers that were not used (or not available) during training.

3.2 Proposed Method

3.2.1 Problem formulation

Let $X \in \mathbb{R}^{m \times n \times c}$ denote a face image having c channels each of height m and width n . Let $f_G(X)$ denote a binary gender classifier that returns a value in the range $[0, 1]$, where 1 indicates a “Male” and 0 indicates a “Female”. Let $f_M(X_1, X_2)$ denote a face matcher that computes the match score between a pair of face images, X_1 and X_2 . The goal of this work is to construct a model $\phi(X)$, that perturbs an input image X such that the perturbed image $X' = \phi(X)$ has the following characteristics: (a) from a human perspective, the perturbed image X' must look similar to the original input X ; (b) the perturbed image X' is most likely to be misclassified by an arbitrary gender classifier $f_G(X)$; (c) the match scores, as assessed by an arbitrary biometric matcher f_M , between perturbed image X' and other unperturbed face images from the same subject, are not impacted thereby retaining verification accuracy.

This goal can be expressed as the following objective function, which minimizes a loss function J consisting of three disjoint terms corresponding to the three characteristics listed above:

$$\begin{aligned} J(X, y, X'; f_G, f_M) = \\ \lambda_D J_D(X, X') + \lambda_G J_G(y, X'; f_G) + \lambda_M J_M(X, X'; f_M), \end{aligned} \tag{3.1}$$

where, X is the input image, y is the gender label of X , and X' is the perturbed image. The term $J_D(X, X')$ measures the dissimilarity between the input image and the perturbed image produced by a decoder $\phi(X)$ to ensure that the perturbed images still appear as realistic face images. The second term, $J_G(y, X'; f_G)$, measures the loss associated with correctly predicting gender of perturbed image X' using f_G , to ensure that the accuracy of the gender classifier on the perturbed image X' is reduced. The third term, $J_M(X, X'; f_M)$, measures the loss associated with the match

score between X and X' computed by f_M . This term ensures that the matching accuracy as assessed by f_M is not substantially diminished due to the perturbations introduced to confound the gender classifier.

In order to optimize this objective function, *i.e.*, minimizing gender classifier accuracy while maximizing the biometric matching accuracy and generating realistic looking images, we design a novel convolutional neural network architecture that we refer to as a semi-adversarial convolutional autoencoder.

3.2.2 Semi-adversarial network architecture

The semi-adversarial network introduced in this chapter is significantly different from Generative Adversarial Networks (GANs). A typical GAN has two components: a discriminator and a generator. The *generator* learns to generate realistic looking images from the training data, while the *discriminator* learns to distinguish between the generated images and the corresponding training data [53, 126]. In contrast to regular GANs consisting of a generator and a single discriminator, the proposed semi-adversarial network attaches two independent classifiers to a generative subnetwork. Unlike the generator subnetwork of GANs that is trained based on the feedback of one classifier, the semi-adversarial configuration proposed in this chapter learns to generate image perturbations based on the feedback of two classifiers, where one classifier acts as an adversary of the other. Hence, the semi-adversarial network architecture we propose consists of the following three different subnetworks (Fig. 3.1): (a) a trainable generative component in form of a convolutional autoencoder (subnetwork I) for adversarial learning; (b) an auxiliary CNN-based gender classifier (subnetwork II); (c) an auxiliary CNN-based face matcher (subnetwork III).

The auxiliary gender classifier as well as the auxiliary matcher¹ are detachable parts in this

¹The term “auxiliary” is used to indicate that these subnetworks do not correspond to pre-trained gender classifiers

network architecture used only during the *training* phase. In contrast to GANs, the generative component of this proposed network architecture is a convolutional autoencoder (section 3.2.2), which is initially pre-trained to produce an image that closely resembles an image from the training set after incorporating gender prototype information (section 3.2.2). Then, during further training, feedback from both an auxiliary CNN-based gender classifier and an auxiliary CNN-based face matcher are incorporated into the loss function (see Eqn. (3.1)) to perturb the regenerated images such that the error rate of the auxiliary gender classifier increases while that of the auxiliary face matcher is not unduly affected.

An overview of this semi-adversarial architecture is shown in Fig. 3.1, and the details are further described in the following subsections.

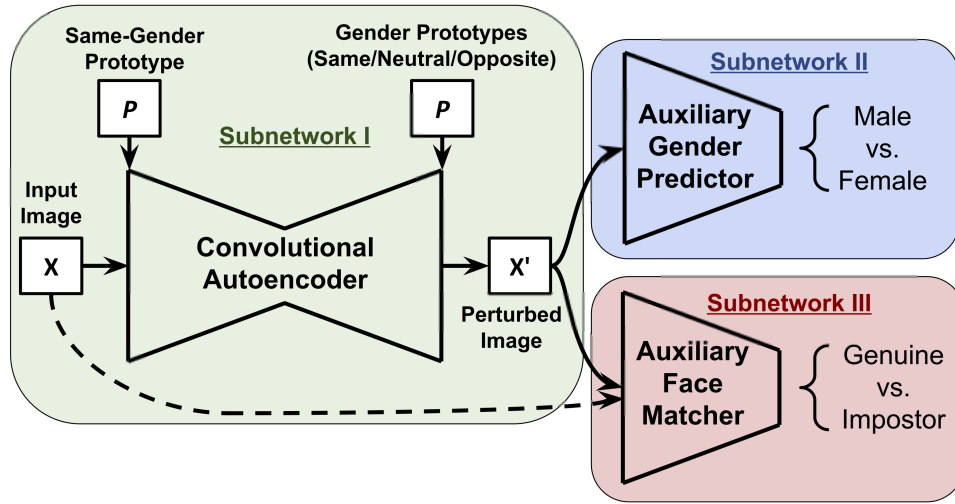


Figure 3.1: Schematic representation of the semi-adversarial neural network architecture designed to derive perturbations that are able to confound gender classifiers while still allowing biometric matchers to perform well. The overall network consists of three sub-components: a convolutional autoencoder (subnetwork I), an auxiliary gender classifier (subnetwork II), and an auxiliary matcher (subnetwork III).

Convolutional autoencoder

The architecture of the convolutional autoencoder sub-network or face matchers, but rather classifiers that are generated from the training data. Note that such a formulation makes the semi-adversarial network generalizable.

that modifies and reconstructs the input image in three different ways is shown in Fig. 3.2. The input to this sub-network is a gray-scale face image of size 224×224 concatenated with a same-gender prototype, P_{SM} (Fig. 3.3). The input is then processed through the encoder part consisting of two convolutional layers; each layer is followed by a leaky ReLU activation function and an average pooling layer, resulting in feature maps of size $56 \times 56 \times 12$. Next, the outputs of the encoder are passed through a decoder with two convolutional layers each, followed by a leaky ReLU activation and an upsampling layer using two-dimensional nearest neighbor interpolation. The output of the decoder is a $224 \times 224 \times 128$ dimensional feature map.

The feature maps from the decoder output are then concatenated with either same-gender (P_{SM}), neutral-gender (P_{NT}), or opposite-gender (P_{OP}) prototypes in the *proto-combiner* module (see Fig. 3.2 and Fig. 3.3). The proto-combiner module is followed by a final convolutional layer and a sigmoid activation function yielding a reconstructed image X'_{SM} , X'_{NT} , or X'_{OP} , depending on the gender-prototype used. The autoencoder described in this section contains five trainable layers. Those layers are pre-trained using an information bottleneck approach [65] to retain the relevant information from both the original image and the same-gender prototype. This is sufficient to reconstruct realistic looking images by minimizing $J_D(X, X')$, which measures the dissimilarity between the gray-scale input images and the perturbed images by computing the sum of the element-wise cross entropy between input and output (perturbed) images. After pre-training, this subnetwork is further trained by passing its reconstructed images to two other sub-networks: the auxiliary gender predictor and the auxiliary face matcher (Fig. 3.1). The gender prototypes, as well as the two subnetworks, are described in the following subsections.

Gender prototypes The 224×224 male and female RGB gender prototypes (P_{male} , P_{female}) were computed as the average of all 65,160 male images and 92,190 female images, respectively, in the CelebA training set [86]. Then, the same-gender (P_{SM}) and opposite-gender (P_{OP}) pro-

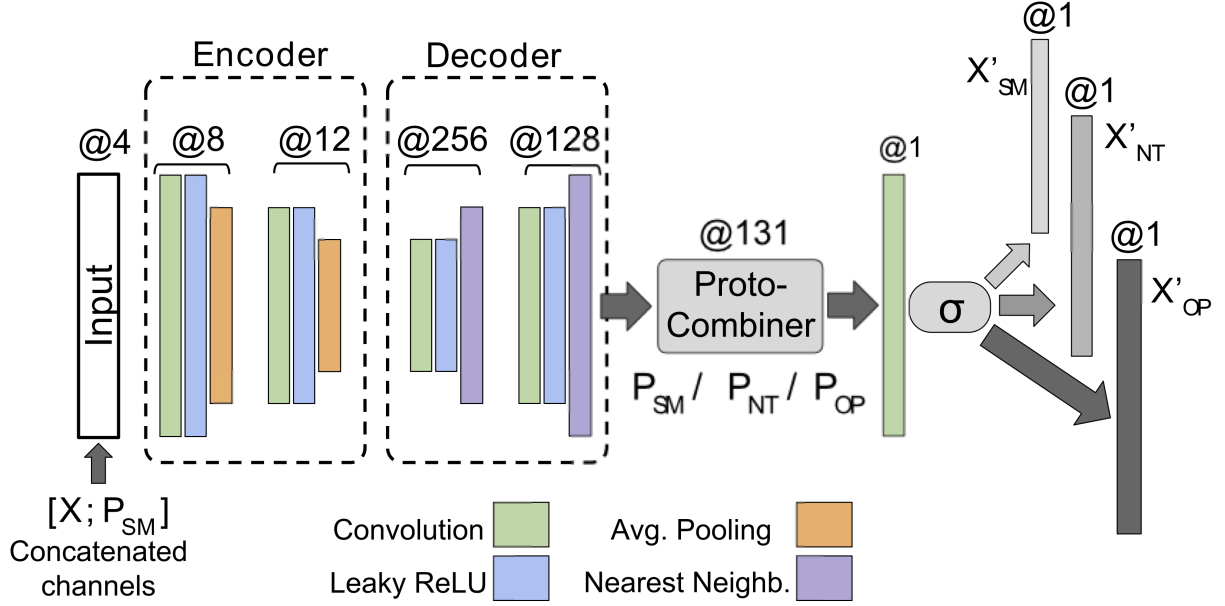


Figure 3.2: Architecture of the autoencoder augmented with gender-prototype images. The encoder receives a one-channel gray-scale image as input, which is concatenated with the RGB channels of the same-gender prototype image. After the compressed representation is passed through the decoder part of the autoencoder for reconstruction (128 channels), the proto-combiner concatenates it with the RGB channels of a same-, neutral-, or opposite-gender prototype resulting in 131 channels that are then passed to a final convolutional layer.

totypes, which are being concatenated with the input image and combined with the autoencoder output (Fig. 3.2), are constructed based on the ground-truth label y , while the neutral-gender prototype is computed as the weighted mean of male and female prototypes (Fig. 3.3):

- Same-gender prototype, P_{SM} : $yP_{\text{male}} + (1 - y)P_{\text{female}}$
- Opposite-gender prototype, P_{OP} : $(1 - y)P_{\text{male}} + yP_{\text{female}}$
- Neutral prototype, P_{NT} : $\alpha_F P_{\text{female}} + \alpha_M P_{\text{male}}$

Here, α_F is the proportion of females in the CelebA training set and α_M is the proportion of males. The convolutional autoencoder network (summarized in Fig. 3.1 and further illustrated in Fig. 3.2) is provided with same-gender prototype images (female or male corresponding to

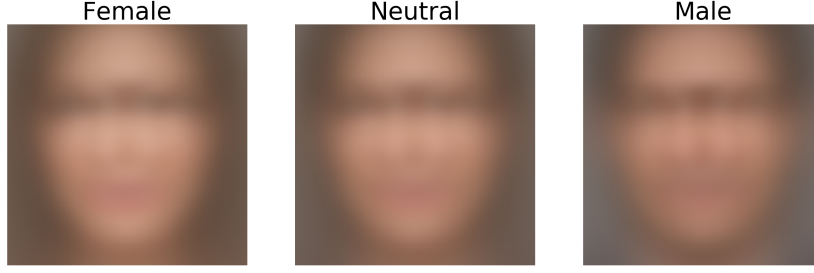


Figure 3.3: Gender prototypes used to confound gender classifiers while maintaining biometric matching during the semi-adversarial training of the convolutional autoencoder.

the ground truth label of the input image), which are concatenated with the input image before being transmitted to the encoder module in order to derive a compressed representation of the original image along with the same-gender prototype information. After the decoder reconstructs the original images, the three different gender-prototypes are added as additional channels via the proto-combiner (Fig. 3.2).

The final convolutional layer of the autoencoder produces three different perturbed images: X'_{SM} (obtained when the same-gender prototype is used), X'_{NT} (when the neutral prototype is used), and X'_{OP} (when the opposite-gender prototype is used).

Pre-training: During pre-training, to ensure that the convolutional autoencoder is capable of reconstructing the original images, only the same gender perturbations (X'_{SM}) were considered in the cross-entropy cost function.

Training: For the further training of the autoencoder, to confound the auxiliary gender classifier and ensure high matching accuracy of the auxiliary matcher, both the perturbed outputs using same- and opposite-gender prototypes were passed through the auxiliary gender classifier, to ensure that the perturbation made using the same-gender prototype produces accurate gender prediction while perturbations made using the opposite-gender prototype confounds the gender prediction. The perturbed outputs due to the neutral prototypes are not incorporated in the loss

function, and are only used for evaluation purposes.

3.2.3 Auxiliary CNN-based gender classifier

The architecture of the auxiliary CNN-based gender classifier, which consists of six convolutional layers and two fully connected (FC) layers, is summarized in Fig. 3.4. Each convolutional layer is followed by a leaky ReLU activation function and a max-pooling layer that reduces the height and width dimensions by a factor of 2, resulting in feature maps of size $4 \times 4 \times 256$. Passing the output of the second FC layer through a sigmoid function results in class-membership probabilities for the two labels: 0:“Female” and 1:“Male”. This network was independently trained on the CelebA-train dataset by minimizing the cross-entropy cost function, until its convergence after five epochs; the gender prediction accuracy of the auxiliary network when tested on the CelebA-test set was 96.14%. During training, two dropout layers with drop probability of 0.5 were added to the FC layers for regularization. However, these dropout layers were removed when this subnetwork was used for deriving perturbations as part of the three-subnetwork neural network architecture shown in Fig 3.1.

As this CNN-based gender classifier was only used for training the convolutional autoencoder for generating perturbed face images, and not for further evaluation of this model, it is referred to as *auxiliary gender classifier* to distinguish it from the gender classifiers used for evaluation.

3.2.4 Auxiliary CNN-based face matcher

As discussed in Section 3.2.1, the loss function contains a term $J_M(X, X'; f_M)$ to ensure good face matching accuracy despite the perturbations introduced to confound the gender classifier. To provide match scores during the training of the autoencoder subnetwork, we used a publicly

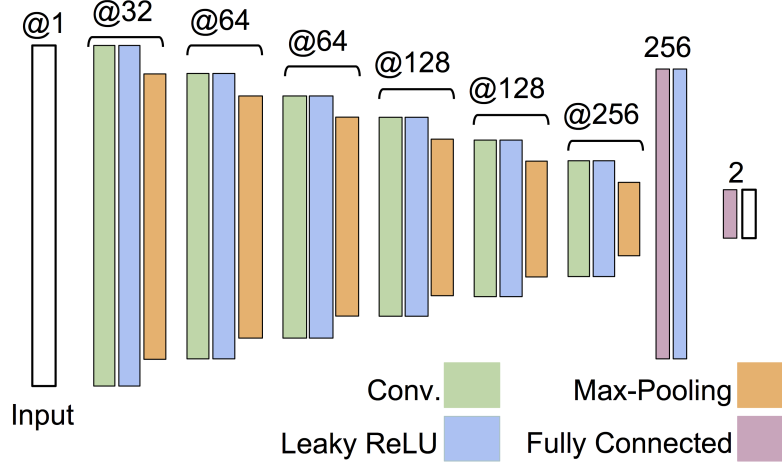


Figure 3.4: Architecture of the CNN-based auxiliary gender classifier that was used during the training of the convolutional autoencoder. This classifier was used as an auxiliary (fixed) component in the final model to derive the image perturbations according to the objective function described in Section 3.2.1.

available VGG model as described by Parkhi *et al.* [109] consisting of 16 weight layers. This VGG subnetwork produces face descriptors which are vector representations of size 2622 extracted from RGB face images. The publicly available weight parameters of this network were used without further performance tuning.

In addition, as the open-source VGG-face network expects RGB images as inputs, we modified the convolutional filters of the first layer by adding the three filter matrices related to the input channels, for compatibility with the single-channel gray-scale input images. As this CNN-based face matcher was only used for training the convolutional autoencoder for generating perturbed face images, and not for further evaluation of this model, it is referred to as *auxiliary matcher* to distinguish it from the commercial matching software used for evaluation.

3.2.5 Loss function

After pre-training the convolutional autoencoder described in Section 3.2.2, it is connected to the other two subnetworks (the auxiliary CNN-based gender classifier described in Section 3.2.3 and the auxiliary CNN-based face matcher described in Section 3.2.4) for further training. During the pre-training stage, the loss term $J_D(X, X')$ was used to ensure that the convolutional autoencoder is capable of producing images that are similar to the input images. The loss term is computed as the element- or pixel-wise cross entropy, S , between input and output (perturbed) images:

$$J_D(X, X'_{SM}) = \sum_{k=1}^{224^2} S\left(X^{(k)}, X'^{(k)}_{SM}\right). \quad (3.2)$$

Next, to generate the perturbed images X'_{SM} , X'_{NT} , or X'_{OP} (based on the type of gender-prototype used) such that gender classification is confounded but biometric matching remains accurate, two loss terms, J_G and J_M , were used. The first loss term is associated with suppressing gender information in X'_{OP} and preserving it in X'_{SM} :

$$J_G(y, X'_{SM}, X'_{OP}; f_G) = S(y, f_G(X'_{SM})) + S(1 - y, f_G(X'_{OP})), \quad (3.3)$$

where, $S(t, \hat{p})$ denotes the cross-entropy cost function using target label t and the predicted class-membership probability \hat{p} . Note that in this loss function, we use the ground truth labels for X'_{SM} so that the gender of X'_{SM} is correctly predicted, while we use flipped labels for X'_{OP} so that the gender of perturbed image X'_{OP} is incorrectly predicted. We found that without the use of this configuration for X'_{SM} and X'_{OP} , the network will perturb the input image, X , such that perturbations are overfit to the auxiliary gender classifier that is used during training.

The second loss term, J_M , measures the matching similarity between input image X and the

perturbed image X'_{SM} generated from the same-gender prototype:

$$J_M(X, X'_{SM}; R_{vgg}) = \|R_{vgg}(X'_{SM}) - R_{vgg}(X)\|_2^2, \quad (3.4)$$

where, $R_{vgg}(X)$ indicates the vector representation of image X obtained from the VGG-face network [109]. The total loss is then the weighted sum of the two loss terms J_G and J_M :

$$\begin{aligned} J_{total}(X, y, X'_{SM}, X'_{OP}; f_G, R_{vgg}) = \\ \lambda_G J_G(y, X'_{SM}, X'_{OP}; f_G) + \lambda_M J_M(X, X'_{SM}; R_{vgg}). \end{aligned} \quad (3.5)$$

J_{total} was then used to derive the loss gradients with respect to the parameter weights of the convolutional autoencoder during the training stage, to generate perturbations according to the objective function (Section 3.2.1). Note that the coefficients λ_M and λ_G in Eqn 3.5 constitute additional tuning parameters to re-weight the contributions of J_G and J_M toward the total loss. In this work, we did not optimize λ_M and λ_G , however, and used a constant of 1 to weight both J_G and J_M equally.

3.2.6 Datasets

The original dataset source used in this work is the large-scale CelebFaces Attributes (CelebA) dataset [86], which consists of 202,599 face images in JPEG format for which gender attribute labels were already available with the dataset. The dataset was randomly divided into 162,079 training images (CelebA-train) and 40,520 images for testing (CelebA-test). The CelebA-train dataset was used to train the gender classifier (Section 3.2.3), as well as the convolutional autoencoder (Section 3.2.2).

In addition to the CelebA-test dataset, three publicly available datasets were used for evalua-

Table 3.1: Sizes of the datasets used in this study for training and evaluation. CelebA-train was used for training only, while the other four datasets were used to evaluate the final performance of the trained model.

Dataset	Train	# Images	# Male	# Female
CelebA-train	yes	157,350	65,160	92,190
CelebA-test	no	39,411	16,318	23,093
MUCT	no	3754	131	145
LFW	no	12,969	4205	1448
AR-face	no	3286	76	60

tion only: MUCT [94], LFW [68] and AR-face [90] databases. The final compositions of these datasets, after applying a preprocessing step using a deformable part model (DPM) as described by Felzenszwalb *et al.* [44] to ensure that all images have the same dimensions (224×224), are summarized in Table 3.1. The resulting perturbed images obtained from the CelebA-test, MUCT, LFW, and AR-face datasets, were used to measure the effectiveness of modifying the gender attribute as assessed by a commercial gender classifier (G-COTS) and a commercial biometric matcher (M-COTS, excluding AR-images labeled as occluded due to sunglasses or scarfs).

3.2.7 Implementation details and software

The convolutional autoencoder (Section 3.2.2), auxiliary CNN-based gender classifier (Section 3.2.3) and the auxiliary CNN-based face matcher (Section 3.2.4) were implemented in TensorFlow [2] based on custom code for the convolutional layers and freezing the parameters of the gender classifier and face matcher during training of the autoencoder subnetwork [117].

3.3 Experiments and Results

After training the autoencoder network using the CelebA-train dataset as described in Section 3.2.2, the model was used to perturb images in other, independent datasets: CelebA-test, MUCT,

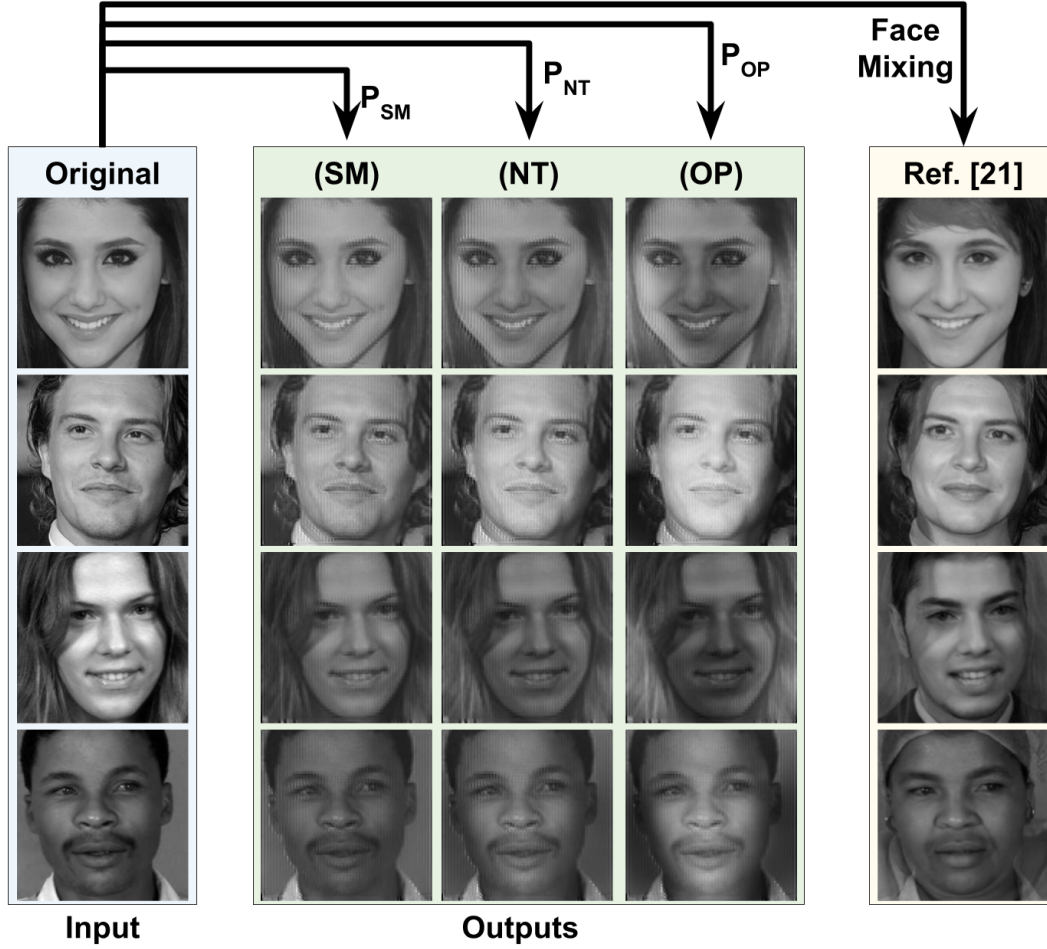


Figure 3.5: Example input images with their reconstructions using same, neutral, and opposite gender prototypes from the CelebA-test (first two rows) and MUCT (last two rows) datasets.

LFW, and the AR-face database. For each face image in these datasets, a set of three output images was reconstructed using same-gender, neutral-gender, and opposite-gender prototypes. Furthermore, our results are compared with the face-mixing approach proposed in [107]. Examples of these reconstructed outputs for two female face images, and two male face images are shown in Fig. 3.5.

3.3.1 Evaluation and verification

The previously described auxiliary CNN-based gender classifier (Section 3.2.3) and auxiliary CNN-based face matcher (Section 3.2.4) were not used for the evaluation of the proposed semi-adversarial autoencoder as these two subnetworks were used to provide semi-adversarial feedback during training. The performance of the semi-adversarial autoencoder is expected to be optimally biased when tested using the auxiliary gender classifier and auxiliary face matcher. Thus, we used independent gender classification and face matching software for evaluation and verification instead, to represent a real-world use case scenario.

Two sets of experiments were conducted to assess the effectiveness of the proposed method. First, two independent software for gender classification were considered: the popular research software IntraFace [142] as well as a state-of-the-art commercial software, which we refer to as *G-COTS*. Second, a state-of-the-art commercial matcher that has shown excellent recognition performance on challenging face datasets was used to evaluate the face matching performance; we refer to this commercial face matching software as *M-COTS*.

3.3.2 Perturbing gender

In order to assess the effectiveness of the proposed scheme in perturbing gender, the reconstructed images using the proposed semi-adversarial autoencoder from the four datasets were analyzed. The Receiver Operating Characteristic (ROC) curves for predicting gender using IntraFace and G-COTS from the original images and the perturbed images are shown in Fig. 3.6.

We note that gender prediction via IntraFace is heavily impacted when using different gender prototypes for image reconstruction. We observe that the performance of IntraFace on AR-face images after opposite-gender perturbation is very close to random (as indicated by the near-diagonal

Table 3.2: Error rates in gender prediction using IntraFace and G-COTS gender classification softwares on the original datasets before and after perturbation. Note the substantial increase in the prediction error upon perturbation via the convolutional autoencoder model using opposite-gender prototypes.

Software	Dataset	Original (before)	Perturbed (after OP)	Ref. [107]
IntraFace	CelebA-test	19.7%	39.3%	44.6%
	MUCT	8.0%	39.2%	57.7%
	LFW	33.4%	72.5%	70.9%
	AR-face	16.9%	53.8%	54.2%
G-COTS	CelebA-test	2.2%	13.6%	42.4%
	MUCT	5.1%	25.4%	53.9%
	LFW	2.8%	18.8%	46.1%
	AR-face	9.3%	26.9%	40.6%

ROC curve in Fig. 3.6(a)-(d)). The performance of G-COTS proves to be more robust towards perturbations, compared to IntraFace; however, the ROC curve corresponding to the opposite-gender prototype, shows a substantial deviation from the ROC curve of the original images (Fig. 3.6(e)-(h)). This observation indicates that the opposite-gender prototype perturbations have a substantial, negative impact on the performance of state-of-the-art G-COTS software, thereby extending gender privacy.

The exact error rates in predicting the gender attribute of face images using both IntraFace and G-COTS software are provided in Table 3.2 for the original images and the perturbed images using opposite-gender prototypes. The quantitative comparison of the error rates indicates a substantial increase in the prediction error rates when image datasets were perturbed using opposite-gender prototypes. Note that in the case of G-COTS software, perturbations made by the face mixing scheme proposed in [107] result in higher error rates. On the other hand, the additional advantage of our approach is in preserving the identity, as we will see in the next section.

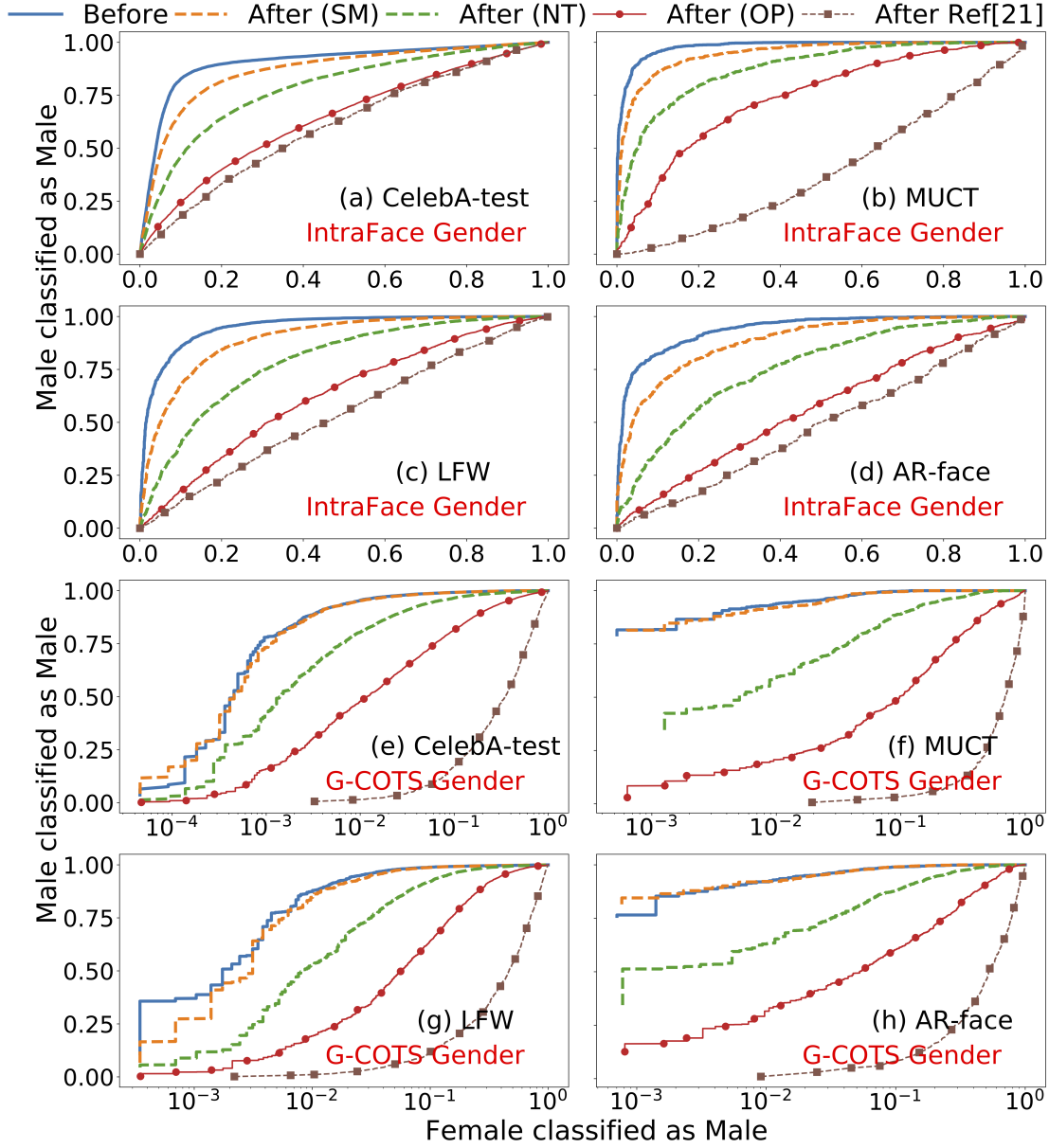


Figure 3.6: ROC curves comparing the performance of IntraFace (a-d) and G-COTS (e-h) gender classification software on original images (“Before”) as well as images perturbed via the convolutional autoencoder model (“After”) on four different datasets: CelebA-test, MUCT, LFW, and AR-face.

3.3.3 Retaining matching accuracy

The match scores were computed using a state-of-the-art M-COTS software and the resulting ROC curves are shown in Fig. 3.7. While the matching term, J_M , in the loss function is directly applied

to reconstructed outputs from same-gender prototype, X'_{SM} , the reconstructions that use neutral- or opposite-gender prototypes are not directly subject to this loss term (see Section 3.2.5). As a result, the ROC curve of the reconstructed images coming from same-gender prototype appear much closer to the original input compared to the reconstructed images from neutral- and opposite-gender prototypes. Overall, we were able to retain a good matching performance even when using opposite-gender prototype. On the other hand, the ROC curves obtained from outputs of the mixing approach proposed in [107] are heavily impacted, resulting in de-identified outputs (which is not desirable in this work).

Finally, the True Match Rate (TMR) values at a False Match Rate of 1% are reported in Table 3.3. The perturbed images from all three datasets show TMR values that are very close to the value obtained from the unperturbed original dataset.

Table 3.3: True (TMR) and false (FMR) matching rates (measured at values of 1%) of the independent, commercial M-COTS matcher after perturbing face images via the convolution autoencoder using same (SM), neutral (NT), and opposite (OP) gender prototypes, indicating that the biometric matching accuracy is not substantially affected by confounding gender predictions.

Dataset	Original (before)	Perturbed		
		(SM)	(NT)	(OP)
MUCT	99.88 %	99.79%	99.57%	98.44%
LFW	90.29%	90.02%	88.47%	83.45%
AR-face	94.97%	94.11%	91.95%	90.81%

3.4 Summary and Future Work

In this work, we focused on developing a semi-adversarial network for imparting demographic privacy to face images. In particular, our semi-adversarial network perturbs an input face image such that gender prediction is confounded while the biometric matching utility is retained. The proposed method uses an auxiliary CNN-based gender classifier and an auxiliary CNN-based face

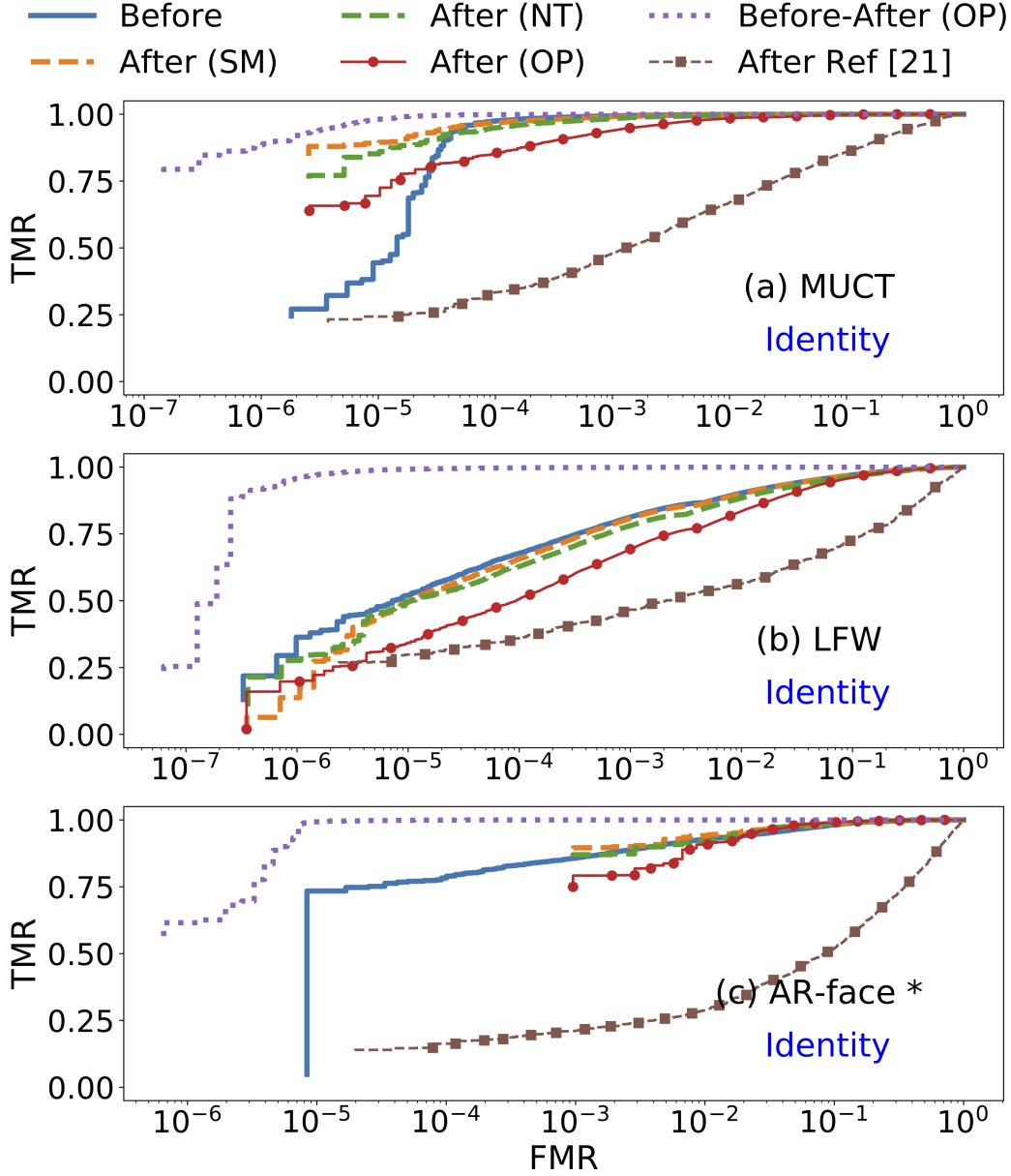


Figure 3.7: ROC curves showing the performance (true and false matching rates) of M-COTS biometric matching software on the original images (“Before”) compared to the perturbed images (“After”) generated by the convolutional autoencoder model using same-, neutral-, or opposite-gender prototypes for three different datasets: (a) MUCT, (b) LFW, and (c) AR-face.

matcher for training the convolutional autoencoder. The trained model is evaluated using two independent gender classifiers and a state-of-the-art commercial face matcher which were unseen during training. Experiments confirm the efficacy of the proposed architecture in imparting gender

privacy to face images, while not unduly impacting the face matching accuracy.

While the proposed SAN model was shown to be successful in confounding unseen gender classifier, we need to find the vulnerable points of the SAN model. One such vulnerabilities is the issue of generalizability, which means whether the perturbed face images are able to confound an arbitrary unseen gender classifier or not. In the next chapter, we provide a solution using an ensemble model to address this issue.

Chapter 4

On the Generalization Ability of Gender Privacy using Semi-Adversarial Networks

Portions of this chapter have been published in:

- V. Mirjalili, S. Raschka, A. Ross, "Gender privacy: An ensemble of semi adversarial networks for confounding arbitrary gender classifiers", 9th International Conference on Biometrics Theory, Applications and Systems (BTAS 2018).

4.1 Introduction

In chapter 2, we investigated the possibility of utilizing adversarial images for imparting gender privacy. The researchers were able to generate image perturbations targeting a specific gender classifier and showed that these perturbations could confound the gender classifier, while preserving the performance of a commercial face-matcher. Although perturbed adversarial images have shown to be effective in confounding a *particular* classifier, the issue that these images may not adversarially affect other *unseen* classifiers limits their effectiveness in practical privacy applications.¹ Adversarial images generated for a particular gender classifier may not generalize to another. Furthermore, in a real-world application, the knowledge of a gender classifier may not

¹The term "unseen" indicates that the classifier or matcher was not used during the training stage.

be available in advance; as a result, generating adversarial images for an unseen gender classifier would be difficult. To address this issue, in Chapter 3, developed an autoencoder called Semi Adversarial Network (SAN) for generating perturbed face images that could potentially generalize across unseen gender classifiers. They trained the SAN model using an auxiliary gender classifier and an auxiliary face matcher and evaluated the success of their model in producing output images that could confound two unseen gender classifiers, while preserving the recognition accuracy of an unseen face matcher. Although the accuracy of the two unseen gender classifiers were indeed confounded, yet, generalizability to a large number of unseen gender classifiers remains an open problem (see Section 4.3). Furthermore, a human observer may be able to correctly classify the gender of the perturbed images generated by their model (see Figure 4.6), which means that, in principle, there exists an unseen gender classifier that can correctly recognize the gender of the perturbed images. In this chapter, we formulate an *ensemble* technique to address the limitations of the previous SAN model and facilitate its generalizability to a large number of unseen gender classifiers.² In the context of this work, the *generalizability* of a SAN model is defined as its ability to perturb face images in such a way that an *arbitrary* unseen gender classifier is confounded while an arbitrary unseen face matcher retains its utility.

The major contributions of this chapter are as follows:

- Designing an ensemble of SANs to address the problem of generalizability across unseen gender classifiers.
- Conducting large-scale experiments that convey the practicality and efficacy of the proposed approach.

²The acronym SAN was simultaneously coined by two independent research groups. Cao *et al.* [25] defined Selective Adversarial Networks for partial style transfer, and Mirjalili *et al.* [101] defined Semi Adversarial Networks for imparting privacy to face images. Here, we use SAN to refer to the latter.

- Ensuring that race and age attributes are retained in the perturbed face images.

4.2 Proposed Method

Previous SAN model [101]: The overall architecture of the individual SAN models in the ensemble is similar to the SAN model proposed in [101] as shown in Figure 4.2, but with a few modifications. The SAN model consists of a **convolutional autoencoder** to perturb face images, a convolutional neural network (CNN) as an **auxiliary face matcher**, and a CNN as an **auxiliary gender classifier**. The pre-trained, publicly available VGG-face CNN [109] is used as the auxiliary face matcher. The input gray-scale image is first fused with a **face prototype** belonging to the same gender as the input image. Then 128 feature maps are obtained from the last layer of the decoder, which are combined with the face prototype of the opposite gender using 1×1 convolutions. The final image is then passed to both the auxiliary face matcher and the auxiliary gender classifier to compute its match score with the original input and its gender probability, respectively. During training, each input image is reconstructed by the autoencoder using both same-gender and opposite-gender prototypes to obtain two different outputs. Then, three different cost functions are used based on these outputs. First, a pixel-wise similarity measure between the input and the output from the same-gender prototype is used as a cost function to ensure that the autoencoder is able to construct realistic images. The second cost function is the L^2 distance between the face vector of the input image and those of the outputs to make the autoencoder learn to perturb face images such that the accuracy of the face matcher is retained. The third cost term is the cross-entropy loss applied to the gender probabilities of the two outputs as computed by the auxiliary gender classifier, where the ground-truth label of the input image is used for the output of the same-gender prototype but the reverse is used for the output of the opposite-gender prototype.

4.2.1 Ensemble SAN Formulation

We assume that there exists a large set of gender classifiers $\mathcal{G} = \{G^1, G^2, \dots, G^n\}$, where each $G^i(X)$ predicts the gender of a person based on a 2D face image, X . Furthermore, we assume a set of face-matchers denoted by $\mathcal{M} = \{M^1, M^2, \dots, M^m\}$, where each $M^i(X_a, X_b)$ computes the match score between a pair of face images, X_a and X_b . The goal of the work is to design an *ensemble* of t SAN models, $\mathcal{S} = \{S^1, S^2, \dots, S^t\}$, that can be shown to generalize to arbitrary gender classifiers. In particular, we demonstrate that for each face image X , \mathcal{S} produces a set of outputs $\mathcal{S}(X) = \{Y^1, Y^2, \dots, Y^t\}$ such that for each $G^i \in \mathcal{G}$, there exists at least one output $Y^j = S^j(X)$ that is able to confound G^i . At the same time, the outputs, $\mathcal{S}(X)$, can be successfully used for face recognition by the matchers in \mathcal{M} .

4.2.2 Diversity in Autoencoder Ensembles

One of the key aspects of neural network ensembles is diversity among the individual network models [63]. Several techniques have been proposed in the literature for enhancing diversity among individual networks in an ensemble, such as seeding the networks with different random weights, choosing different network architectures, or using bootstrap samples of the training data [132, 39].

In the context of SAN models, autoencoder diversity can be imposed in two ways: (a) through training on different datasets, and (b) by utilizing different auxiliary gender classifiers. Intuitively, an ensemble of classifiers can only be useful if individual classifiers do **not** make similar errors on the test data [132, 63, 80]. To benefit from ensembles, it is thus critical to ensure error diversity, which can be accomplished by assembling the ensemble from a diverse set of classifiers. A number of approaches to explicitly measure ensemble diversity have been reported in the literature [80].

Among the novel contributions of this work is the development of ensemble methods for SANs

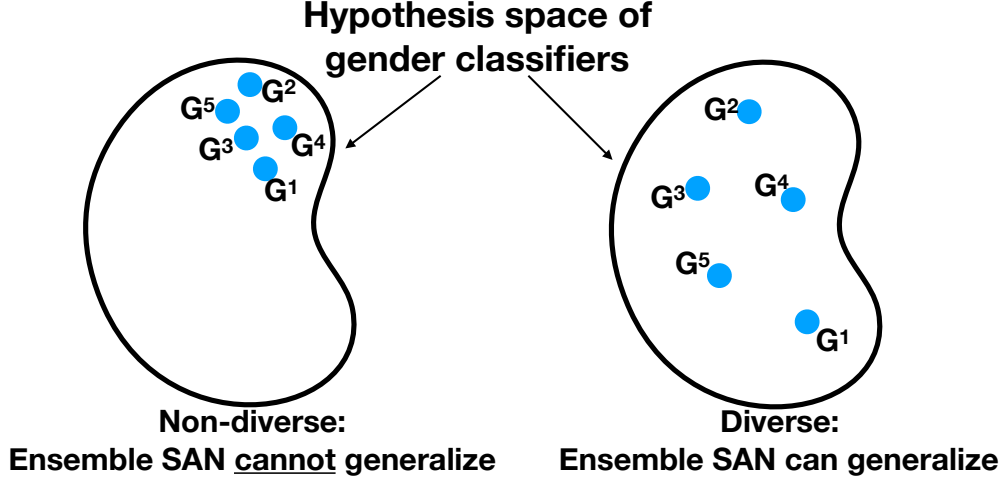


Figure 4.1: Diversity in an ensemble SAN can be enhanced through its auxiliary gender classifiers (see Figure 4.2). When the auxiliary gender classifiers lack diversity, ensemble SAN cannot generalize well to arbitrary gender classifiers.

using oversampling and data augmentation techniques. As shown in Figure 4.1, if auxiliary gender classifiers that are used to build a SAN lack diversity, the ensemble SAN *cannot* generalize to arbitrary classifiers. Therefore, in order to ensure generalizability, we (1) diversify the auxiliary gender classifiers and (2) diversify the autoencoder component of the SANs during the training phase.

4.2.3 Ensemble SAN Architecture

The original SAN model used single-attribute prototype images, which were computed by averaging over all male and female images, respectively, in the training dataset [101]. However, this approach does not take other demographic attributes into account, such as race and age, which increases the risk of introducing a systematic bias to the perturbed images if certain attributes are over- or under-represented in the training dataset. This issue is addressed in the current work.

Proposed Ensemble Model: The overall architecture of the proposed model is shown in Figure 4.3. The ensemble consists of t individual SAN models that are trained independently as

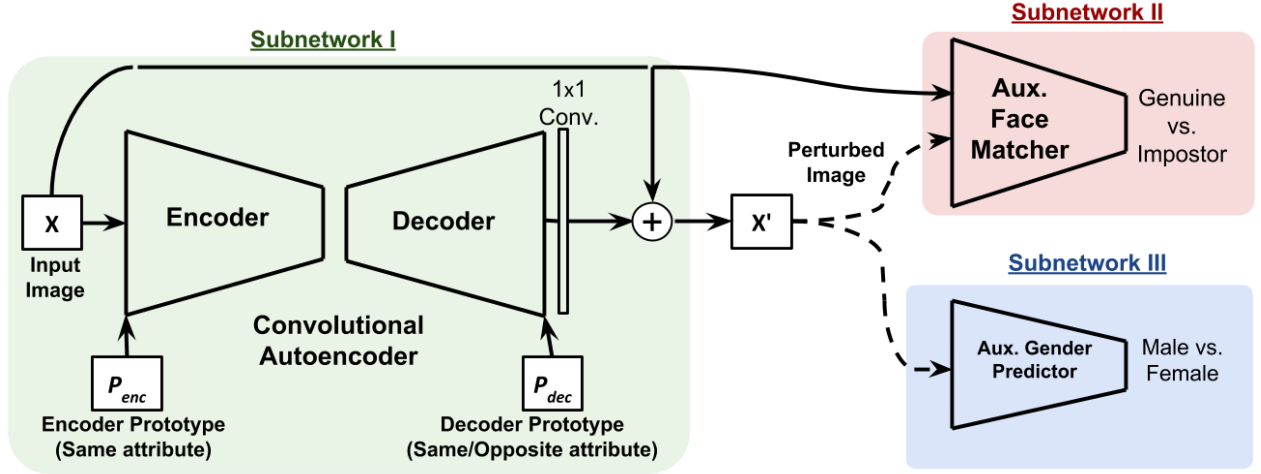


Figure 4.2: Architecture of the original SAN model [101].

will be discussed later. Each model is associated with an individually pre-trained auxiliary gender classifier and a pre-trained auxiliary face matcher.³ After the training of a SAN model has been completed, the auxiliary networks (gender classifier and face matcher) are discarded, and each SAN model S^j is used to generate an output image Y^j ($j \in \{1, \dots, t\}$) from an input image X , which results in a total of t output images.

We further propose that taking attributes other than just the attribute of interest (i.e., gender) into account reduces side-effects such as modifications to the race and age of an input image. Considering three binary attributes, gender (male, female), age (young, old), and race (black, white), we can categorize an input image into one of eight disjoint groups. For each group, we generate a prototype image, which is the average of all face images from the training dataset that belong to that group. Hence, given eight distinct categories or groups, eight different prototypes are computed. Next, an opposite-attribute prototype is defined by flipping one of the binary attribute labels of an input image. For example, if the input image had the attribute labels {young, female, white},

³The term auxiliary is used to indicate that these gender classifiers and face matchers are *only used during training* and *not* associated with any of the “unseen” gender classifiers and face matchers that will be used in the evaluation phase.

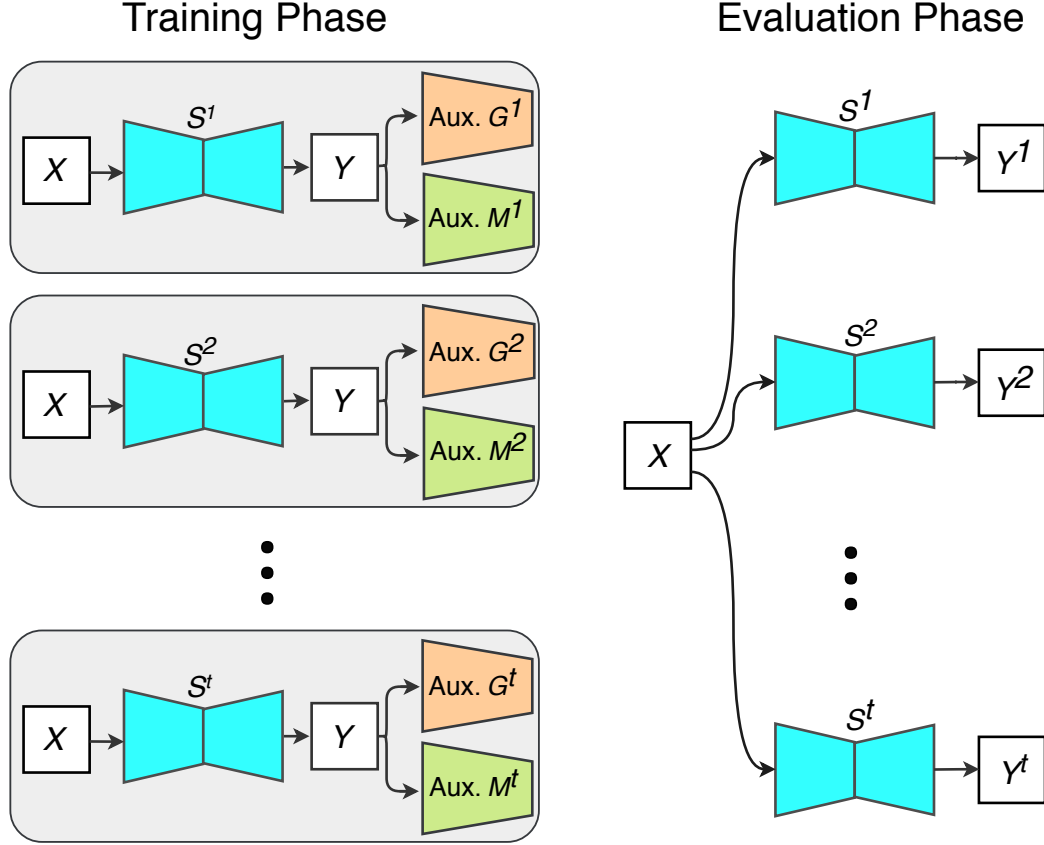


Figure 4.3: Schematic of the proposed ensemble of t SAN models. During the training phase, each SAN model, S^i , is associated with an auxiliary gender classifier G^i and an auxiliary face matcher M^i (common across all SANs). During the evaluation phase, the trained SAN models are used to generate t outputs $\{Y^1, Y^2, \dots, Y^t\}$.

the opposite-gender prototype chosen for gender perturbation would be {young, male, white}. The face prototype for each group is shown in Figure 4.4, and is computed by aligning the corresponding faces onto the the average face shape of each group.

The similarities and differences between the originally proposed SAN model and the ensemble SANs developed in this work are summarized below:

- The autoencoder, auxiliary gender classifier, and auxiliary face matcher architectures are similar to the original SAN model.
- In contrast to the original SAN model, we construct face image prototypes to reduce alter-

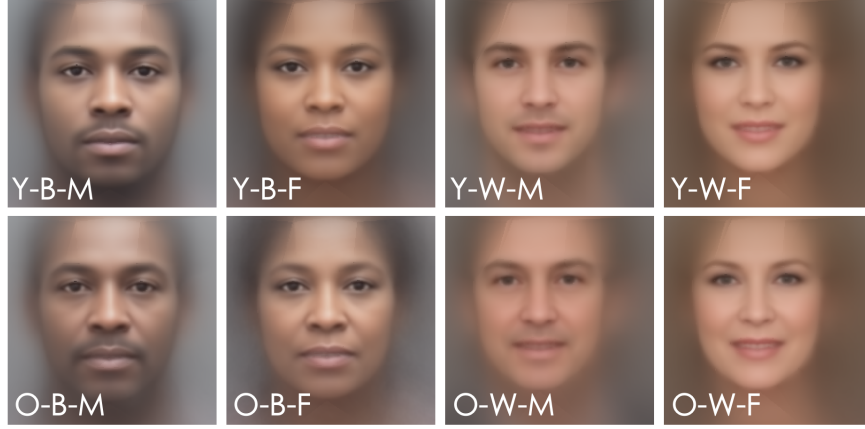


Figure 4.4: Face prototypes computed for each group of attribute labels. The abbreviations at the bottom of each image refer to the prototype attribute-classes, where Y=young, O=old, M=male, F=female, W=white, B=black.

ations to non-target attributes such as age and race.

- Instead of training a single SAN model, we create an ensemble of diverse SAN models that extend the range of arbitrary gender classifiers that can be confounded while preserving the utility of arbitrary face matchers.

4.2.4 Ensemble of SANs: Training Approach

To obtain a diverse set of SAN models, we trained the individual SAN models using different initial random weights. Further, we enhanced the diversity among the models by designing three different *training schemes* for the auxiliary gender classifier component of the SAN model as illustrated in Figure 4.5 and further described below.

- **E1 (regular):** Consists of five SANs, where the auxiliary gender classifier in each SAN model was initialized with different initial random weights. The models were trained on the CelebA training partition without resampling.

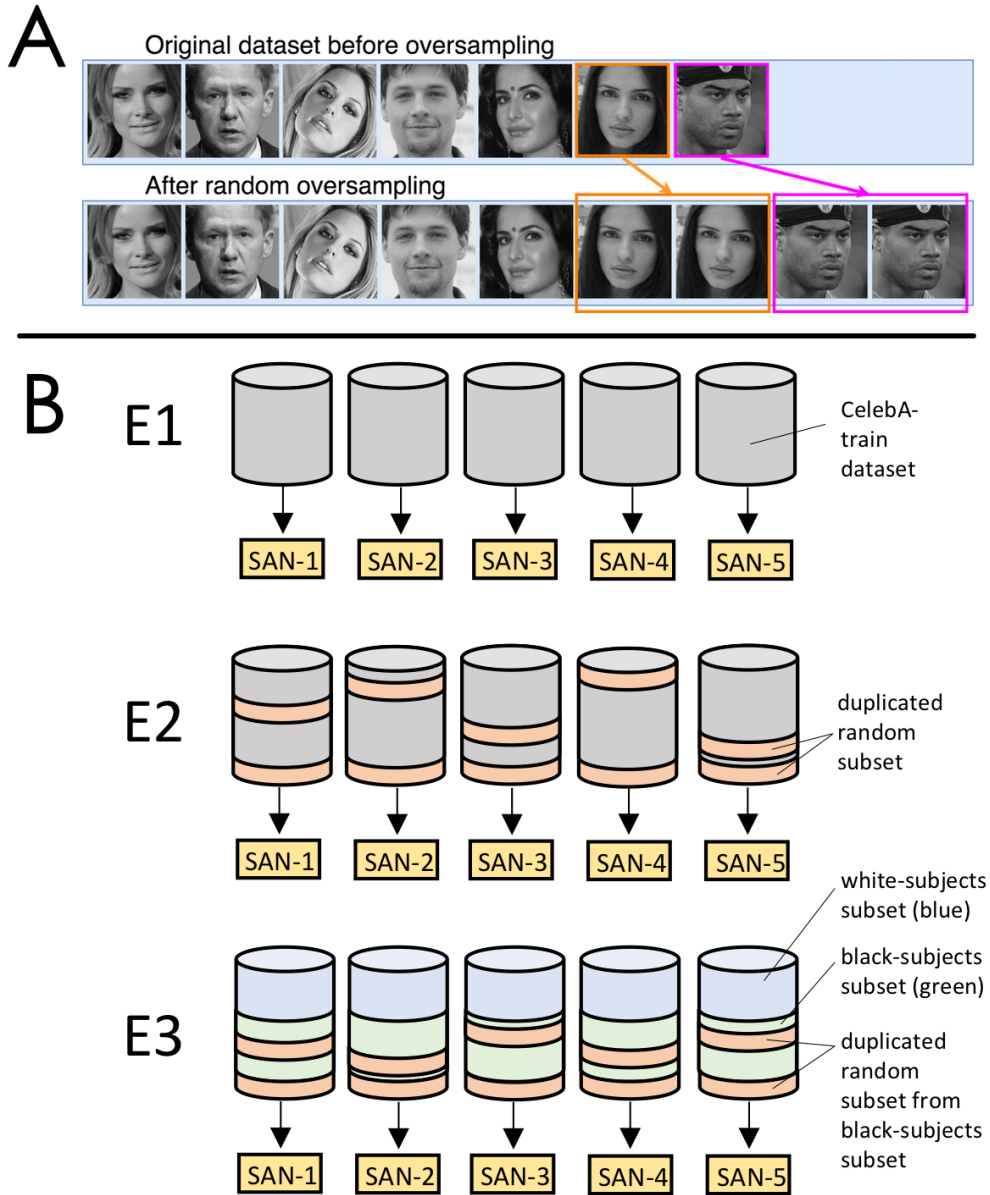


Figure 4.5: An example illustrating the oversampling technique used for enforcing diversity among SAN models in an ensemble. A: A random subset of samples are duplicated. B: Different Ensemble SANs (E1, E2, and E3) are trained on the CelebA-train dataset. SANs of the E1 ensemble are trained on the same dataset with different random seeds. In addition to using different random seeds, E2 SAN models are trained on datasets created by resampling the original dataset (duplicating a random subset of the images). Finally, for E3, a random subset of black subjects was duplicated for training the different SANs in the ensemble.

- **E2 (subject-based oversampling):** Consists of five SANs similar to E1, but in addition to choosing different initial random weights for the auxiliary gender classifiers, we applied a resampling technique by duplicating each sample from a random subset of subjects (representing 10% of the images in the training set). The selected subjects are disjoint across the five models, and the samples are duplicated four times.
- **E3 (race-based oversampling):** Five SANs were trained, similar to E1 and E2, but instead of resampling a random subset of subjects as in E2, we resampled instances of the minority race represented in the CelebA dataset to balance the racial distribution in the training data. In particular, a random 10%-subset of black samples was duplicated 40 times, that is, 10% of the black samples were copied 40 times and appended to the training dataset.

4.2.5 Datasets

We used five face image datasets in this work, viz., CelebA [86], MORPH [122], LFW [68], MUCT [94], and RaFD [83]. The details of the datasets, and how they were used in this work, are summarized in Table 4.1. Furthermore, the CelebA and MORPH datasets were split into non-overlapping training and test partitions, such that the train and test partitions are subject-disjoint (i.e., if a dataset contained multiple poses of the same person, these were all included either in the training set or the test set but not both). CelebA-train was used for training the auxiliary gender classifiers under the three schemes mentioned in the previous section, as well as for training all the individual SAN models. The face prototypes were computed using the CelebA-train and MORPH-train datasets. The remaining datasets were used for evaluating the performance of the SAN models on unseen gender classifiers and unseen face matchers.

Table 4.1: Overview of datasets used in this study. The letters in the “Usage” column indicate the tasks for which the datasets were used. A: training auxiliary gender classifiers, B: SAN training, C: SAN evaluation, D: constructing unseen gender classifiers used for evaluating SAN models.

Dataset	#male subjects / images	#female subjects / images	Usage
CelebA-train	4482 / 73,549	5163 / 103,772	A, B
CelebA-test	502 / 7929	581 / 11,511	C
MORPH-train	10,363 / 41,587	1938 / 7567	D
MORPH-test	1143 / 4643	224 / 863	C
LFW	4205 / 10,064	1448 / 2905	D
MUCT	131 / 1844	145 / 1910	C
RaFD	42 / 1008	25 / 600	C

4.2.6 Obtaining Race Labels

Since race labels are not provided in the face datasets considered in this study, we designed a procedure to efficiently label the face images:

1. Predict the racial labels for individual face images using a *commercial-off-the-shelf* (COTS) software.
2. Aggregate the COTS predictions for each subject (for whom multiple face images with different poses are present in a given dataset) by majority voting. For example, if five face images of a given subject exist and three face images are labeled as *white* and two face images are labeled as *black*, the label *white* was assigned to all five face images of the given subject.
3. Group the subjects based on their predicted majority class label from the previous step. Then, visually inspect one face image per subject and correct the class labels for all face images of a given subject if the class label was assigned incorrectly.

4.3 Experiments and Results

As described in Section 4.2, we trained and evaluated three auxiliary gender classifiers associated with the three ensemble SAN models: E1, E2, and E3. Table 4.2 summarizes the performance of these three models in terms of their gender classification errors on the CelebA-test and MORPH datasets. While the performance of E1 and E2 are similar, E3 outperforms E1 and E2 on MORPH. Given that 77% of the face images in the MORPH dataset have the class label *black*, it is evident that oversampling examples of the under-represented race during training could have helped overcome the algorithmic bias in gender classification.

Based on the results from Table 4.2, the ensemble of auxiliary gender classifiers in E3 achieves higher accuracy on the MORPH-test dataset. In addition, in Table 4.2, we computed the entropy as an empirical measure of ensemble diversity [80], and the results confirm that auxiliary gender classifiers in E3 have higher diversity. Hence, we selected the ensemble SAN E3 for evaluation on unseen gender classifiers and face matchers. Figure 4.6 shows example images with their perturbed outputs from each of the SAN models in E3. In the remainder of the document, SAN-1 to SAN-5 denote the 5 models pertaining to E3.

4.3.1 Unseen Gender Classifiers

In order to assess the performance of the proposed ensemble SAN in confounding an arbitrary gender classifier, we used 9 gender classifiers that were not available to any of the SAN models during training, as noted in Table 4.3. We used five pre-trained models: a commercial-of-the-shelf gender classifier (G-COTS), IntraFace [142], AFFACT [60], and two additional Convolutional Neural Network(CNN)-based gender classifiers from Ref. [13]. In addition to the five existing gender classifiers, we also included CNN-based gender classifiers that were trained on three

Table 4.2: Error rates of the auxiliary gender classifiers on the CelebA / MORPH-test datasets. E3 (95% confidence interval: 5.46%–5.63%) performs significantly better ($p \ll 0.01$) on the MORPH dataset compared to E1 (CI95: 6.24%–6.42%) and E2 (CI95: 6.25%–6.43%). At the end, ensemble diversities are reported [80].

Auxiliary Classifier	E1: Regular	E2: Subject-based	E3: Race-based
G^1	2.25 / 5.56	2.07 / 6.24	2.29 / 5.17
G^2	2.11 / 6.20	2.03 / 6.45	1.97 / 5.28
G^3	2.03 / 6.38	2.06 / 6.46	2.13 / 5.04
G^4	2.21 / 6.97	2.03 / 5.85	1.99 / 6.96
G^5	2.42 / 6.53	2.12 / 6.72	2.02 / 5.28
Average:	2.20 / 6.33	2.06 / 6.34	2.08 / 5.55
Diversity:	0.047 / 0.079	0.044 / 0.076	0.045 / 0.083

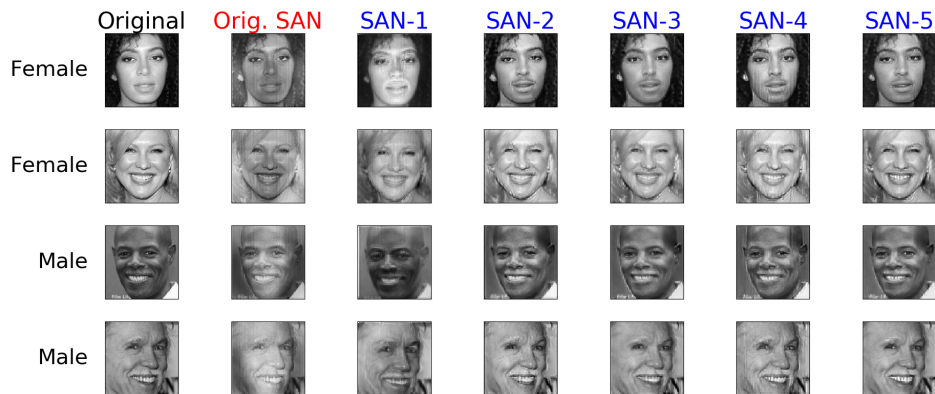


Figure 4.6: Four example images with their perturbed outputs using the original SAN model from Ref. [101] and the outputs of five individual SAN models. Note that the ensemble SAN generates diverse outputs that is necessary for generalizing to arbitrary gender classifiers.

datasets, MORPH-train, LFW, and a merged version of MORPH-train and LFW. The CNN architecture of each of these gender classifiers contain five convolutional layers, each followed by SELU [78] activation units and a max-pooling layer. Inspired by HyperFace [116], the feature maps from the third convolution layer are fused with those of the last convolution layer to provide features with hierarchical receptive fields for classification. The fused feature maps then undergo a global average pooling prior to two fully-connected layers, which were followed by a final sigmoid activation function. Two CNN models, named CNN-LFW and CNN-MORPH, were trained on the MORPH-train and LFW datasets, respectively, after the datasets were balanced by oversampling the female samples. A third CNN model, called CNN-Merged, was trained on the merged MORPH-train/LFW dataset, after balancing the male/female ratio, as well as balancing the size of the two datasets since MORPH-train is almost five times larger than LFW. Furthermore, we also applied data augmentation during training by randomly adjusting illumination and contrast using the Torchvision library and PyTorch software [110]. Finally, for the fourth gender predictor, we used CNN-Merged but applied data augmentation in the evaluation phase as suggested in [60]. Some examples of this data augmentation during evaluation are shown in Figure 4.7. The illumination and contrast of a test sample is varied randomly to obtain seven samples. The augmented face images were then evaluated by the gender predictor, CNN-Merged, and the average score of the seven different modified test samples is reported; this is denoted as CNN-Aug-Eval (examples of the seven augmentation methods are shown in Figure 4.7, columns 2-8).

The performance of all nine unseen gender classifiers is shown in Figure 4.8. The ROC curves of gender prediction on the perturbed images generated by each SAN model is compared with the ROC curves of gender prediction on the original samples from the CelebA-test, MORPH-test, MUCT, and RaFD datasets. The ROC curves indicate that the gender classification performance varies widely across the SAN models. In certain cases, the perturbations made by some of the

Table 4.3: List of the nine unseen gender classifiers used for evaluating the outputs of the proposed ensemble SAN models.

Pre-trained models	In-house trained CNN models
G-COTS	CNN-MORPH
IntraFace [142]	CNN-LFW
AFFACT [60]	CNN-Merged
Ref. [13]-A	CNN-Aug-Eval
Ref. [13]-B	

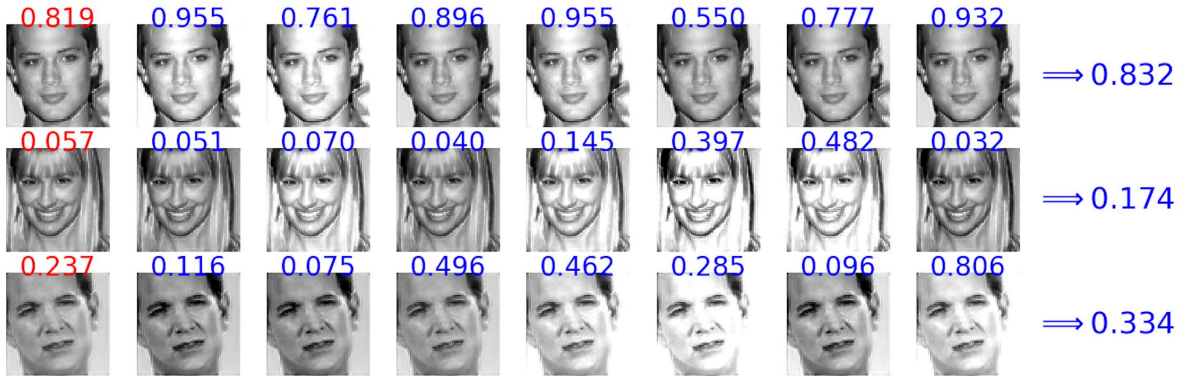


Figure 4.7: Data augmentation at the evaluation phase using random illumination and contrast adjustments. The left column shows the perturbed images before augmentation, and the next seven columns show the samples used for augmentation along with their gender prediction scores. Finally, average prediction scores obtained using the CNN-Merged model on these seven augmented samples are computed and denoted as CNN-Aug-Eval in the text.

SAN models improve the performance of the gender classifier compared to their performance on the original data. In contrast to the original SAN model [101] (also shown in Figure 4.8 for comparison), it is always possible to find at least one SAN model in the ensemble that can effectively degrade the gender classification performance for a given image.

To illustrate the advantage of the proposed ensemble SAN over a single SAN, we did the following. For each unseen gender classifier, we selected the best-perturbed sample for each face image, X , based on the ground-truth gender label as follows:

$$P_{best} = \begin{cases} \min_{i=1..5} P(S^i(X)), & \text{if } X \text{ is male;} \\ \max_{i=1..5} P(S^i(X)), & \text{if } X \text{ is female.} \end{cases} \quad (4.1)$$

The ROC curve using the best-perturbed sample is shown in Figure 4.8 for each gender classifier. The results suggests that diversity among individual SAN models is necessary for confounding unseen gender classifiers.

4.3.2 Unseen Face Matchers

Next, we show the performance of unseen face matchers on the original and perturbed samples. For this analysis, we utilized four face matchers: a commercial-of-the-shelf face matcher (M-COTS) that has shown state-of-the-art performance in face recognition, and face representation vectors obtained from DR-GAN [143], FaceNet [127], and OpenFace [10]. For the latter three choices, we used the cosine-similarity measure between a pair of face vectors to measure their degree of dissimilarity. Figure 4.9 shows the performance of these four matchers on the four evaluation datasets. The performance of M-COTS and DR-GAN on perturbed samples matches closely with that of original samples, except for some minor deviations for the DR-GAN matcher on the RaFD

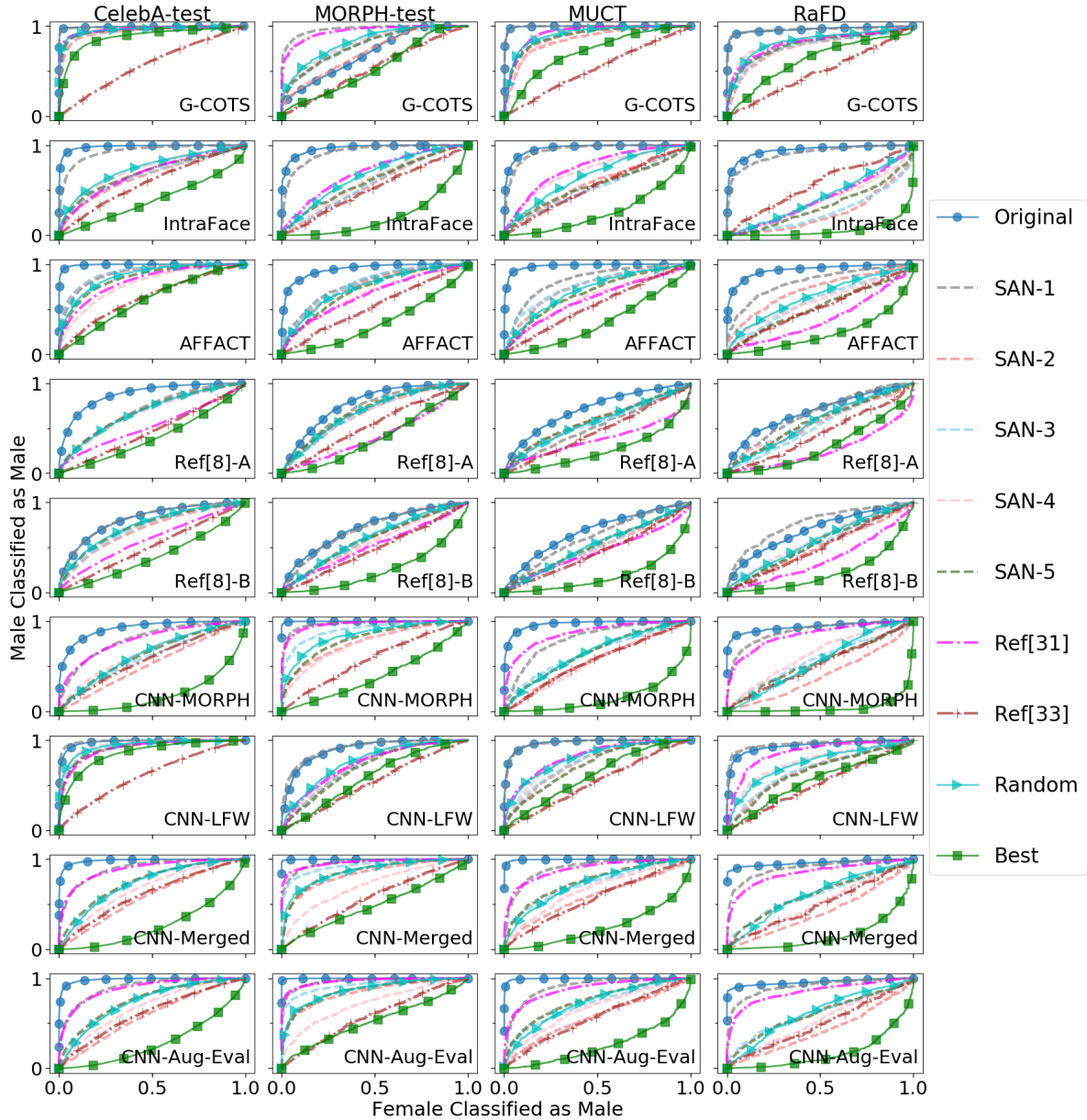


Figure 4.8: ROC curves of the nine unseen gender classifiers (each row corresponds to one classifier) on the perturbed images generated by each SAN model of the E3 ensemble on four evaluation datasets: CelebA-test, MORPH-test, MUCT, and LFW. Note that the gender classification performance shows a wide degree of change on perturbed samples, but in all cases, there is always one output from each ensemble degrading the performance.

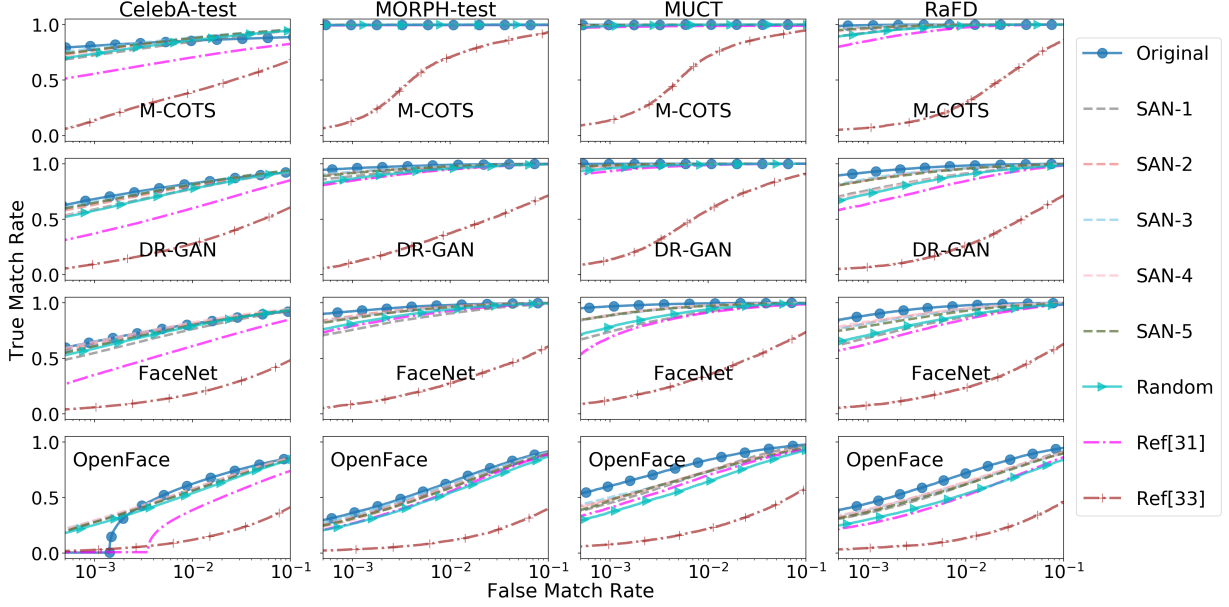


Figure 4.9: ROC curves of the four unseen face matchers (each row corresponds to one matcher) on the perturbed images generated by each SAN model of the E3 ensemble on four evaluation datasets: CelebA-test, MORPH-test, MUCT, and RaFD. Note that the matching performance is mostly retained except for some small degradations in the case of FaceNet and OpenFace.

dataset. Performance of FaceNet and OpenFace on perturbed samples shows marginal deviation from that of original samples. In contrast, the face mixing approach [107] results in significant drop in performance of unseen face matchers, thereby suggesting that these outputs have lost their biometric utility.

Practical Implementation: In a practical application, we may not have *a priori* knowledge about the arbitrary gender classifier. Given an arbitrary gender classifier, one way to utilize the ensemble SAN is by randomly selecting one of the t perturbed images. The result of such a random SAN model selection is shown in Figure 4.8. As the results illustrate, in most cases, randomly selecting a SAN model from the E3 ensemble results in better performance in terms of confounding arbitrary unseen gender classifiers compared to using a single SAN model. Furthermore, randomly selecting one SAN output does not degrade the face matching performance (Figure 4.9).

Randomly selecting a perturbed sample tends to conceal the true gender label, since flipping

the predicted label may or may not result in the true label of the original sample.

While gender recognition from a human perspective was not the main focus of this study, we may consider a human observer as an arbitrary gender classifier. The degree to which the gender information is concealed from human observers will be a subject of future studies.

4.4 Summary and Future Work

In this work, we focused on addressing one of the main limitations of previous gender privacy methods, viz., their inability to generalize across multiple previously unseen gender classifiers. In this regard, we proposed an ensemble technique that generates diverse perturbations for an input face image, and at least one of the perturbed outputs is expected to confound the gender information with respect to an arbitrary gender classifier. We showed that randomly selecting perturbations for face images stored in a biometric database is an effective way for enabling gender privacy. In addition, we have showed that the face matching accuracy is retained for all perturbed outputs, thereby preserving the biometric utility of the face images.

In the next chapter, we will extend the ensemble SAN model introduced in this chapter. In particular, we propose an algorithm to stack SAN models in order to further improve the generalization ability of SAN models.

Chapter 5

FlowSAN: Privacy-enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers

Portions of this chapter have been published in:

- V. Mirjalili, S. Raschka, A. Ross, "FlowSAN: privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers", IEEE Access, Vol. 7, pp. 99735-99745, 2019.

5.1 Introduction

Previously, in Chapter 3 we developed Semi-Adversarial Networks (SAN) [101] for imparting demographic privacy to face images, where a face image is modified such that the matching utility of the modified face image is retained while the automatic extraction of gender information is confounded. We empirically showed that the ability to predict gender information, using an unseen gender classifier from outputs of the SAN model, is successfully diminished. In Chapter 4, we defined the generalizability of the SAN model as its ability to confound arbitrary unseen¹ gender

¹The term “unseen” indicates that a certain classifier (or face matcher) was not used during the training stage. On the contrary, the term “auxiliary” in this work refers to the classifier (or face matcher) that is used during the training

classifiers. Generalizability is an important property for real-world privacy applications since the lack thereof implies that there exists at least one gender classifier that can still reliably estimate the gender attribute from outputs of the SAN model and, therefore, jeopardizes the privacy of users. In order to address the generalizability issue of SAN models, in this chapter, we propose the FlowSAN model, that progressively degrades the performance of unseen gender classifiers. Extensive experiments on a variety of independent gender classifiers and face image datasets show that the proposed FlowSAN method (Fig. 5.1) results in a substantially improved generalization performance compared to the original SAN method with regard to concealing gender information while retaining face matching utility.

In this work, we address the generalization issue of the SAN method using a novel stacking paradigm that will successively enhance the perturbations for confounding an arbitrary unseen gender classifier as illustrated in Fig. 5.1. We refer to this method as FlowSAN. The primary contributions of this work are as follows:

- Designing the FlowSAN model that can successively degrade the performance of arbitrary unseen gender classifiers;
- Generalizing the FlowSAN model to multiple arbitrary gender classifiers;
- Demonstrating the practicality and efficacy of the proposed approach in confounding the gender information for real-world privacy applications via extensive experiments involving broad and diverse sets of datasets.

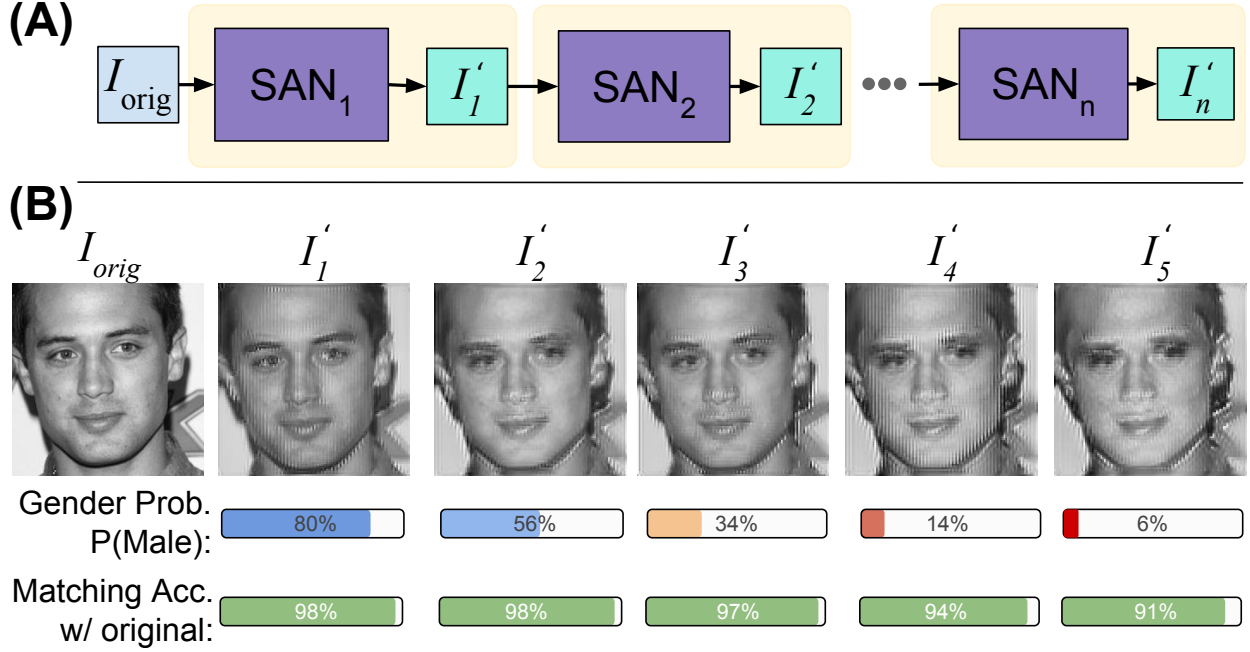


Figure 5.1: Illustration of the FlowSAN model, which sequentially combines individual SAN models in order to sequentially perturb a previously unseen gender classifier, while the performance of an unseen face matcher is preserved. A: An input gray-scale face image I_{orig} is passed to the first SAN model (SAN_1) in the ensemble. The output image of SAN_1 , I'_1 , is then passed to the second SAN model in the ensemble, SAN_2 , and so forth. B: An unmodified face image from the CelebA [86] dataset (I_{orig}) and the perturbed variants I'_i after passing it through the different SAN models sequentially. The gender prediction results measured as probability of being male ($P(\text{Male})$) as well as the face match score between the original (I_{orig}) and the perturbed images (I'_i) are shown.

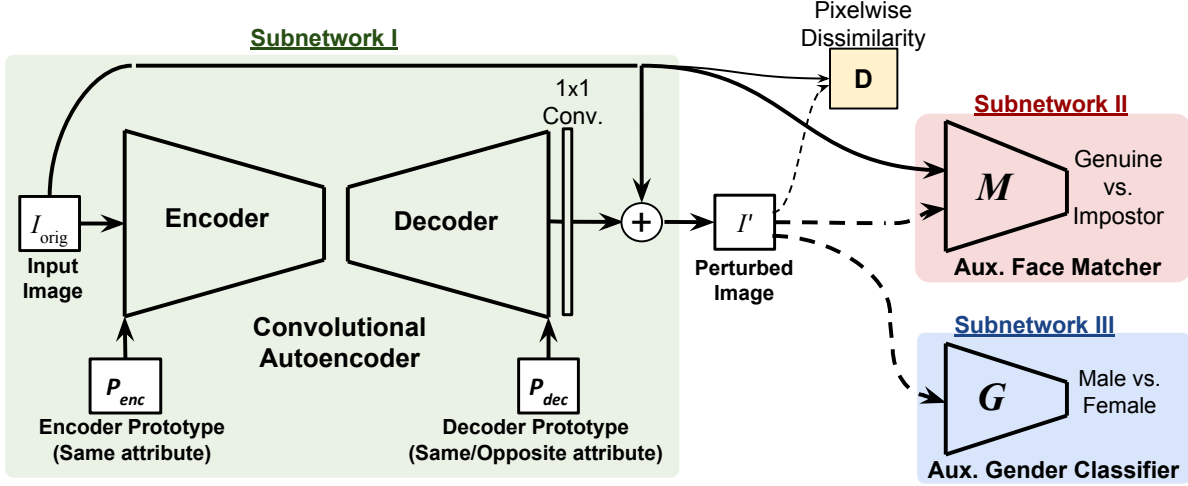


Figure 5.2: Architecture of the original SAN model [101] composed of three subnetworks: I: a convolutional autoencoder [16], II: an auxiliary face matcher (M), and III: an auxiliary gender classifier (G). In addition, the unit D computes the pixelwise dissimilarity between input and perturbed images during model training.

5.2 Proposed Method

Original SAN model [101]: The SAN model for imparting gender privacy to face images was first proposed in [101], and the overall architecture is shown in Fig. 5.2. The SAN model leverages pre-computed face prototypes, which are average face images for each gender. SAN consists of three subnetworks: 1) a convolutional autoencoder that perturbs an input face image via face prototypes, 2) an auxiliary face matcher, which is a convolutional neural network (CNN), and 3) a CNN-based auxiliary gender classifier. The input to the convolutional autoencoder is a gray-scale² face image I_{orig} , of size $224 \times 224 \times 1$, fused with a face prototype belonging to the same gender (P_{sm}). After the fused input image was passed through the encoder and decoder networks, the face prototypes (P_{sm} prototype face image from the same gender as input image, or P_{op} the prototype face image of the opposite gender) are added as additional channels to the resulting 128-channel feature-map phase.

²Since most face-matchers work with gray-scale face images, we used gray-scale images in all experiments to allow for a fair comparison between matchers based on the same input data.

representation. Finally, a 1×1 -convolutional operation is used to reduce the number of channels in the resulting feature-maps to a $224 \times 224 \times 1$ -dimensional output image, which is denoted as I'_{sm} or I'_{op} , depending on the type of prototype used by the decoder:

$$I'_{\text{sm}} = \text{SAN}(I_{\text{orig}}; P_{\text{sm}}), \text{ and} \quad (5.1)$$

$$I'_{\text{op}} = \text{SAN}(I_{\text{orig}}; P_{\text{op}}).$$

These output images, I'_{sm} and I'_{op} , are then passed to both the auxiliary face matcher and the auxiliary gender classifier. The auxiliary face matcher predicts whether the original and the perturbed face images belong to the same individual via a face match score. The gender classifier predicts the gender of the input and output images via gender probabilities for male and female.³ For the auxiliary face matcher, the pre-trained, publicly available VGG-face model [109] is used, which computes the face representation vectors for an input face image, and the similarity between two face representation vectors determines the associated match-score.

Three different loss functions are defined based on the outputs from the autoencoder, the auxiliary gender classifier, and the auxiliary face matcher. The first component of the loss function, \mathcal{J}_D , measures the pixelwise dissimilarity between the input and the output from the same-gender prototype I'_{sm} , which is used to ensure that the autoencoder subnetwork is able to construct realistic face images:

$$\mathcal{J}_D(I_{\text{orig}}, I'_{\text{sm}}) = \frac{1}{h \times w} \sum_{i=1}^{h \times w} \mathcal{H}(I_{\text{orig}}^{(i)}, I'_{\text{sm}}^{(i)}), \quad (5.2)$$

³As explained in Chapter 1, throughout this work we have assumed binary labels for gender; however, it must be noted that societal and personal interpretation of gender can result in many more classes.

where \mathcal{H} indicates the cross-entropy function for the binary case, defined as

$$\mathcal{H}(p, q) = -(p \log(q) + (1 - p) \log(1 - q)). \quad (5.3)$$

The second loss term, \mathcal{J}_M , is the squared L^2 distance between the face representation vectors obtained from the auxiliary face matcher (VGG-face network [109]) for the input image and the perturbed output, making the autoencoder learn how to perturb face images such that the accuracy of the face matcher is retained:

$$\mathcal{J}_M(I_{\text{orig}}, I'_{\text{op}}) = \|\mathcal{R}_M(I_{\text{orig}}) - \mathcal{R}_M(I'_{\text{op}})\|_2^2, \quad (5.4)$$

where $\mathcal{R}_M(I)$ and $\mathcal{R}_M(I'_{\text{op}})$ indicate the face representation vectors for the input image and the perturbed output based on the opposite-gender prototype.

Finally, the third loss term, \mathcal{J}_G , is the cross-entropy loss function applied to the gender probabilities computed by the auxiliary gender classifier, G , on the two perturbed output images. Here, the ground-truth label y of the input image is used for I'_{sm} , but the reverse $(1 - y)$ is used for I'_{op} :

$$\mathcal{J}_G(y, I'_{\text{sm}}, I'_{\text{op}}) = \mathcal{H}(y, G(I'_{\text{sm}})) + \mathcal{H}(1 - y, G(I'_{\text{op}})). \quad (5.5)$$

The total loss, \mathcal{J}_{tot} , is the weighted sum of the three individual loss functions described in the previous paragraphs,

$$\mathcal{J}_{\text{tot}} = \lambda_1 \mathcal{J}_D + \lambda_2 \mathcal{J}_M + \lambda_3 \mathcal{J}_G, \quad (5.6)$$

where the parameters λ_i are the relative weighting terms that can be chosen uniformly or adjusted via hyperparameter optimization.

In the remaining part of the chapter, we use notation I' for the output of a SAN model on a face image I_{orig} when using the opposite-gender prototype, i.e., $I' = \text{SAN}(I_{\text{orig}}; P_{\text{op}})$.

Based on our previous study [99], we employed a data augmentation and resampling scheme for training the auxiliary gender classifiers as a means to diversify the SAN models. In particular, by resampling the instances belonging to the underrepresented race in the CelebA [86] dataset, we aimed to balance the racial distribution in the training data. In this regard, we generated five resampled training datasets, where in each one a random disjoint subset of samples from the underrepresented race was replicated 40 times. This is an effort to enhance the diversity among the SAN models in an ensemble. The resampling approaches that are used to mitigate the imbalances in the different training datasets employed in this study are described in [99].

5.2.1 Training and Evaluation of an Ensemble SAN model

In our previous work [99], we proposed an ensemble approach for generalizing SAN models to unseen gender classifiers. The objective of an ensemble SAN was to create n SAN models such that their union can span a larger subset of the hypothesis space compared to a single SAN model. Therefore, for a new test image and an arbitrary unseen gender classifier, G , it is likely that at least one of these SAN models in the ensemble is able to confound G . For training an ensemble of SANs, we start with n auxiliary gender classifiers, $\mathcal{G} = \{G_1, G_2, \dots, G_n\}$, which were trained using different data augmentation schemes (to achieve higher diversity among classifiers), and a pre-trained face matcher M . Then, we train n SAN models, where SAN_i is associated with the auxiliary gender classifier G_i , as shown in Fig. 5.3. According to the original SAN model proposed in [101], the loss function for training each model is composed of three components: gender loss, matching loss, and pixelwise dissimilarity loss (Eq. 5.6). Note that the ensemble of SAN models described with this setting can be trained in parallel since each SAN model is independent

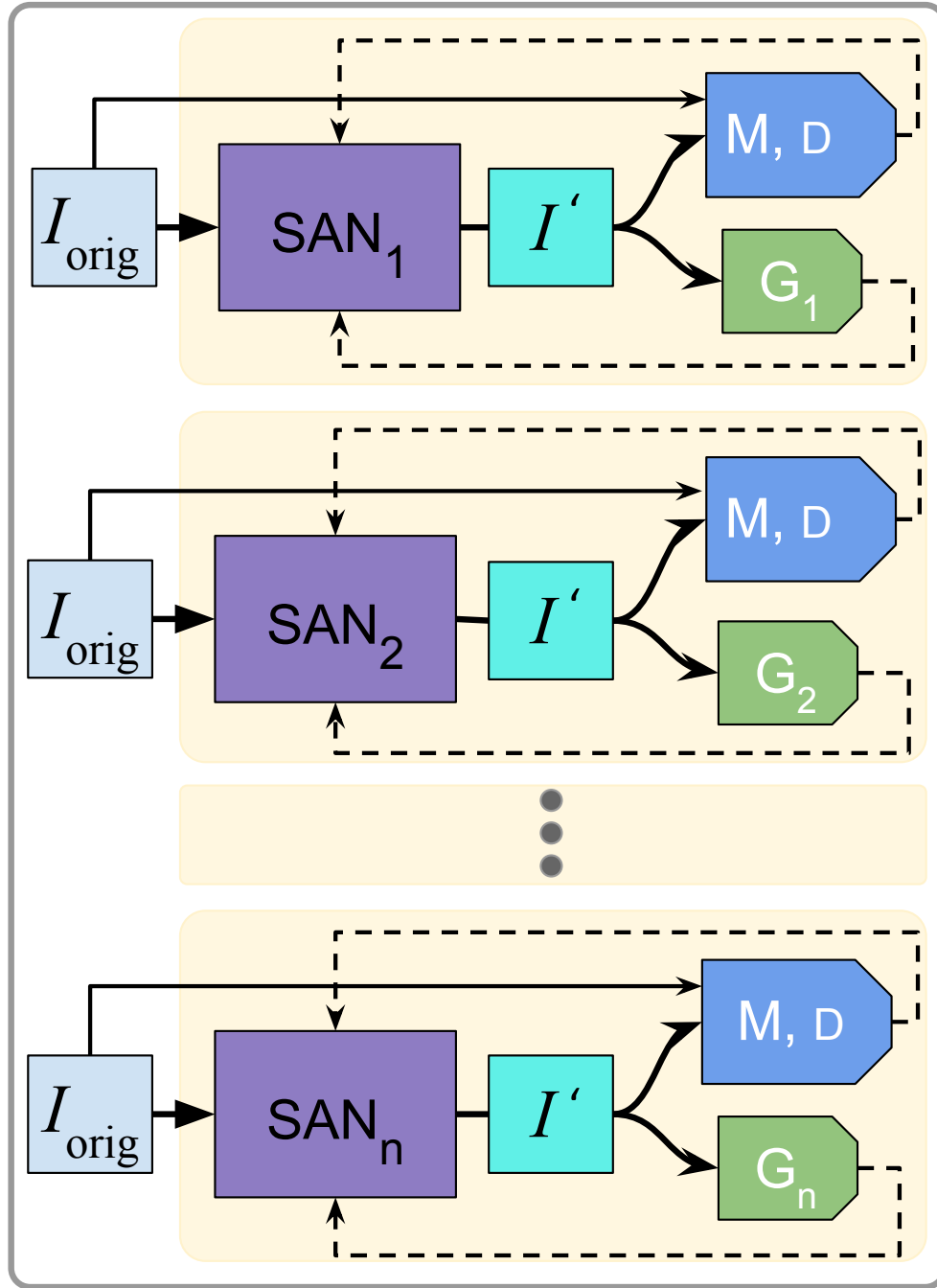


Figure 5.3: Illustration of an ensemble SAN, where individual SAN models are trained *independent* of each other using n diverse, pre-trained, auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, \dots, G_n\}$), and a face matcher M that computes face representation vectors for both input face image I_{orig} and the output of the SAN model. D refers to a module that computes pixelwise dissimilarity between an input and output face image.

of others, and each individual SAN model takes unmodified images as input (Fig. 5.3).

Evaluation of an ensemble of models, that were trained independently, can be performed in two ways:

1. Averaging: Evaluating the ensemble of SANs by computing the average output image from the set of n outputs as shown in Fig. 5.4-A.
2. Gibbs: Randomly selecting the output of one SAN model (Fig. 5.4-B).

These two ensemble-based methods serve as a basis for the comparison with the proposed FlowSAN method, which is described in the following section.

5.2.2 FlowSAN: Connecting Multiple SAN Models

Assume there exists a large set of gender classifiers $\mathcal{G} = \{G_1, G_2, \dots, G_g\}$, where each $G_i(I)$ predicts the probability that a face image I belongs to a male individual. Furthermore, suppose there exists a set of m face-matchers denoted by $\mathcal{M} = \{M_1, M_2, \dots, M_m\}$, where each $M_i(I_a, I_b)$ computes the match score between a pair of face images, I_a and I_b . Our goal is to design an ensemble of n SAN models, $\mathcal{E} = \langle S_1, S_2, \dots, S_n \rangle$, that, once they are sequentially stacked together, can be shown to generalize to confound unseen gender classifiers in \mathcal{G} . We hypothesize that stacking diverse SANs sequentially would have a cumulative effect, where each SAN adds perturbations to an input image that confound a particular gender classifier. Therefore, stacking SANs would enhance their generalizability in terms of decreasing the performance of multiple, diverse gender classifiers.

We define a recursive function $\Psi_{\mathcal{E}}(I_{\text{orig}}, t)$ for stacking SAN models in $\mathcal{E} = \{\text{SAN}_1, \dots, \text{SAN}_n\}$,

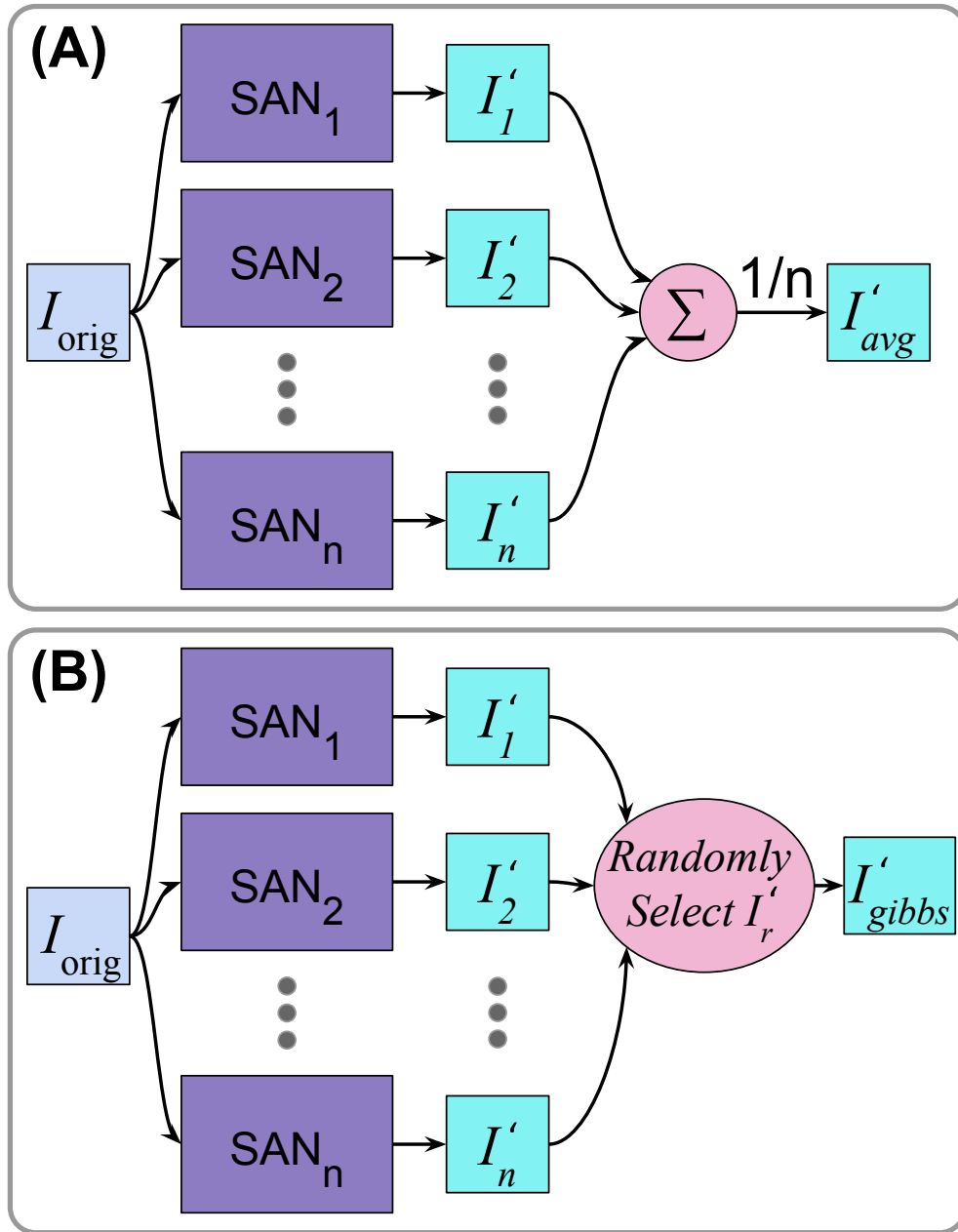


Figure 5.4: Two approaches for evaluating an ensemble of SAN models: Combining a set of n SAN models trained in the ensemble by (A) averaging n output images, and (B) randomly selecting an output (Gibbs).

as follows:

$$\Psi_{\mathcal{E}}(I_{\text{orig}}, t) = \begin{cases} \text{SAN}_1(I_{\text{orig}}) & \text{if } t = 1, \\ \text{SAN}_t(\Psi_{\mathcal{E}}(I_{\text{orig}}, t-1)) & \text{otherwise.} \end{cases} \quad (5.7)$$

By varying t from 1 to n , $\Psi_{\mathcal{E}}(I_{\text{orig}}, t)$ produces a sequence of n output images $\langle I'_1, I'_2, \dots, I'_n \rangle$:

- $t = 1 \rightarrow I'_1 = \Psi_{\mathcal{E}}(I_{\text{orig}}, 1) = \text{SAN}_1(I_{\text{orig}}),$
- $t = 2 \rightarrow I'_2 = \Psi_{\mathcal{E}}(I_{\text{orig}}, 2) = \text{SAN}_2(\text{SAN}_1(I_{\text{orig}})),$
- ...
- $t = n \rightarrow I'_n = \Psi_{\mathcal{E}}(I_{\text{orig}}, n) = \text{SAN}_n(\dots \text{SAN}_1(I_{\text{orig}})).$

In particular, we hypothesize that for each $G_i \in \mathcal{G}$, the stacking of SAN models will progressively confound G_i . Since the individual SAN models were trained to have a minimal impact on face matching performance, we further hypothesize that the perturbations introduced in the output face images $\langle I'_1, \dots, I'_n \rangle$ from the stacked SAN models should not substantially affect the face recognition performance of the matchers in \mathcal{M} .

5.2.3 Training Procedure for Stacking SAN Models

The goal of this work is to develop a model that leverages the image perturbations induced by individual, diverse SAN models to broaden the spectrum of diverse gender classifiers that can successfully be confounded. To accomplish this goal, we designed and evaluated the FlowSAN model, where multiple individually-trained SAN models were sequentially combined.

This section describes the training procedure for the FlowSAN model, where SAN models $i = 1, \dots, n$ are trained in sequential order, each with their corresponding auxiliary gender classifier

and an auxiliary face matcher, which is common among all SANs. The first SAN model, $\text{SAN}_1 \in \mathcal{E} = \{\text{SAN}_1, \dots, \text{SAN}_n\}$, takes the original image as input and generates a perturbed output, I'_1 , while using the auxiliary gender classifier G_1 during its training. Then, once SAN_1 is trained, the entire training dataset is transformed by SAN_1 , and the transformed data is then used for training the next SAN model while using its corresponding auxiliary gender classifier. This process is repeated for SAN models $i = 1, \dots, n$, to obtain n SAN models that are trained in sequential order. Note that the matching loss is computed between face representation vectors (generated by a face matcher) of the SAN output with that of the corresponding original face image, as opposed to the input to the SAN model (which is already perturbed for $i \geq 2$). This is to ensure that the matching performance does not substantially decline as the sequence is expanded. Furthermore, we considered three different scenarios for the pixelwise dissimilarity loss:

1. Omitting the pixelwise dissimilarity loss term;
2. pixelwise dissimilarity with respect to the input, i.e., I'_{i-1} for SAN_i ;
3. pixelwise dissimilarity loss with respect to the original image I_{orig} for each of SAN models $i = 1, \dots, n$.

We evaluated all three different pixelwise loss function schemes listed above. However, we were unable to observe any noticeable differences except for some cases where the third scheme slightly outperformed the other two. Therefore, we only report the results of the third case in this chapter. The training procedure is illustrated in Fig. 5.5.

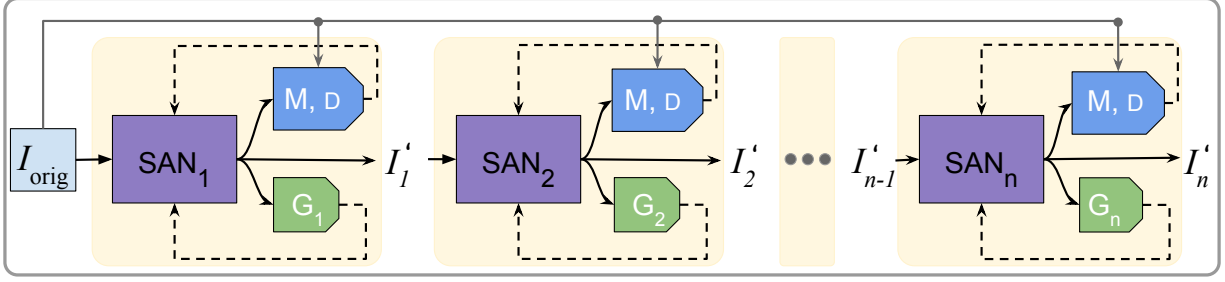


Figure 5.5: An illustration of a FlowSAN model: n SAN models are trained sequentially using n auxiliary gender classifiers ($\mathcal{G} = \{G_1, G_2, \dots, G_n\}$), and a face matcher M that computes face representation vectors for both input image I and the output of SAN model. Both auxiliary face matcher and the dissimilarity unit (D) use the original image along with the output of their corresponding SAN.

5.2.4 Evaluating the FlowSAN Model

During the model evaluation, the auxiliary networks (the auxiliary gender classifiers and auxiliary face matchers) from the individual SANs are discarded, and the n SAN models are stacked in the same sequence they were trained, in order to enhance their generalizability to arbitrary gender classifiers. In the FlowSAN model, the first SAN model (SAN_1) takes an original image (I_{orig}) as input and generates a perturbed output image I'_1 . This output image is then passed into the next SAN model in the sequence to obtain I'_2 , and so forth. In general, the i th SAN model (SAN_i for $i = 2, \dots, n$) takes the output of the previous SAN model (I'_{i-1}) as input and generates the perturbed output I'_i .

5.3 Experiments and Results

We designed two different protocols for training n SAN models:

- (a) Training an ensemble of SANs independent of each other as described in [99] (see Section 5.2.1);

(b) Training the FlowSAN model using the sequential procedure described in Section 5.2.2.

Protocol (a) was adapted from [99] and is further described in Section 5.2.1. For evaluating models trained in the ensemble, we applied two techniques: 1) taking the average output from SAN models which we denote as Ens-Avg, and 2) randomly selecting the output which we denote as Ens-Gibbs. In addition, similar to [99], we also define the *oracle best-perturbed* sample for a specific gender classifier, G :

$$\text{best}(I; \mathcal{E}, G) = \begin{cases} \arg \min_{\text{SAN}_i \in \mathcal{E}} G(\text{SAN}_i(I)) & \text{if } y = 1, \\ \arg \max_{\text{SAN}_i \in \mathcal{E}} G(\text{SAN}_i(I)), & \text{otherwise.} \end{cases} \quad (5.8)$$

The results of best-perturbed samples are denoted as Ens-Best. This analysis indicates which output from the ensemble model \mathcal{E} has resulted in the highest prediction error for a particular gender classifier G if the best output is selected.

The training of the FlowSAN model was initiated from the pre-trained individual SAN models in [99] and then trained for 10 additional epochs on the CelebA-train subset [86] (see Table 5.1) using the training procedure described in Section 5.2.2. Then, the models were stacked successively to generate a sequence of perturbed output images, $\langle I'_1, \dots, I'_n \rangle$.

As the FlowSAN model conceals the gender information in face images incrementally, it naturally produces a sequence of perturbed face images, where the length of this sequence is determined by its ensemble size. By varying the size of the ensemble, we can have a fair comparison between the ensemble approach vs. the FlowSAN model, such that the number of SANs used to obtain an output from the ensemble model is consistent with the number of SANs that are used to generate the output from the FlowSAN model.

For model evaluation and comparison, we used four test datasets: CelebA-test [86], MORPH-

Table 5.1: Overview of datasets used in this study. The letters in the “Usage” column indicate the tasks for which the datasets were used. a: training auxiliary gender classifiers, b: SAN training, c: SAN evaluation, d: constructing unseen gender classifiers used for evaluating SAN models.

Dataset	#male	#female	Usage
CelebA-train	73,549	103,772	a, b
CelebA-test	7,929	11,511	c
MORPH-train	41,587	7,567	d
MORPH-test	4,643	863	c
LFW	10,064	2,905	d
MUCT	1,844	1,910	c
RaFD	1,008	600	c

test [122], MUCT [94], and RaFD [83]. The number of male and female individuals in each dataset is listed in Table 5.1.

5.3.1 Performance in Confounding Unseen Gender Classifiers

In order to evaluate the generalization performance of the three ensemble-based methods discussed in the previous section (Ens-Avg, Ens-Gibbs, Ens-Best) as well as the proposed FlowSAN model, we considered six independent gender classifiers. The experiments designed in this section assess how well the proposed models are able to confound gender classifiers that were unseen during training. These six gender classifiers include three models that were already trained: a commercial-of-the-shelf gender classifier (G-COTS), IntraFace [142], AFFACT [60], and three CNN models built in-house, which we refer to as CNN-1, CNN-2 (trained using MORPH-train and LFW, respectively), and CNN-3 (trained on the union of MORPH-train and LFW). Note that these three CNN models have shown a similar level of performance on the original test-sets, compared to the other three pre-trained gender predictors.

Fig. 5.6 shows the area under the ROC curve as a performance metric for evaluating the generalization performance of each unseen gender classifier on the four independent test datasets.

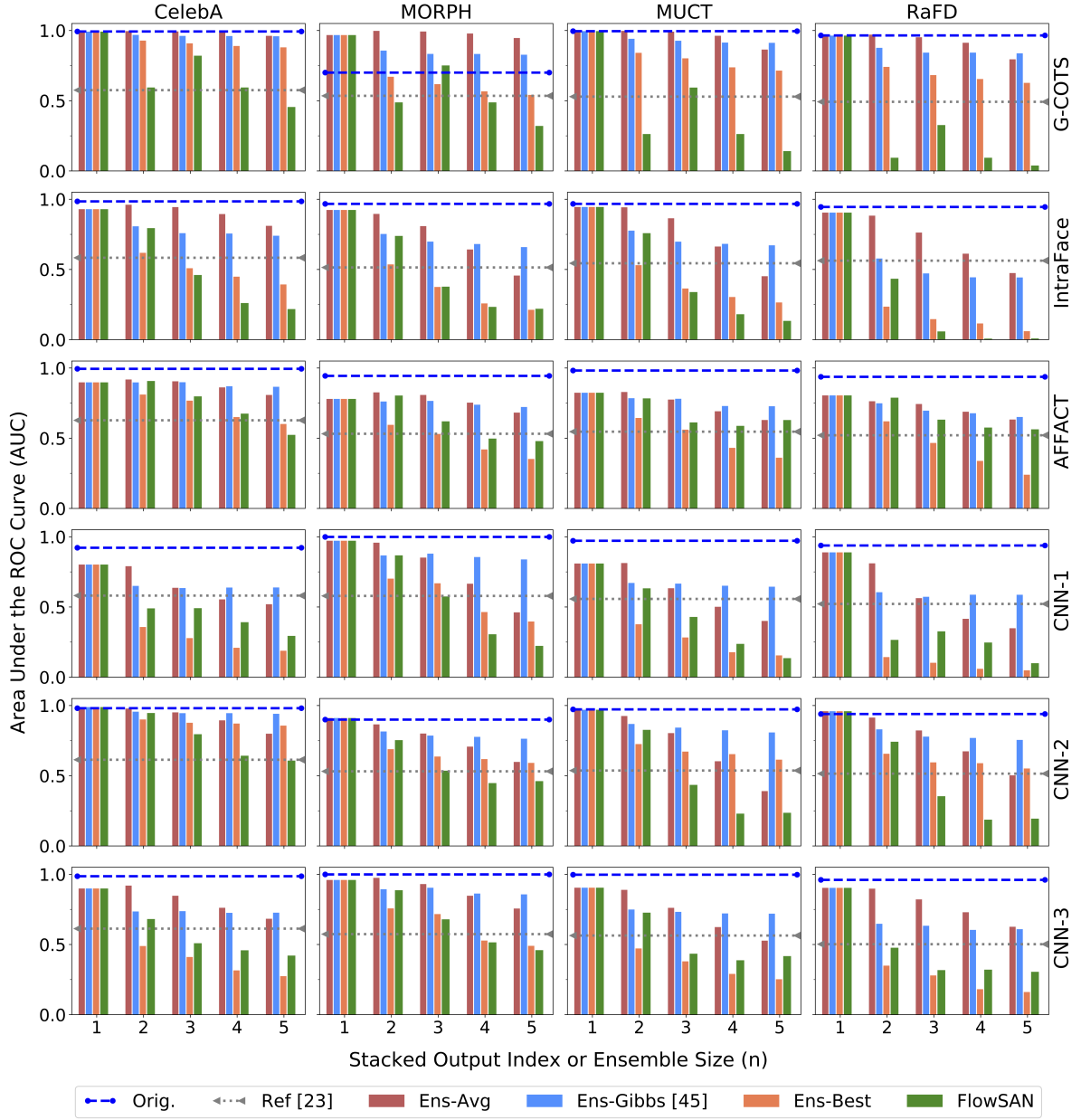


Figure 5.6: Area under the ROC curve (AUC) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images (“Orig.”) is shown (blue dashed line) as well as the perturbed samples using the three ensemble-based models (Ens-Avg, Ens-Gibbs, Ens-Best) the proposed FlowSAN model, and the face mixing approach [107] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, \dots, I'_5 \rangle$ obtained by varying the ensemble size, n . In almost all cases, stacking three SAN models results in an AUC of approximately 0.5 (a perfectly random gender prediction).

The performance of these gender classifiers on the original images (before perturbations), as well as the outputs from the mixing approach by [107], is also shown for comparison.

In all cases, the FlowSAN approach results in lower AUC values (lower is better) of predictions made by unseen gender classifiers (Fig. 5.6) compared to the ensemble models Ens-Avg and Ens-Gibbs. In fact, the results of the stacking SAN models are almost on par with the oracle best-perturbed samples (Ens-Best) for each gender classifier. In some cases, the FlowSAN model even outperforms Ens-Best. **It is important to note that selecting the best-perturbed sample (from the individual SAN models) for each gender classifier without *a priori* knowledge of the classifier is infeasible in practice. Yet, we are able to outperform the best result using the FlowSAN model in several cases.**

Note that in a real privacy application, reaching a near random gender prediction performance ($AUC \approx 0.5$, and Equal Error Rate (EER) ≈ 0.5) is desired for gender anonymization. As it can be seen in Fig. 5.6, both Ens-Avg and Ens-Gibbs methods produce samples that are mostly incapable of lowering the AUC of the unseen gender classifiers below 0.75 AUC. Based on the results shown in Fig. 5.6 (and the EER results shown in Fig. 5.9), it is evident that, in the majority of cases, a sequential stacking of three SAN models via FlowSAN produces the desired behavior in terms of face gender-anonymization, i.e., $AUC \approx 0.5$ (similarly, $EER \approx 0.5$). Although, in some cases, the 5th output from Ens-Avg and Ens-Gibbs resulted in a low, desired AUC of ≈ 0.5 , it also has a substantially detrimental effect on the face matching performance, as discussed in Section 5.3.2.

As a result, we conclude that stacking three SAN models in FlowSAN is sufficient to achieve the best gender label anonymization performance across a set of different, unseen gender classifiers and face image datasets. Stacking fewer than three models affects unseen gender classifiers substantially less, and stacking more than three models induces such strong perturbations that flipping the predicted labels could again de-anonymize the perturbed face images with respect to their

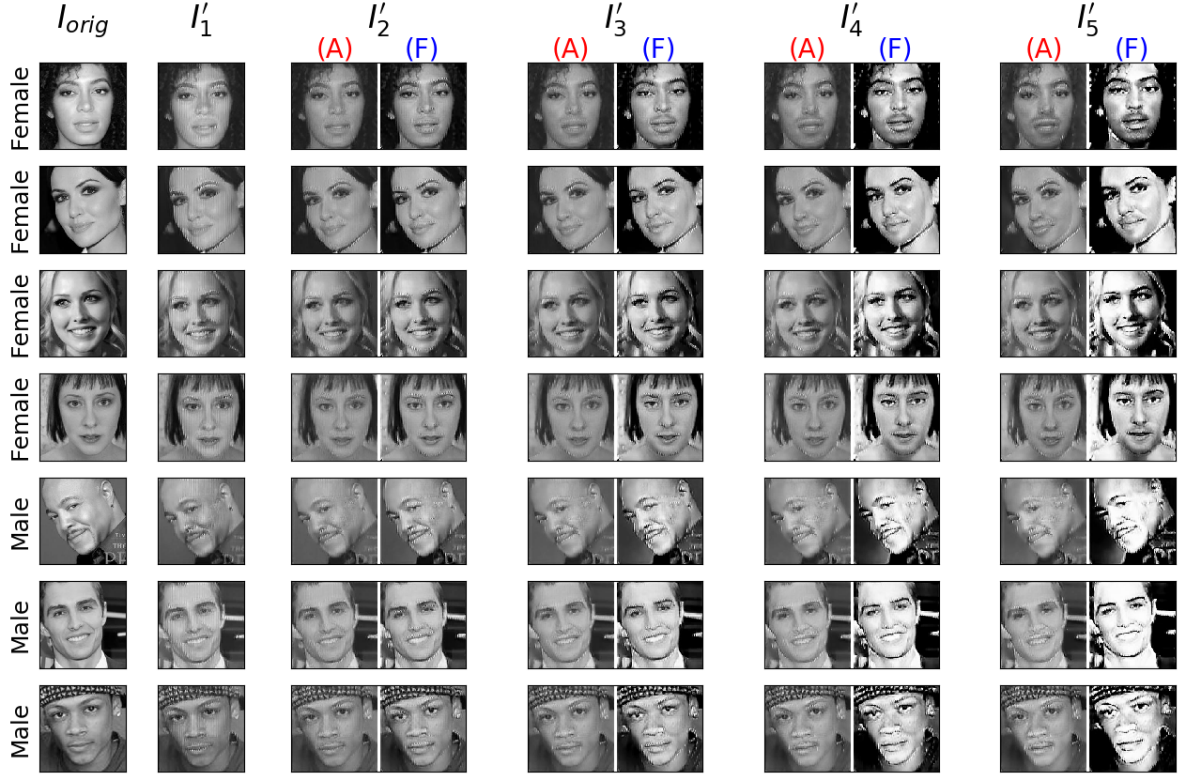


Figure 5.7: A randomly selected set of examples showing input face images and their outputs from I'_1 to I'_5 using (A) the ensemble model, Ens-Avg, and (F) using the FlowSAN model.

gender labels.

We shall note that our study was not the first to confound gender classifiers to produce random predictions. In [107], researchers proposed a face mixing approach that also leads to successful gender anonymization (approximately 0.5 AUC gender prediction performance for a specific gender classifier); however, this approach was unable to retain the face matching utility. In different studies, the researchers were able to retain face matching utility but without generalizing to arbitrary gender classifiers [100, 31]. Thus, the FlowSAN model we propose in this chapter presents the first successful approach for satisfying both objectives: concealing gender information and retaining matching performance to a satisfactory degree across a variety of independent gender classifiers and face matchers.

5.3.2 Retaining the Performance of Unseen Face Matchers

To assess the effect of the gender perturbations on the matching accuracy, we considered four different unseen face matchers. This includes a commercial-of-the-shelf face matcher (M-COTS), which has shown state-of-the-art performance in face recognition, as well as three publicly available algorithms that provide face representation vectors: DR-GAN [143], FaceNet [127], and OpenFace [10]. For the latter three models, we measured the cosine similarity between face representation vectors obtained from the original images and face representation vectors obtained from the SAN-perturbed output images.

Fig. 5.8 shows the True Match Rate (TMR) values at False Match Rate (FMR) of 0.1% for different ensemble methods. In most cases, the performance of the face matchers regarding the first three outputs (I'_1 , I'_2 , and I'_3) is similar and relatively close to the matching performance on original images. We note that stacking three SANs in FlowSAN yields the desired performance with regard to confounding unseen gender classifiers. Therefore, the evaluation of the face matching performance for stacking more than three SANs I'_3 (i.e., I'_4 and I'_5) is only included for completeness.

Comparing the performance of face matchers for equal values of n , we observe that the face matchers appear to perform slightly better on outputs produced by the ensemble model compared to the FlowSAN model. However, the extent to which the gender classification performance is reduced by the two models is not the same for equal values of n (Table 5.2). The ensemble model requires at least $n = 5$ individual SAN models to be able to confound unseen gender classifiers to reach the same level of gender anonymization as the FlowSAN model with $n = 3$. Therefore, if we compare the ensemble models with $n = 5$ to the FlowSAN model with $n = 3$, the face matchers perform substantially better on the face image outputs by the FlowSAN model (Fig. 5.8). Further, note that the performance of M-COTS on CelebA on the original images is already as low as

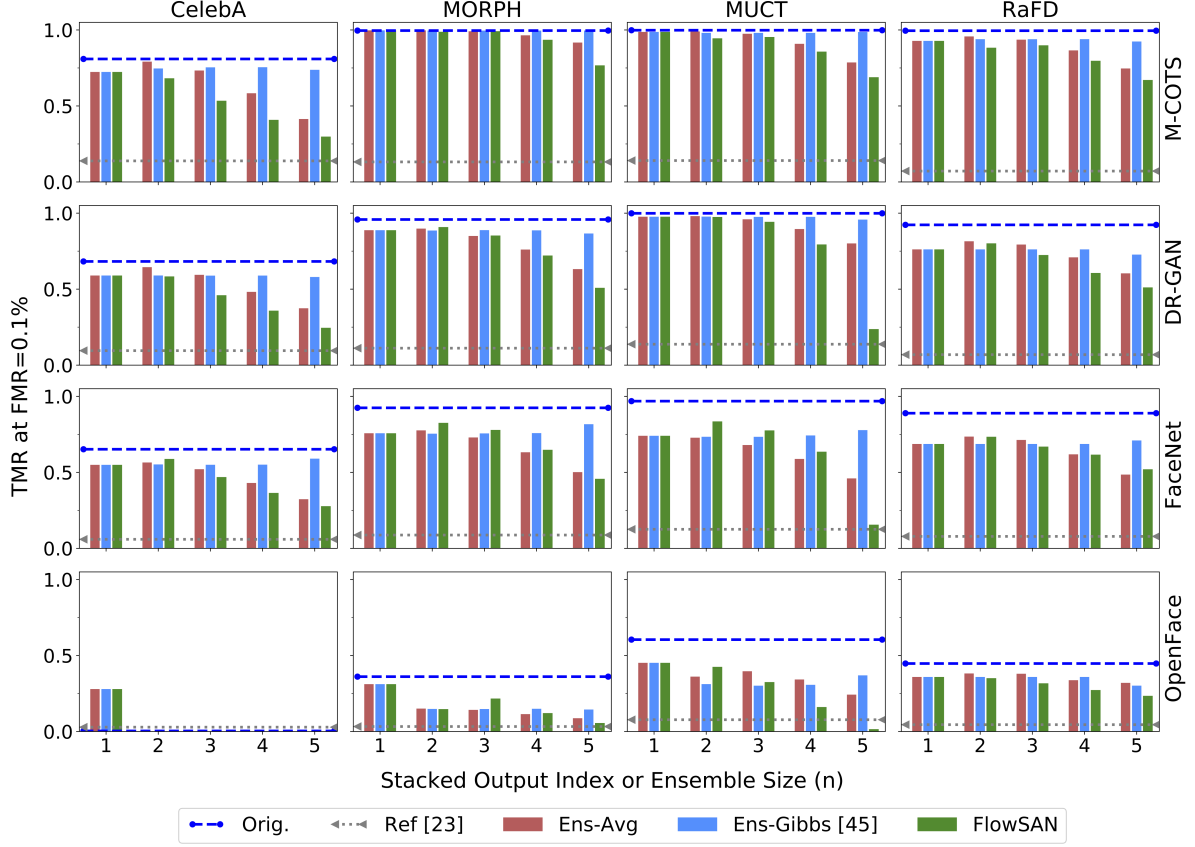


Figure 5.8: True Match Rate (TMR) values at False Match Rate (FMR) of 0.1% obtained from four unseen face matchers, M-COTS, DR-GAN, FaceNet, and OpenFace on the original images as well as perturbed outputs after applying stacking SAN models and the ensemble models (Ens-Avg and Ens-Gibbs). Note that the matchers' performance obtained after applying the first three SANs in the FlowSAN model is close to the original performance, but it further diminishes when the sequence is extended.

85.6%. In fact, all matchers perform poorly on the CelebA dataset, which may be due to different face orientations captured in the wild.

5.3.3 Preserving Privacy

The overall average performance considering the two target objectives of this study, i.e., confounding gender classifiers and retaining the matching utility of face images, is provided in Table 5.2. In this analysis, the average EER results of all six gender classifiers over all four evaluation datasets

Table 5.2: Comparing the overall average performance of six unseen gender classifiers and four unseen face matchers over the four evaluation datasets using $n = 3$ or $n = 5$ SAN models. This shows that stacking 3 SAN models results in gender anonymization $EER \approx 0.5$, while the the average matching performance is still comparable to the unmodified images as well as the matching performance on the outputs form other existing methods.

	Gender: EER		Matching: TMR at FMR=0.1%	
Orig.	10%		76.3%	
Ref [107]	46%		9.1%	
	$n = 3$	$n = 5$	$n = 3$	$n = 5$
Ens-Avg	23%	40%	64.9%	48.1%
Ens-Gibbs	29%	31%	65.2%	65.6%
Ens-Best	48%	57%	–	–
FlowSAN	49%	64%	61.9%	35.4%

were computed for original images, outputs from Ref. [107], as well as outputs from the stacking and the ensemble models using $n = 3$ and $n = 5$. The results clearly show that the FlowSAN model outperforms the ensemble-based methods, including the oracle-best results. On the other hand, the average true matching rate (TMR) values, at a false matching rate (FMR) of 0.1%, are also computed similarly, and the results indicate that the Ens-Gibbs method has the highest performance for both ensemble sizes, while the performance of the FlowSAN model at $n = 3$ is ranked as second, but it is very close to that of Ens-Gibbs. The detailed EER results for each gender classifier is provided in Table 5.3.

5.3.4 Computational Efficiency

The overall computational cost for training the ensemble-based approach and the FlowSAN model is similar, except that FlowSAN requires an additional data transformation step between each consecutive SAN training. However, the ensemble approach comes with a bigger advantage that the

individual SAN models can be trained in parallel, while the SAN models in the FlowSAN model have to be trained sequentially.

5.4 Summary and Future Work

In this work, we address one of the main limitations of previous gender privacy methods, namely, their inability to generalize across multiple previously unseen gender classifiers. In this regard, we propose the FlowSAN method that sequentially combines diverse perturbations for an input face image to confound the gender information with respect to an arbitrary gender classifier. We compared the performance of the proposed FlowSAN model with two ensemble-based approaches: 1) using the average output of SAN models trained independent of each other (Ens-Avg); 2) randomly selecting the output from the SAN models in the ensemble (Ens-Gibbs).

Our experiments show that the FlowSAN method outperforms the other ensemble-based approaches in terms of confounding gender attribute for a range of gender classifiers. More importantly, while gender classification is successfully confounded, face matching accuracy is retained for all perturbed output face images, thereby preserving the biometric utility of the gender-anonymous face images.

For future work, we will extend the proposed privacy-preserving scheme to multiple demographic attributes including age and race, and design a SAN model that can confound a selected combination of attributes while preserving matching performance. This is expected to enhance the privacy of individuals whose biometric data is stored in central databases.

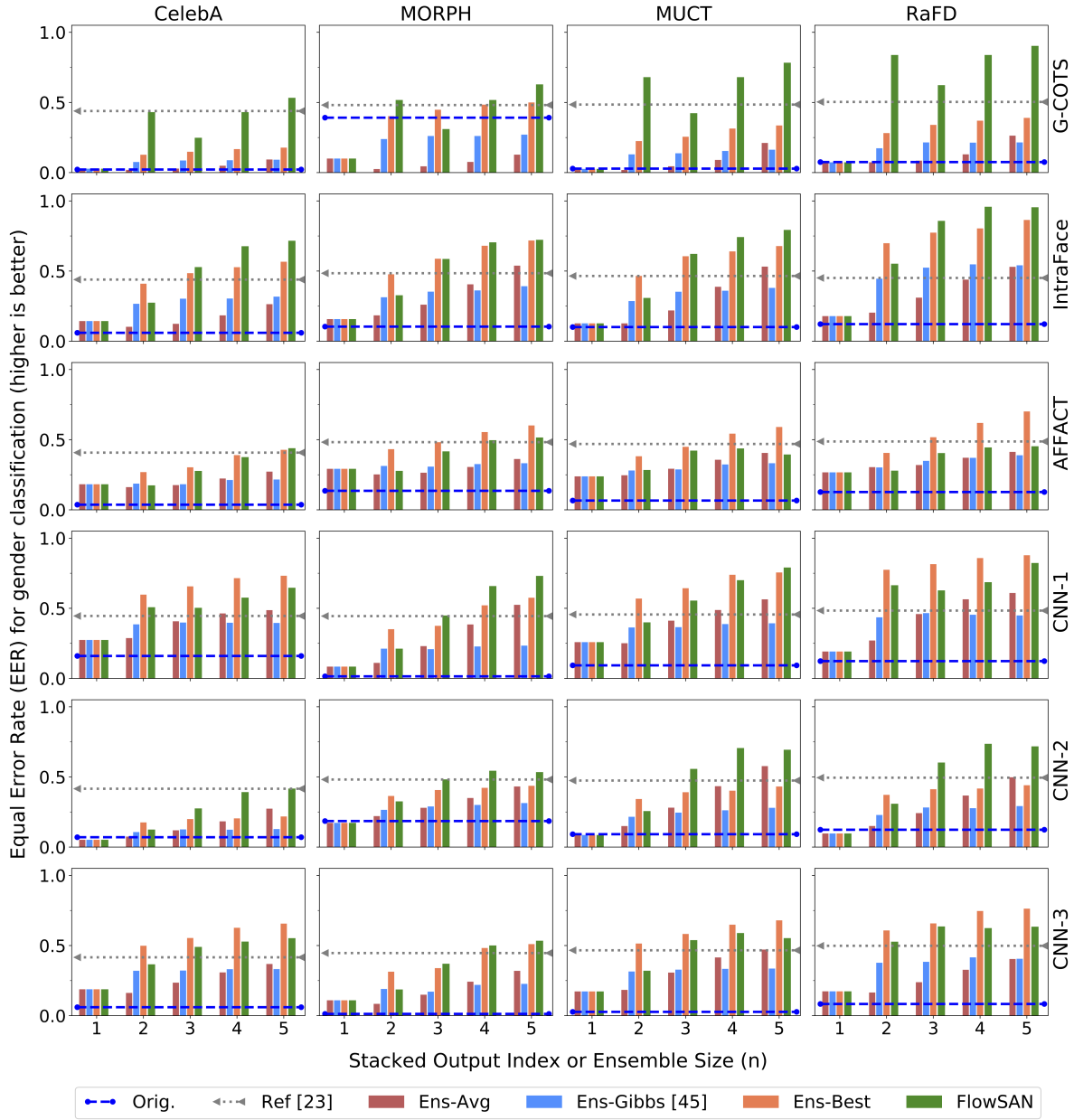


Figure 5.9: Equal Error Rate (EER) measured for the six unseen gender classifiers (CNN-3, CNN-2, CNN-1, AFFACT, IntraFace, and G-COTS) on the test partitions of the four different datasets (CelebA, MORPH, MUCT, and RaFD). The gender classification performance on the original images ("Orig.") is shown (blue dashed line) as well as the perturbed samples using the three ensemble models (Ens-Avg, Ens-Gibbs, Ens-Best), the proposed FlowSAN model, and the face mixing approach [107] (gray dashed line). The index (1, 2, ..., 5) on the x-axis indicates the sequence of outputs $\langle I'_1, I'_2, \dots, I'_5 \rangle$ obtained by varying the ensemble size, n .

Table 5.3: Comparing the overall average Equal Error Rate (EER) of six unseen gender classifiers averaged over all four evaluation datasets (CelebA-test, MORPH-test, MUCT, and RaFD), higher is better. Note that the Ens-Best method is the result of “oracle best” selected classifier from an ensemble of multiple SANs, which assumes knowledge of the gender classifier. While this is impractical in a real-world privacy application, we show the results for comparison purposes.

Part-A: ($n = 3$)						
Gender Classifier	Orig.	Ref. [107]	Ens-Avg	Ens-Gibbs	Ens-Best	FlowSAN
G-COTS	0.13	0.48	0.05	0.18	0.30	0.40
IntraFace	0.10	0.46	0.23	0.38	0.61	0.65
AFFACT	0.09	0.46	0.26	0.28	0.44	0.38
CNN-1	0.10	0.46	0.38	0.36	0.62	0.53
CNN-2	0.12	0.47	0.23	0.23	0.35	0.48
CNN-3	0.05	0.46	0.23	0.30	0.53	0.51
Average	0.10	0.46	0.23	0.29	0.48	0.49

Part-B: ($n = 5$)						
G-COTS	0.13	0.48	0.17	0.18	0.35	0.71
IntraFace	0.10	0.46	0.47	0.41	0.71	0.80
AFFACT	0.09	0.46	0.36	0.32	0.58	0.45
CNN-1	0.10	0.46	0.55	0.38	0.74	0.75
CNN-2	0.12	0.47	0.45	0.25	0.38	0.59
CNN-3	0.05	0.46	0.39	0.32	0.65	0.57
Average	0.10	0.46	0.40	0.31	0.57	0.64

Chapter 6

PrivacyNet: Semi-Adversarial Networks for Multiattribute Face Privacy

Portions of this chapter have been published in:

- V. Mirjalili, S. Raschka, A. Ross, "PrivacyNet: Semi-adversarial networks for multi-attribute face privacy", IEEE Transactions on Image Processing, Vol. 29, pp. 9400-9412, 2020.

6.1 Introduction

So far, we developed Semi-Adversarial Networks (SAN) which is a deep-learning model for imparting gender privacy to face images. While the original SAN model has successfully been shown to conceal gender attributes from face images while being able to retain satisfactory face matching accuracy, it did not apply to a broader range of soft biometric characteristics.

In this chapter, we propose a new method for imparting multi-attribute privacy to face images including age, gender, and race, which we refer to as *PrivacyNet*. Our overall objective is shown in Fig. 6.1, where PrivacyNet can perturb the soft biometric information contained in an input face image across three orthogonal axes corresponding to age, gender, and race. While soft biometric information can be successfully concealed, the matching utility of transformed faces is preserved.

In previous chapters, we developed a deep learning model to generate perturbed examples for

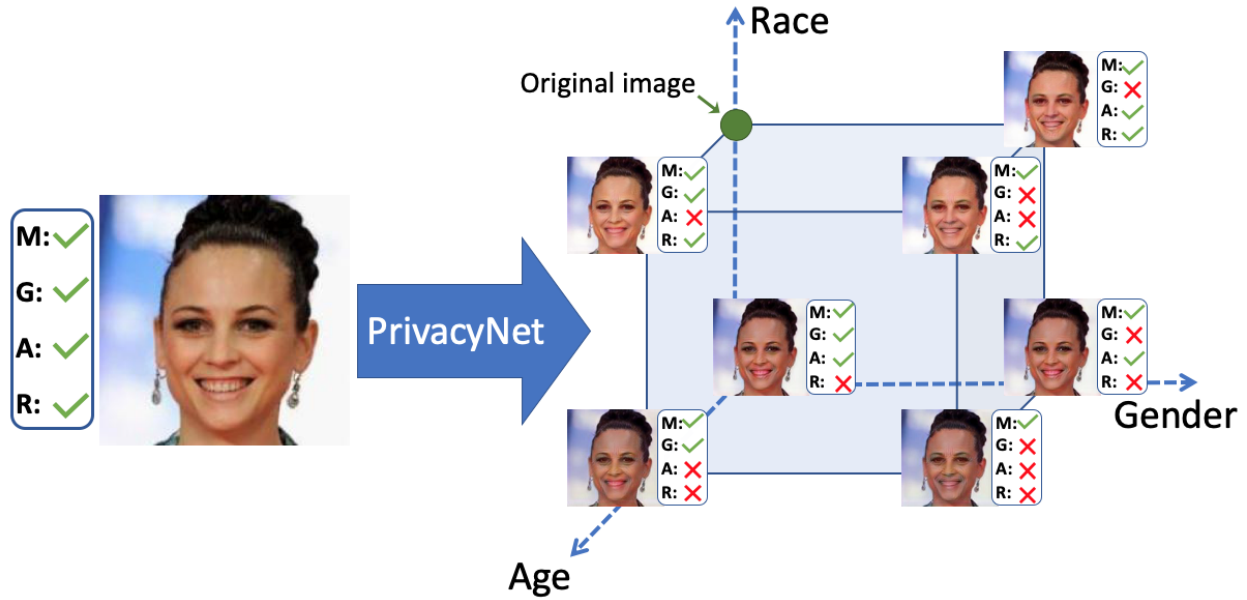


Figure 6.1: The overall idea of this work: transforming an input face image across three dimensions for imparting multi-attribute privacy selectively while retaining recognition utility. The abbreviated letters are M: Matching, G: Gender, A: Age and R:Race.

confounding gender information in face images [101]. The neural network was coined Semi-Adversarial Network (SAN) and is composed of a convolutional autoencoder for synthesizing face images such that the gender information in the synthesized images is confounded while their matching utility is preserved. The SAN model is trained using an auxiliary gender classifier and an auxiliary face matcher. After training, the auxiliary subnetworks are discarded and the convolutional autoencoder is used for performance evaluation, and it was shown that this model is able to confound gender information as assessed by some unseen¹ attribute classifiers while the matching utility, assessed by unseen face matchers, was retained. Furthermore, using an ensemble of SAN models, it was empirically shown that the face perturbations for concealing soft-biometric information generalize to arbitrary unseen gender classifiers [99]. Most recently, imparting privacy to face representation vectors have also emerged, where extracting sensitive information from the face

¹In contrary to “auxiliary” classifiers, the term “unseen” indicates that the classifier (or face matcher) was not used during the training stage.

representation vectors is confounded [139, 103]. SensitiveNet [103] was also proposed where the proposed model generates agnostic face representations for face recognition such that the sensitive information including gender and race are removed from these representations [71].

Here, we present a new model, PrivacyNet, for enabling multi-attribute privacy. Existing techniques including controllable face privacy [131] work on well-posed cropped face images and the face area needs to be cropped. Furthermore, confounding multiple attributes requires adding a sequence of perturbations which can potentially result in further drift in matching performance. Perturbation-based method proposed by [31] have limited applicability to real-world privacy problems since their perturbed faces do not generalize to unseen attribute classifiers, and can only confound the specific classifiers where the perturbations are derived for. On a different avenue, Generative Adversarial Networks (GAN) [53, 88] and its variants have shown remarkable performance in many computer vision tasks such as image-to-image translation, and face image synthesis [33, 155, 76, 69, 11]. However, GAN models are not considered a viable solution for imparting soft-biometric privacy since GANs are trained to generate realistic face images from random samples of an arbitrary distribution rather than altering specific input images. Hence, a face image generated by a GAN is not specifically related to a given input image but rather the distribution of all training set images. Therefore, as the previously proposed SAN model has shown to successfully overcome the limitations of existing techniques in confounding unseen attribute classifiers, while maintaining the recognition capability, in this work we design PrivacyNet, a multi-attribute privacy model empowered with GAN for confounding age, race and gender in a controllable fashion, where the users or the system can decide what attributes to flip and what to conceal.

In summary, the contributions of this work is the following

- the design of the multi-attribute PrivacyNet model to provide controllable soft-biometric privacy including gender, age, and race;

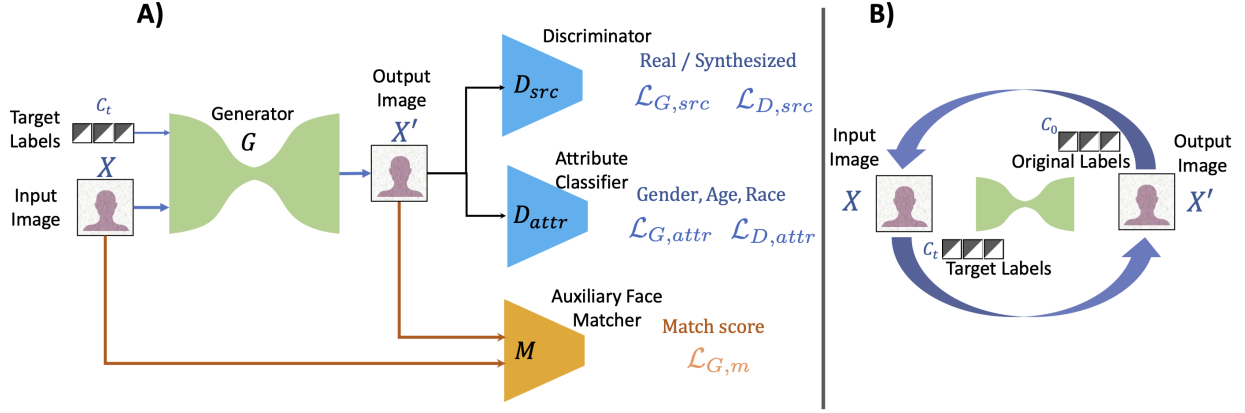


Figure 6.2: Schematic representation of the architecture of PrivacyNet for deriving perturbations to obfuscate three attribute classifiers – gender, age and race – while allowing biometric face matchers to perform well. (A) Different components of the PrivacyNet: generator, source discriminator, attribute classifier, and auxiliary face matcher. (B) Cycle-consistency constraint applied to the generator by transforming an input face image to a target label and reconstructing the original version.

- a solution for making GAN models useful in biometric applications by ensuring that generated face images still match with their original counterpart;
- performance assessments using *unseen* attribute classifiers which provide empirical evidence for the efficacy of the proposed model in imparting soft-biometric attribute privacy to face images.

6.2 Proposed method

6.2.1 Problem Formulation

Given a face image X , let S_{obf} be a set of face attributes to be obfuscated and S_{keep} be a set of attributes to be preserved. The overall objective is to find function ϕ that applies some perturbations to the input image X such that $X' = \phi(X)$ has the following properties:

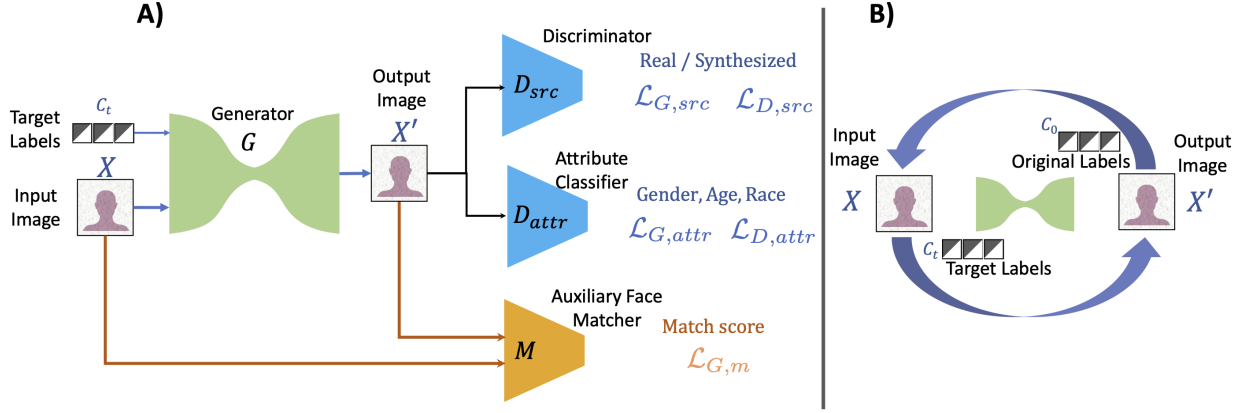


Figure 6.3: Schematic representation of the architecture of PrivacyNet for deriving perturbations to obfuscate three attribute classifiers – gender, age and race – while allowing biometric face matchers to perform well. (A) Different components of the PrivacyNet: generator, source discriminator, attribute classifier, and auxiliary face matcher. (B) Cycle-consistency constraint applied to the generator by transforming an input face image to a target label and reconstructing the original version.

- For a soft biometric attribute $a \in S_{\text{obf}}$, the performance of an unseen attribute classifier f_a is substantially reduced.
- For the remaining set of attributes $b \in S_{\text{keep}}$, the performance of an arbitrary classifier f_b is not noticeably adversely affected; that is, the performance of an attribute classifier f_b on perturbed image X' is close to its performance on the original face image X .
- The primary biometric utility, which is face recognition, must be retained for the modified face image, X' . In other words, given pairs of image examples before ($\langle X_1, X_2 \rangle$) and after ($\langle X'_1, X'_2 \rangle$) perturbations, the matching performance as assessed by an arbitrary face matcher (f_M) is not substantially affected, i.e.,

$$f_M(X_1, X_2) \approx f_M(X'_1, X'_2) \approx f_M(X_1, X'_2) \approx f_M(X'_1, X_2).$$

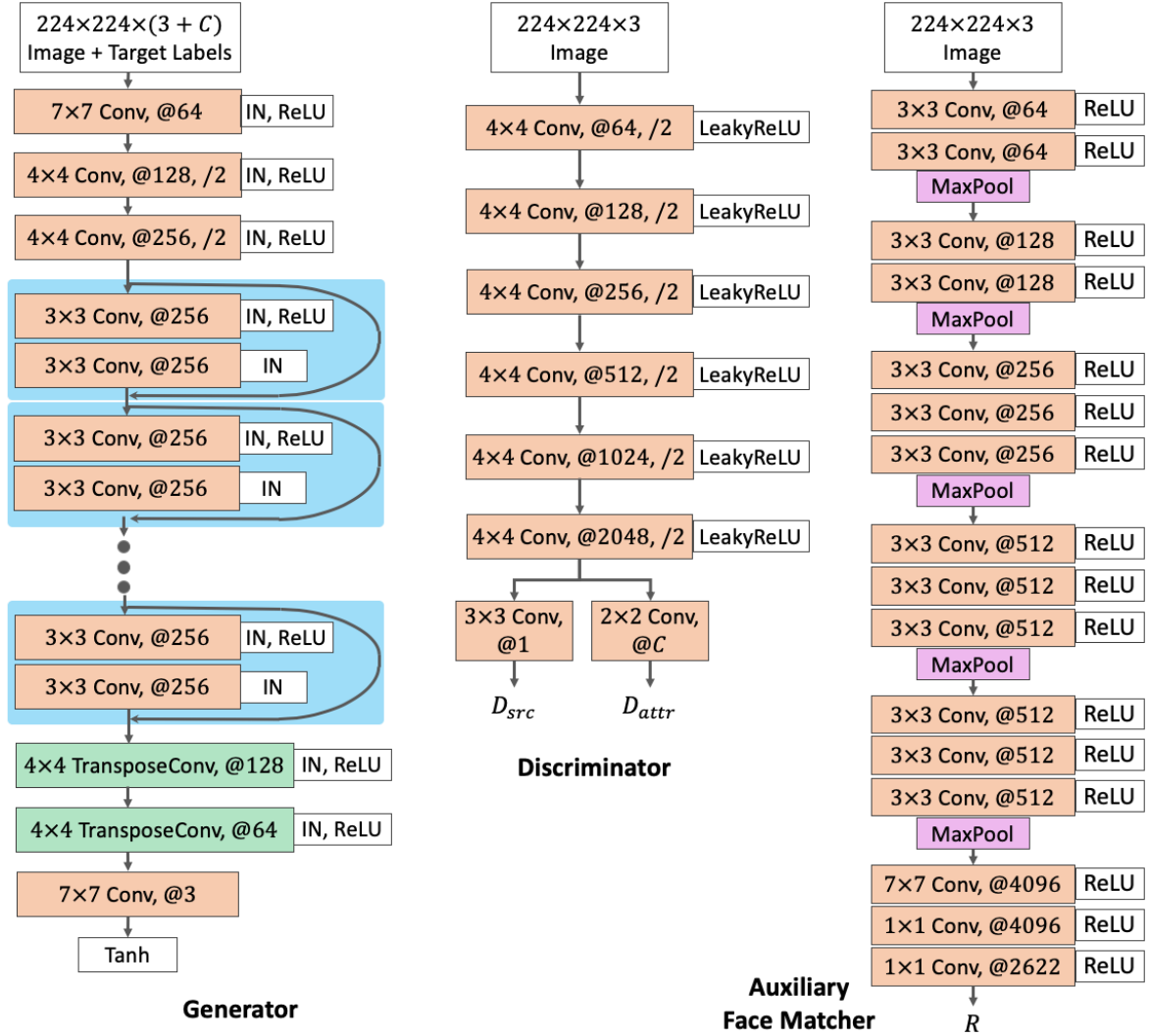


Figure 6.4: The detailed neural network architecture of the four sub-networks of PrivacyNet: the generator G , the discriminators D_{src} and D_{attr} , and the pre-trained auxiliary face matcher M . Note that D_{src} and D_{attr} share the same convolutional layers and only differ in their respective output layers.

6.2.2 PrivacyNet

According to the objectives described in Section 6.2.1, the PrivacyNet neural network architecture (Fig. 6.3A) is composed of four sub-networks: A generator (G) that modifies the input image, a source discriminator (D_{src}) which determines if an image is real or modified, an attribute classifier (D_{attr}) for predicting facial attributes, and an auxiliary face matcher (M) for biometric face recognition. Along with the input image, both generator and discriminator receive the attribute labels as conditional variables, that are spanned to the same width and height as the input image, (224×224). Together, these subnetworks form a cycle-consistent GAN [155] as illustrated in Fig. 6.3B. Given an RGB input face image X , the attribute label vector $\mathcal{V}_0 \in \mathbb{Z}^c$ corresponds to the ground truth attribute labels of the original face image. The target label vector $\mathcal{V}_t \in \mathbb{Z}^c$ (c is the total number of attributes) denotes the desired facial attributes for modifying the face image. Given a target vector $\mathcal{V}_t \neq \mathcal{V}_0$, the objective of the generator G is to synthesize a new image $X' = G(X, \mathcal{V}_t)$ such that X' is mapped to the target label vector \mathcal{V}_t by an attribute classifier D_{attr} . The other component of the GAN model is a source discriminator D_{src} , which is trained to distinguish real images from those synthesized by the generator.

The total loss terms for training the discriminator ($\mathcal{L}_{D,tot}$) and the generator ($\mathcal{L}_{G,tot}$) are as follows:

$$\mathcal{L}_{D,tot} = \mathcal{L}_{D,src} + \lambda_{D,attr} \mathcal{L}_{D,attr}, \quad (6.1)$$

and

$$\begin{aligned} \mathcal{L}_{G,tot} = & \mathcal{L}_{G,src} + \lambda_{G,attr} \mathcal{L}_{G,attr} + \\ & \lambda_m \mathcal{L}_{G,m} + \lambda_{rec} \mathcal{L}_{G,rec}, \end{aligned} \quad (6.2)$$

where, λ coefficients are hyperparameters representing the relative weights for the corresponding loss terms. The individual terms of the total loss for the discriminator ($\mathcal{L}_{D,tot}$) and the generator ($\mathcal{L}_{G,tot}$) are described in the following paragraphs.

For the discriminator, the loss term associated with source discrimination (i.e., discriminating between real and synthesized images) is defined as

$$\begin{aligned} \mathcal{L}_{D,src} = & \mathbb{E}_{X,\mathcal{V}_0} \left[-\log (D_{src}(X, \mathcal{V}_0)) \right] + \\ & \mathbb{E}_{X,\mathcal{V}_t} \left[-\log (1 - D_{src}(G(X, \mathcal{V}_t), \mathcal{V}_t)) \right], \end{aligned} \quad (6.3)$$

where, $\mathbb{E}_{X,\mathcal{V}}[f(X, \mathcal{V})]$ represents the expected value of the random variable $f(X, \mathcal{V})$ taken over distribution of X given the conditional variable \mathcal{V} . Similarly, the loss associated with the source discrimination for the generator subnetwork is defined as

$$\mathcal{L}_{G,src} = \mathbb{E}_{X,\mathcal{V}_t} \left[\log(1 - D_{src}(G(X, \mathcal{V}_t), \mathcal{V}_t)) \right], \quad (6.4)$$

where, $D_{src}(X)$ returns the estimate of the probability that the input image X is real or was synthesized by the generator.

Next, the loss terms for attribute classification are defined as

$$\mathcal{L}_{D,attr} = \mathbb{E}_{X,\mathcal{V}_0} \left[-\log (D_{attr}(\mathcal{V}_0|X)) \right] \quad (6.5)$$

and

$$\mathcal{L}_{G,attr} = \mathbb{E}_{X,\mathcal{V}_t} \left[-\log (D_{attr}(\mathcal{V}_t|G(X, \mathcal{V}_t))) \right], \quad (6.6)$$

where, $D_{attr}(\mathcal{V}|X)$ is the probability that input image X belongs to attribute class \mathcal{V} .

The loss term for optimizing the performance of the biometric face matcher \mathcal{M} on the perturbed images is defined as the squared L_2 distance between the normalized features of the original face image X and those of the synthesized image $G(X, \mathcal{V}_t)$:

$$\mathcal{L}_{G,m} = \mathbb{E}_{X, \mathcal{V}_t} \left[\|R_{\mathcal{M}}(X) - R_{\mathcal{M}}(G(X, \mathcal{V}_t))\|_2^2 \right], \quad (6.7)$$

where, $R_{\mathcal{M}}(X)$ is the normalized face descriptor of face image X after applying a face matcher \mathcal{M} .

Lastly, a reconstruction loss term is used to form a cycle-consistent GAN that is able to reconstruct the original face image X from its modified face image $X' = G(X, \mathcal{V}_t)$:

$$\mathcal{L}_{G,rec} = \mathbb{E}_{X, \mathcal{V}_0, \mathcal{V}_t} \left[\|X - G(G(X, \mathcal{V}_t), \mathcal{V}_0)\|_1 \right]. \quad (6.8)$$

Note that the distance term in Eq. 6.8 is computed as the pixel-wise L_1 norm between the original and modified images, which empirically results in less blurry images compared to employing a L_2 norm as the distance measure [69].

6.2.3 Neural Network Architecture of PrivacyNet

The composition of the different neural networks used in PrivacyNet, generator G , real vs. synthetic classifier D_{src} , attribute classifier D_{attr} , and face matcher $R_{\mathcal{M}}$ is described in Fig. 6.4. The generator and the discriminator architectures were adapted from [33] and [155], respectively.

Generator. The generator G receives as input an RGB face image X of size $224 \times 224 \times 3$ along with the target labels \mathcal{V}_t concatenated as extra channels. The first two convolutional layers, with stride 2, reduce the size of the input image to a 32×32 with 128 channels. The convolutional

layers are followed by instance normalization layers (InstanceNorm) [145]. The layer activations are computed by applying the non-linear ReLU activation function to the InstanceNorm outputs. Then, 6 residual blocks [64] are applied, followed by two transposed convolution for upsampling the image size to 224×224 . Finally, the output image X' is constructed by a 1×1 convolution layer and the hyperbolic tangent ($Tanh$) activation function, which returns pixels in the range $(-1, 1)$ (the input image pixels are also scaled to be in range $[-1, 1]$).

Discriminator and Attribute Classifier. The discriminator, as shown in Fig. 6.4, combines the source discriminator D_{src} and the attribute classifier D_{attr} into one network where all the layers except the last convolution layer are shared among the two tasks. All the shared convolution layers are followed by a Leaky ReLU non-linear activation with a small negative slope of $\alpha = 0.01$. In the last layer, separate convolutional layers are used for the two tasks, where D_{src} returns a scalar score for computing the loss according to Wasserstein GAN [12], and D_{attr} returns a vector of probabilities for each attribute class.

Face Matcher. Lastly, the auxiliary face matcher is adapted from the publicly available pre-trained VGG-Face CNN model that receives input face images of size $224 \times 224 \times 3$ and computes their face descriptors of size 2622 [109].

6.2.4 Datasets

We have used five datasets in this study: CelebA [86], MORPH [122], MUCT [94], RaFD [83], and UTK-face [154]. Table 6.1 shows the number of examples in each dataset, including the number of examples for each face attribute. Since the race label distribution in CelebA is heavily skewed towards Caucasians, and MORPH is heavily skewed towards persons with African ancestry, we combined CelebA and MORPH for training. Both the CelebA and MORPH datasets are split into training and evaluation sets in a subject-disjoint manner. The two training subsets from CelebA

Table 6.1: Overview of datasets used in this study, with the number of face images corresponding to each attribute. Samples which belong to a race other than the two categories shown below, as well as those whose age-group could not be determined, are omitted.

Dataset	Gender		Race		Age groups		
	Male	Female	African-descent	Caucasian	Young	Midle-aged	Old
CelebA	84,434	118,165	11,119	142,225	79,848	91,373	16,337
MORPH	47,057	8,551	42,897	10,736	25,009	26,614	3,985
MUCT	1,844	1,910	1,030	1,480	1,326	1,807	620
RaFD	1,008	600	0	1,608	1,276	332	0
UTK-face	12,582	11,522	4,558	10,222	12,980	6,068	5,056

and MORPH are merged to train the PrivacyNet model with a relatively balanced race distribution. The other three datasets, MUCT, RaFD, and UTK-face are used only for evaluation. While all five datasets provide binary attribute gender labels ², each dataset lacks the ground-truth labels for at least one of the other attributes, age or race.

Gender Attribute: All the five datasets considered in this study provide ground-truth labels for the gender attribute. Furthermore, since gender is a well-studied topic, there are several face-based gender predictors available for evaluation. In this study, we have considered three gender classifiers for evaluation: a commercial-off-the-shelf software G-COTS, IntraFace [142], and AFFACT [60].

Race Labels: We consider binary labels for race: Caucasians and African descent. Samples that do not belong to these two race groups are omitted from our study since the other race groups are under-represented in our training datasets. We have used the ground-truth labels provided in the MORPH and UTK-face datasets, but for the other three datasets, we labeled the samples in multiple stages. First, an initial estimate of the race attribute is computed using commercial software R-COTS. Next, the predictions made by R-COTS from all samples of the same subject are aggregated, and subjects that show discrepant predictions for different samples are visualized

²In this paper we treat gender as a binary attribute with two labels, male and female; however, it must be noted that societal and personal interpretation of gender can result in many more classes.

and the discrepant labels are manually corrected. Finally, one random sample from every subject is visually inspected to verify the predicted label. Furthermore, note that since RaFD did not have any sample from the African-descent race group, we did not use this dataset for race prediction analysis.

Age Information: The ground-truth age information is only provided in the MORPH and UTK-face datasets. Therefore, for the remaining datasets (CelebA, MUCT, and RaFD) we used the commercial-off-the-shelf A-COTS software to obtain the class labels of the original images. For the evaluation of our proposed model, we use the Mean Absolute Error (MAE) metric to measure the change in predicted age on the output images of PrivacyNet from the predicted age on the original face images. Therefore, the combination of all five datasets shows both changes in age prediction with respect to the original (for CelebA, MUCT, and RaFD) as well as the ground-truth age values (for MORPH and UTK-face datasets). For training the PrivacyNet model, we create three age groups based on the age values:

$$y_{\text{age}} = \begin{cases} 0 & \text{age} \leq 30; \\ 1 & 30 < \text{age} \leq 45; \\ 2 & 45 < \text{age}. \end{cases} \quad (6.9)$$

Due to the non-stationary nature of patterns in face aging [30, 106], creating age groups does not fully capture the non-linearity in the textural changes. However, this scheme is consistent with the treatment of the other two attributes, gender and age. Further, it should be emphasized that our objective is *not* to synthesize face images in particular age groups (which is known as age synthesis); instead, the goal of the proposed method is to disturb the performance of arbitrary age predictors.

Identity Information: For matching analysis, we exclude the UTK-face dataset since the sub-

Table 6.2: Summary of the datasets used in this study, with the number of subjects and samples in the train-test partitions. The “Excluded Experiments” column indicate datasets that were removed from an experiment for the reasons given in the text.

Datasets	Train		Test		Excluded Experiments
	# Subj	# Samples	# Subj	# Samples	
CelebA	8,604	150,530	167	2,795	–
MORPH	11,176	45,512	1,968	8,038	–
MUCT	–	–	185	2,508	–
RaFD	–	–	67	1,608	Race
UTK-face	–	–	NA	14,182	Matching

ject information is not provided. We used three face matchers, a commercial-off-the-shelf software M-COTS, and two publicly available face matchers DR-GAN [143] and SE-ResNet-50 [67] (SE-Net for short) which were trained on the VGGFace2 dataset [23].

A summary of the datasets and the number of subjects and samples in each dataset is provided in Table 6.2.

6.3 Experimental Results

The proposed PrivacyNet model is trained on the joint training subsets of CelebA and MORPH as explained in Section 6.2.4. Due to the memory-intensive training process, we used a batch-size of 16. The models were trained for 200,000 iterations. The optimal hyperparameter settings for the weighting coefficients of the attribute loss terms were $\lambda_{attr,d} = 1$ and $\lambda_{attr,d} = 4$. The matching term coefficient was set to $\lambda_m = 4$, and the hyperparameter for the reconstruction term was set to $\lambda_{rec} = 4$. After training the PrivacyNet model, both the discriminator and the auxiliary face matcher subnetworks are discarded and only the generator is used for transforming the unseen face images in the evaluation datasets.

Additionally, we also trained a cycle-GAN model [33], without the auxiliary face matcher, as a baseline to study the effects of the face matcher. The cycle-GAN model is trained using the same

protocol that was described for training PrivacyNet. In the remainder of this paper, we will refer to this method as “baseline-GAN”. The transformations of five different example images from the CelebA-test dataset are shown in Fig. 6.5.

The following subsections summarize the results of the experiments and analyze how the performance of the attribute classifiers and face matchers is affected by the face attribute perturbations via PrivacyNet.

6.3.1 Perturbing Facial Attributes

The performance assessment of the proposed PrivacyNet model involves three objectives:

1. when an attribute is selected to be perturbed, the performance of unseen attribute classifiers must decrease;
2. the attribute classifiers should retain their performance on attributes that are not selected for perturbation;
3. in all cases, the performance of unseen face matchers must not be drastically affected.

We conducted several experiments to assess whether the proposed PrivacyNet model meets these objectives.

Gender Classification Performance: We considered three gender classifiers: a commercial-off-the-shelf software (G-COTS), AFFACT [60] and IntraFace [142]. For this comparison study, all five evaluation datasets listed in Table 6.2 were considered. The performances of the different gender classifiers on the original and perturbed images are measured using the Equal Error Rate (EER); the results are shown in Fig. 6.6. For a given image, PrivacyNet can produce up to 15 distinct outputs, depending on the combination of attributes that are selected for perturbation.



Figure 6.5: Five example face images from the CelebA dataset along with their transformed versions using PrivacyNet and baseline-GAN models. The rows are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is specified as A0 (young), A1 (middle-aged), or A2 (old).

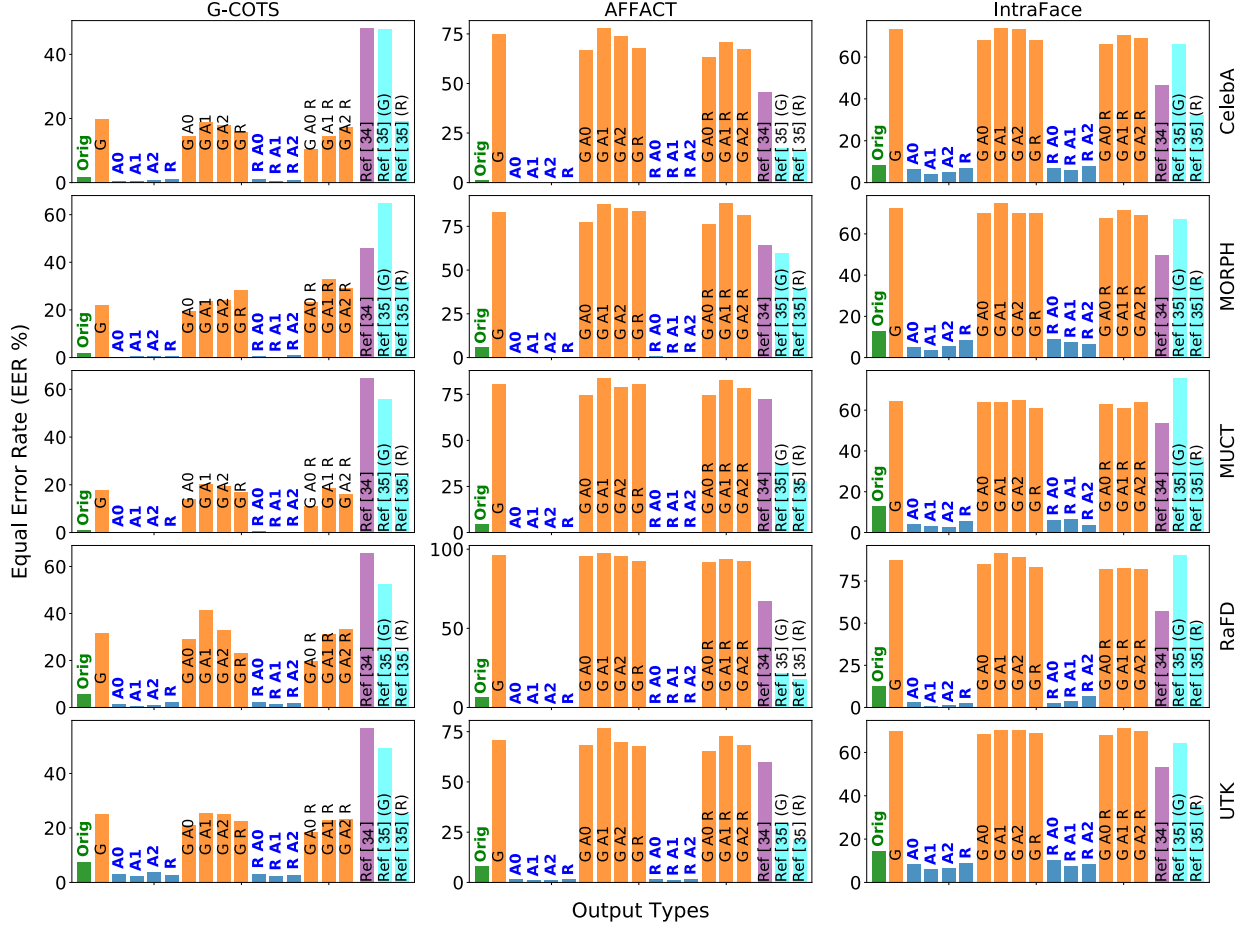


Figure 6.6: Performance of three gender classifiers – G-COTS, AFFACT, and IntraFace – on original images as well as different outputs of the proposed model (the larger the difference the better). The results of a face mixing approach, as described in [107], are also shown. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is abbreviated as A0 (young), A1 (middle-aged), and A2 (old). The outputs of PrivacyNet, where the gender attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.

The EER results shown in Fig. 6.6 indicate that PrivacyNet increases the error rate of the cases where the gender attribute is willfully perturbed, which is desired. At the same time, it can preserve the performance of gender classifiers when gender information is to be retained. The EER of gender classification using G-COTS software on gender-perturbed outputs increases to 20-40%, and the EER of gender classification using AFFACT and IntraFace on these outputs surpasses 60%. Comparisons between the gender prediction results on the outputs of PrivacyNet and the outputs of the face-mixing approach by Othman and Ross [107], as well as the model by Sim and Zhang [131], show that in case of G-COTS, the PrivacyNet results are superior in terms of increasing the EER (Fig. 6.6).

Note that we did not include the results of the GAN model in Fig. 6.6 for readability sake. However, we observed that the GAN model shows larger deviations (which is advantageous) in cases where gender was intended to be perturbed. This is expected since the GAN model does not have the constraints from the auxiliary face matcher. Therefore, there is more flexibility for modifying the patterns of the face. However, a disadvantage of the GAN model is that it also significantly degrades the matching utility as shown in Section 6.3.2.

Next, we consider the distributions of gender prediction scores using G-COTS and AFFACT on original images as well as the outputs of PrivacyNet, as shown in Figure 6.7 and 6.8. In order to measure the change in the distributions of male and female scores, we used KL-divergence. The kl-divergence is computed for each individual label (i.e., distributions of males before and after). Then, the weighted average of the results of KL-divergence of males and females are combined together to get the total KL-divergence. The results indicates that for G-COTS scores, the KL-divergence changes in modest amount, from 0.49 to 1.05 (see Figure 6.7. On the other hand, the KL-divergence obtained from the results of AFFACT shows larger amounts, ranging from 2.9 to 3.8.

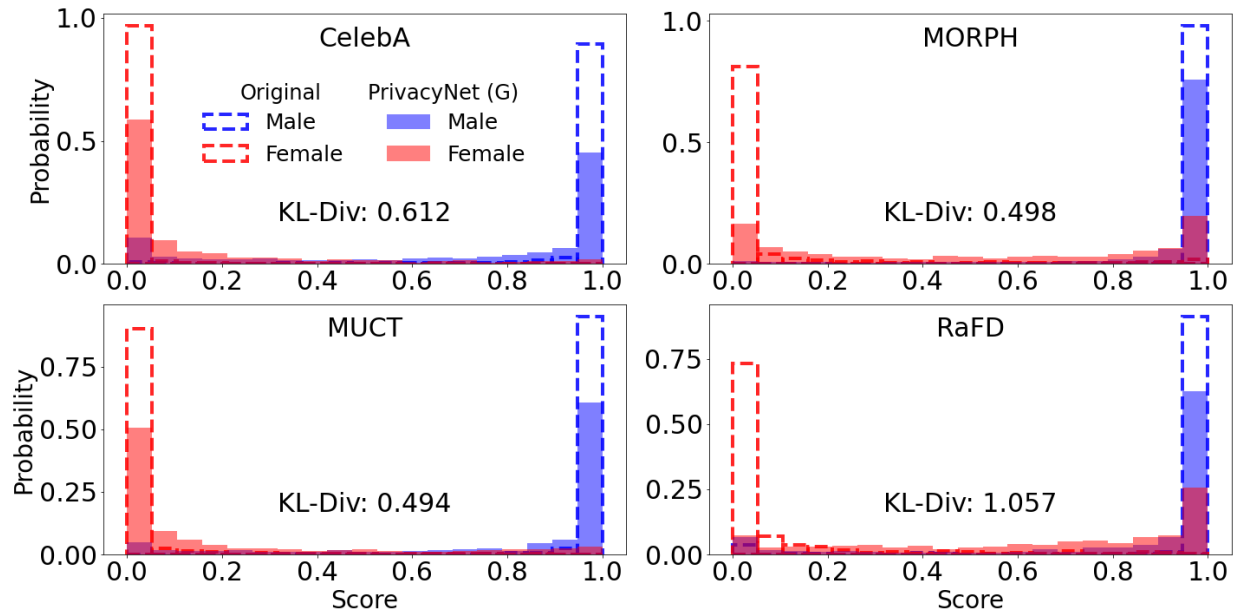


Figure 6.7: Change in distribution of gender prediction scores as assessed by G-COTS on the original images as well as the outputs of PrivacyNet.

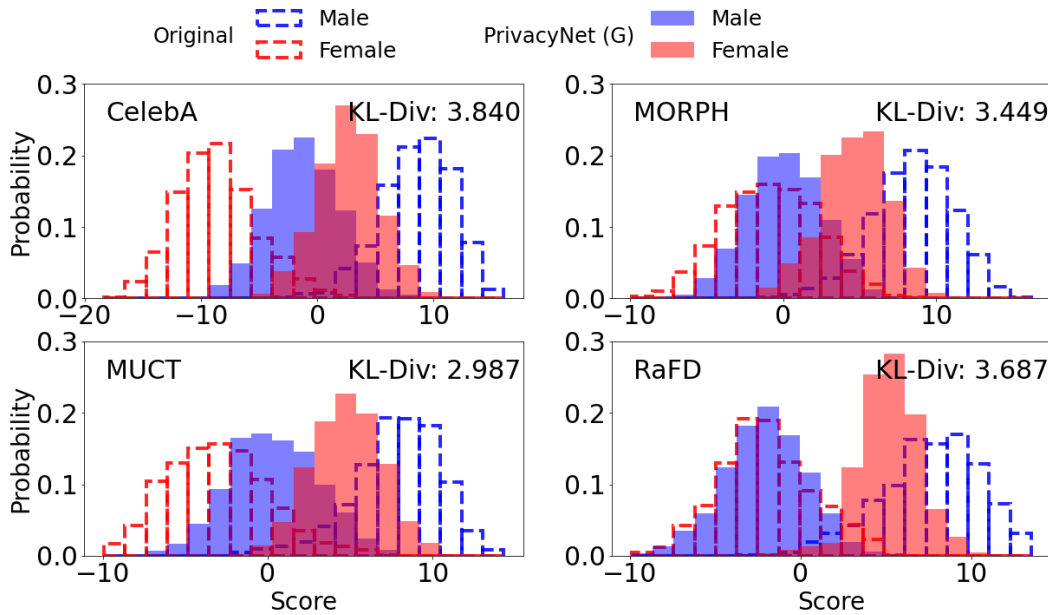


Figure 6.8: Change in distribution of gender prediction scores as assessed by AFFACT on the original images as well as the outputs of PrivacyNet.

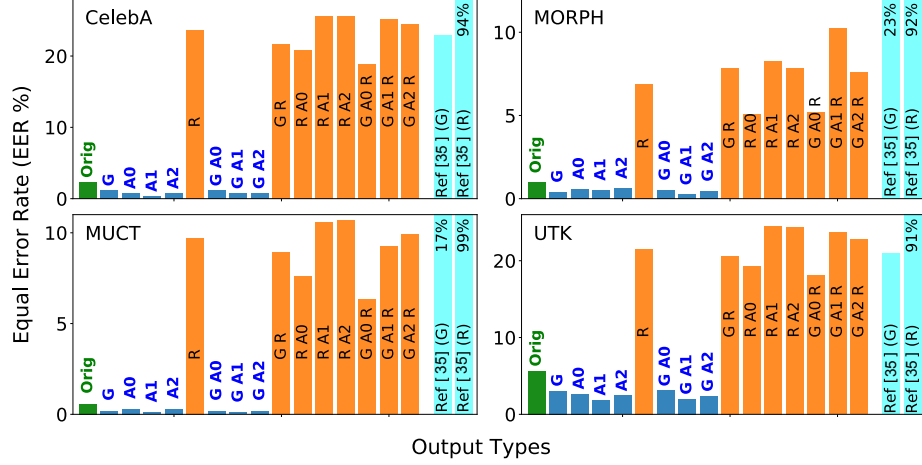


Figure 6.9: Performance of the race classifier, R-COTS, on original images as well as different outputs of the proposed model. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1, and A2 (the larger the difference the better). The outputs of PrivacyNet, where the race attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.

Race Prediction Performance: We conducted the race prediction analysis using a commercial-off-the-shelf software, R-COTS. Similar to the gender classification experiments, we show the EER of race classification on original images as well as the different outputs of the PrivacyNet model in Fig. 6.9. Since the face mixing approach proposed in [107] was only formulated for gender and not race perturbations, we did not include it in this section.

The EER results in Fig. 6.9 show that PrivacyNet successfully meets the objectives of our study for confounding race predictors. The outputs where race is not intended to be perturbed (shown in blue) exhibit low EER values similar to the EER obtained from the original images ($\text{EER} \sim 1\%$). On the other hand, when race is selected to be perturbed, the EER values increase significantly ($\text{EER} \sim 20\%$ for CelebA and UTK-face, and $\text{EER} \sim 10\%$ for MORPH and MUCT datasets). The results of separately perturbing gender and race using the controllable face privacy method proposed in [131] are also shown for comparison. When the race attribute is perturbed according to [131], the performance is slightly higher than our model. However, the disadvantage of the

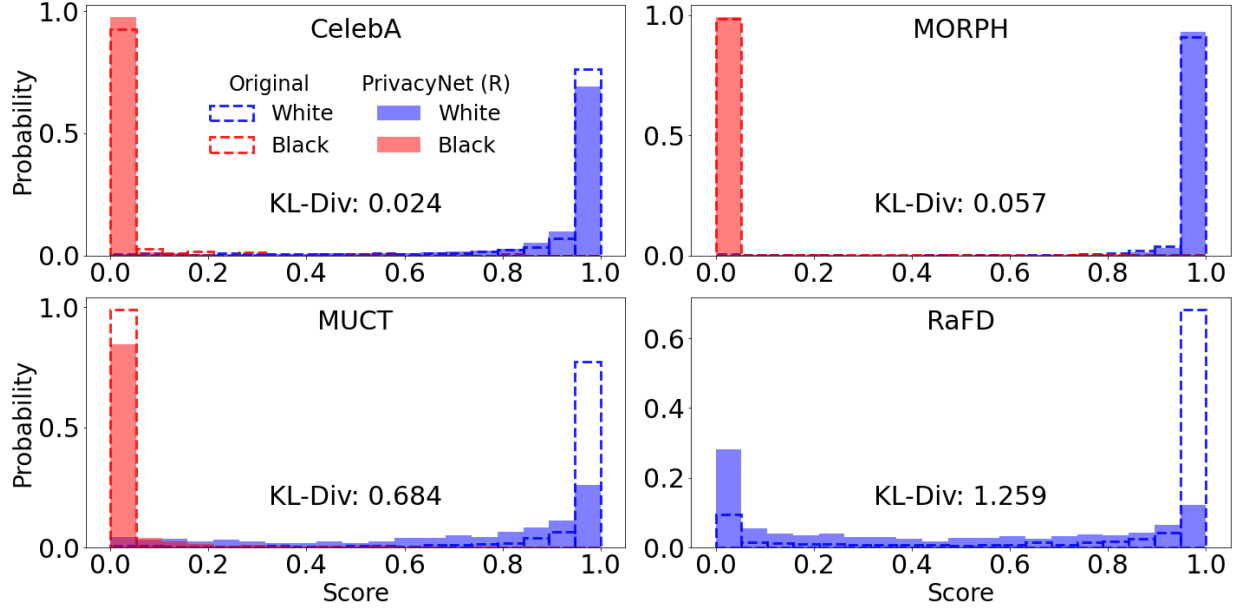


Figure 6.10: Change in distribution of gender prediction scores as assessed by G-COTS on the original images as well as the outputs of PrivacyNet.

controllable face privacy method [131] is that when it perturbs the gender attribute, it also affects the race predictions.

In addition, Figure 6.10 shows the distributions of race prediction on white and black samples from the original images as well as the samples after undergone PrivacyNet perturbations. The KL-divergence values for the four datasets changes 0.024 for the CelebA dataset, to 1.259 for RaFD.

Age Prediction Performance: To assess the ability of PrivacyNet for confounding age information, we used a commercial-off-the-shelf age predictor (A-COTS), which has shown remarkable performance across the different datasets tested in this study (Fig. 6.11). We used the Mean Absolute Error (MAE) values in unit of years to measure the change in age prediction before and after perturbing the images (Fig. 6.11). As mentioned previously (Section 6.2.4), the ground-truth age values for three datasets – CelebA, MUCT, and RaFD – are not provided. Therefore, for these three datasets, the MAE values are computed as the difference between the age predictions on the output

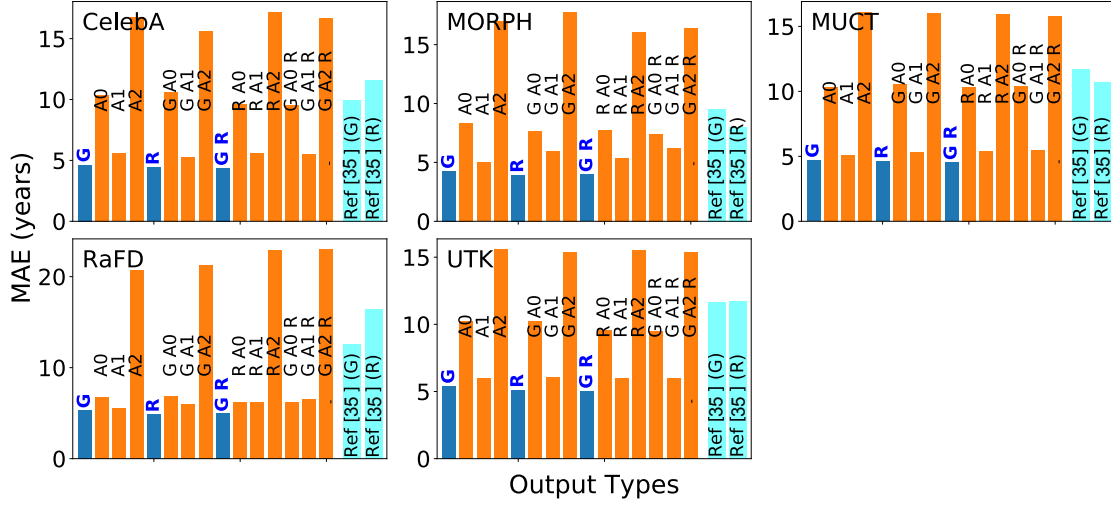


Figure 6.11: Change in age prediction of A-COTS on different outputs of the proposed model. This is with respect to the age predicted on original images for CelebA, MUCT and RaFD, and the ground-truth age values for MORPH and UTK-face. Different outputs are marked by their selected attributes: G: gender, R: race, and A: age, where the specific target age group is denoted as A0, A1 and A2. The outputs of PrivacyNet, where the age attribute is selected for perturbation, are shown in orange, and the rest are shown in blue.

images and the predictions on the original images, while for the other two datasets, MORPH and UTK-face, the ground-truth values are used for computing the MAE values.

The results of age-prediction show that the MAE obtained from the outputs, where age is not meant to be perturbed, remains at approximately 5 years. However, when we intend to modify the age of face images, using the label A2 results in the highest MAE (around 20 years for RaFD and 15 years for the other four datasets) compared to A0 and A1. A possible explanation for this observation is that, due the nature of the aging process, larger textural changes occur in face images belonging to A2. The MAE of the A0 group is also relatively large (except for RaFD), which may be caused by the reversal of the textural changes. However, the results of the middle-age group (A1) is similar to the cases where we did not intend to modify the age. We hypothesize that the small changes in A1 are also due to the non-stationary aspect of aging patterns; the age perturbations via the PrivacyNet model can potentially be improved by using an ordinal regression

approach for age prediction.

6.3.2 Retaining the Matching Utility of Face Images

Besides obfuscating soft-biometric attributes in face images, another objective of this work is to retain the recognition utility of all outputs of PrivacyNet. For this purpose, we conducted matching experiments using three unseen face matchers: commercial-off-the-shelf software (M-COTS) and two publicly available matchers, SE-ResNet-50 trained on the VGGFace2 dataset [23] (SE-Net for short), and DR-GAN [143]. Fig. 6.12 shows the ROC curves obtained from these matching experiments for four datasets – CelebA, MORPH, MUCT, and RaFD. The UTK-face dataset is removed from this analysis since it does not contain subject information. Since PrivacyNet generated 15 outputs for each input face image, the minimum and maximum True Match Rate (TMR) values at each False Match Rate (FMR) value are computed and only the range of values for these 15 outputs are shown. Note that it is expected for the matching utility to be retained in all these 15 outputs. Similarly, the range of TMR values at each FMR obtained from the 15 different outputs of the GAN model that did not have the auxiliary face matcher for training, is also shown for comparison. The ROC curves of PrivacyNet are very close to the ones obtained from the original images for each dataset, compared to the baseline results, which both show significantly larger deviations. It is worth noting that the baseline-GAN is equivalent to removing the matching loss term $\mathcal{L}_{G,m}$ from PrivacyNet. As shown in Figures 6.5, 6.12 and 6.13 (“Baseline-GAN”), the PrivacyNet model produces more realistic-looking faces images without the matching loss term. However, removing the matching loss term results in a severe decline in matching performance, affecting both the true matching rate and identification accuracy (Figs. 6.12 and 6.13). The coefficient λ can be further tuned to control the trade-off between the performance of face-matching and obfuscating the soft-biometric attributes.

In addition to the ROC curves, we have also plotted the Cumulative Match Characteristics (CMC) [38], as shown in Fig. 6.13. According to the CMC curves, the results of PrivacyNet match very closely with the CMC curves obtained from the original images in all cases, which shows that PrivacyNet retains matching utility.

It is worth noting that Ref. [131] has more favorable CMC curves than the other methods evaluated in this study. A plausible explanation is that Ref. [131] aligns and normalizes its inputs to a reference face image, which significantly reduces the intra-class variations. This reduction of intra-class variations increases the number of true positives. However, it also increases the number of false positives, thereby deteriorating the *ROC* performance. One may argue that the difference in performance could be due to the different training datasets that were used to train our model and that of Ref. [131], and, perhaps, re-training Ref. [131] would be necessary for a fair comparison. However, we note that we used the original model for Ref. [131], which was constructed from a carefully curated dataset, and the original authors of [131] recommended against retraining.

6.4 Ablation study on cycle-consistency term

While the presented model in this chapter has shown to be effective in imparting multi-attribute demographic privacy to face images, in this section, we look at the effect of the cycle-consistency loss term in more detail. To do this, we perform a preliminary analysis, in which a new experiment is designed where PrivacyNet model is trained without the cycle-consistency term. This model is denoted as Baseline2, and its performance compared with the results of PrivacyNet and those of Baseline1 (a GAN model without face-matching loss term). Example inputs and outputs of these models are shown in Figure 6.14.

Qualitatively, the results suggest that the cycle-consistency could have been removed without

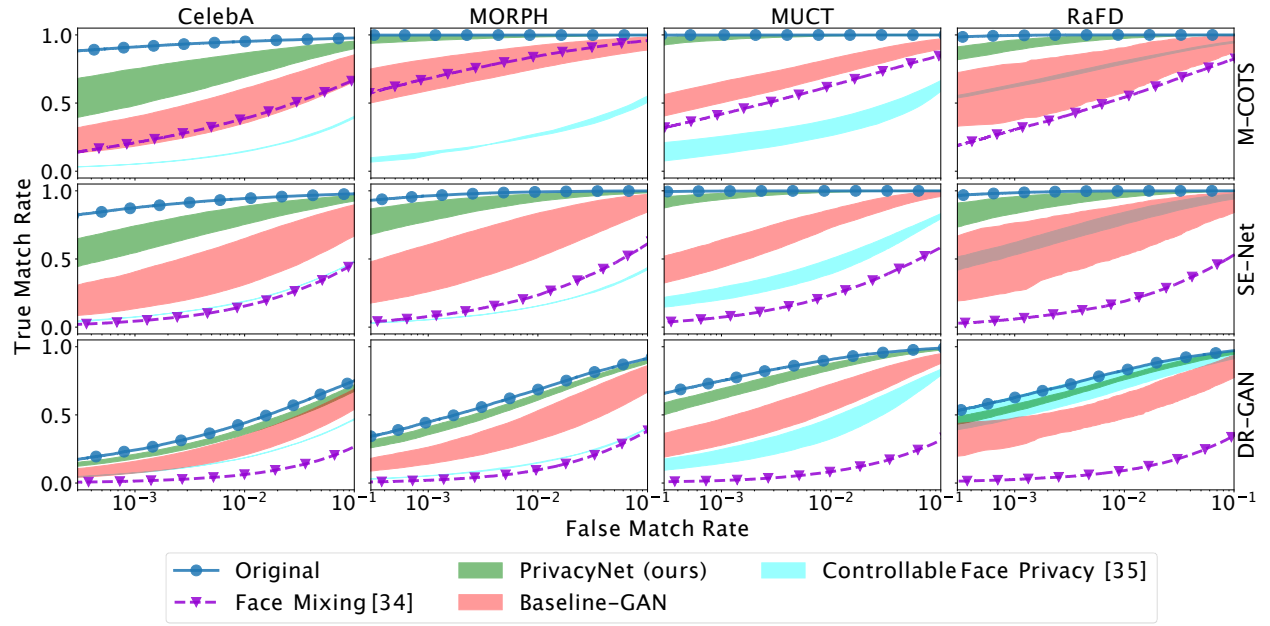


Figure 6.12: ROC curves showing the performance of unseen face matchers on the original images for PrivacyNet, the baseline-GAN model, face mixing [107] approach and the controllable face privacy [131] method. The results show that ROC curves of PrivacyNet have the smallest deviation from the ROC curve of original images suggesting that the performance of face matching is minimally impacted, which is desired.

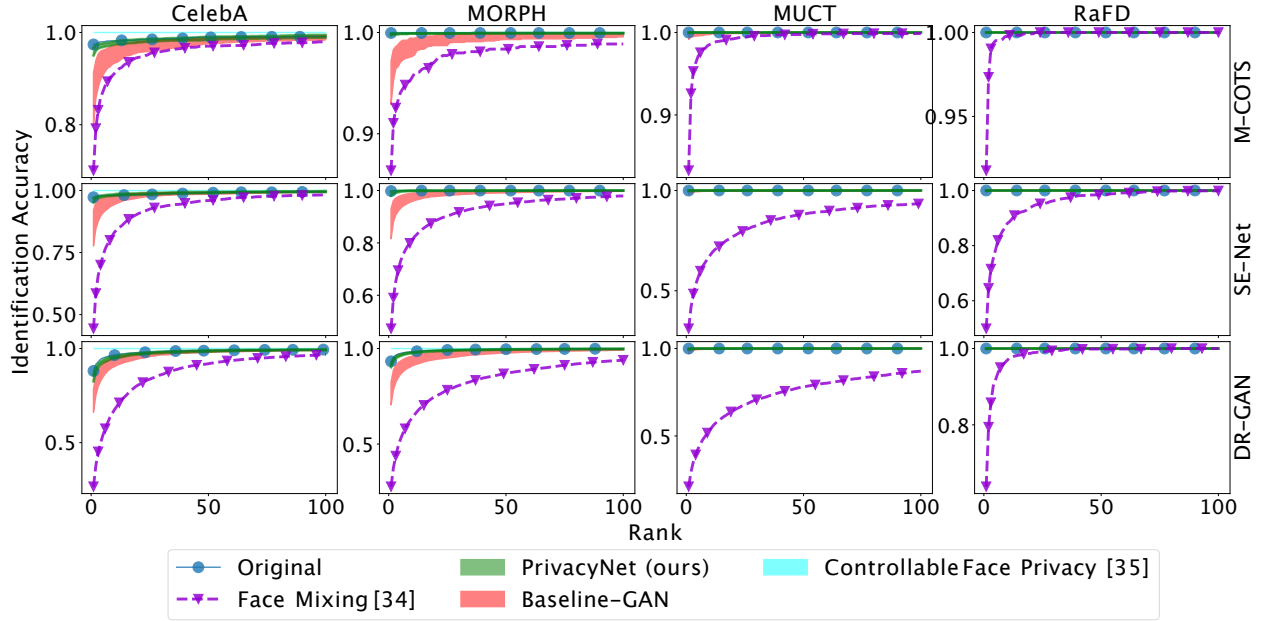


Figure 6.13: CMC curves showing the identification accuracy of unseen face matchers on the original images. Also shown is the range of CMC curves for the PrivacyNet model and the baseline-GAN model, along with that of the face mixing [107] and controllable face privacy [131] approaches. It must be noted that in cases where the results of PrivacyNet or GAN are not visible, the curves overlapped with the CMC curve of the original images: this means that there was no change in matching performance at all (which is the optimal case). The results confirm that transformations made by PrivacyNet preserve the matching utility of face images.



Figure 6.14: Some input and output example images, based on the PrivacyNet, as well as two ablation models (PrivacyNet without cycle-consistency loss, PrivacyNet without the matching loss term).

affecting the main two objectives of this work (i.e., biometric utility is not adversely affected, while improving the obfuscation of the demographic attributes. This is mainly due to the fact that cycle-consistency was included in CycleGAN [155] for improving the visual quality of the generated outputs. However, in this study, the visual quality is of lesser importance, compared to the two other objectives. As a result, removing the cycle-consistency could potentially relax the constraints imposed on the model and thereby, resulting in improving its performance.

Based on the objectives of this work, we look at two measures. First, we consider the ROC curves of gender classification, before and after applying the perturbations using all three models (PrivacyNet, Baseline1, and Baseline2), as shown in Figure 6.15. The results shown in this figure indicates that the performance of Baseline2 is close to that of PrivacyNet. Comparing the area under the ROC curves show that AUC of Baseline2 is 94%, which is slightly higher than that of

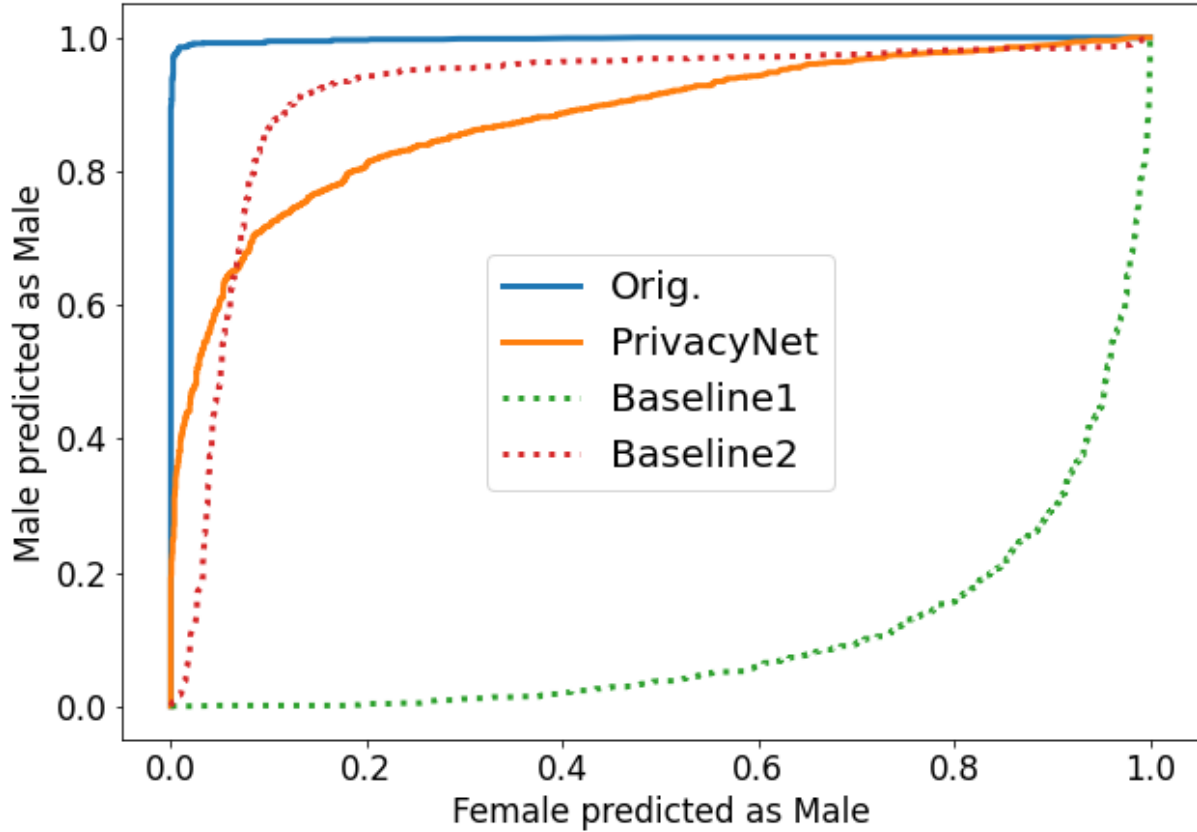


Figure 6.15: Comparison of gender classification on the original images as well as the outputs of PrivacyNet and two ablation experiments: Baseline1 is the model without the matching term, and Baseline2 is the model without the cycle-consistency loss.

PrivacyNet (87%).

Next, we also compare the performance of face-matching as shown in the Figure 6.16. In this case, Baseline2 shows 5% higher TMR at $FPR = 0.1\%$: 79% for Baseline2 vs. 74% for PrivacyNet.

As a result, we conclude that removing the cycle-consistency loss term improves the obfuscation of gender attribute, while it deteriorates the matching performance.

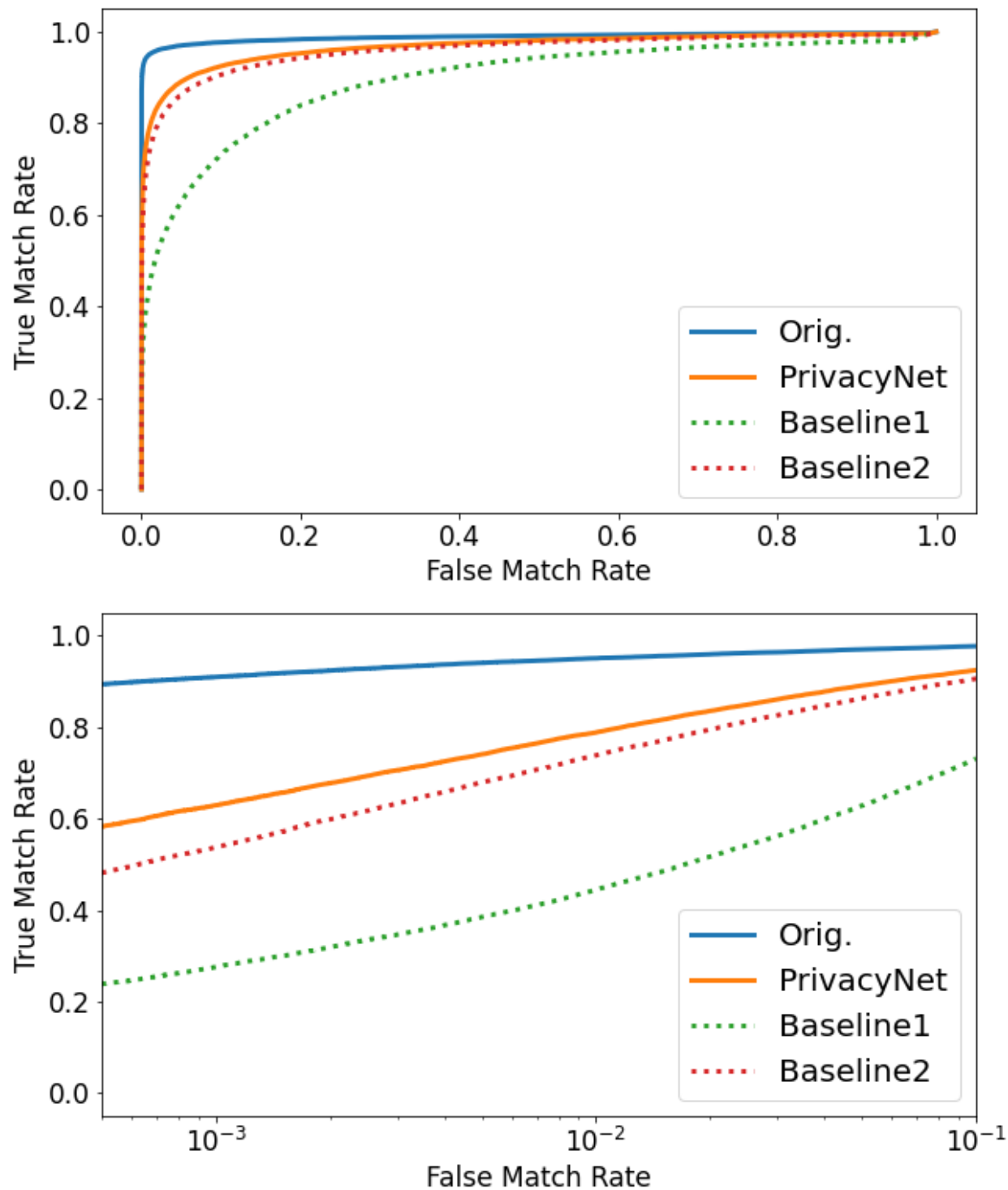


Figure 6.16: Comparison of matching accuracy on the original images, as well as the outputs of the PrivacyNet model along with two ablation experiments: Baseline1 is the model without the matching term, and Baseline2 is the model without the cycle-consistency loss (Top in normal scale, bottom in log-scale).

6.5 Debiasing face recognition

Face recognition models are known to suffer from algorithmic bias [9]. This algorithmic bias has typically results in certain demographics having poorer face recognition accuracy than others. In this regard, one application of removing facial attributes from face images is to debias face-recognition systems. Sixue *et al.* [51] proposed DebFace, a model that removes demographic attribute information from face representation vectors, in order to debias the face recognition systems. Given that our proposed PrivacyNet also obfuscate demographic attributes from face images, in this section, we investigate whether the proposed algorithm can potentially be used for debiasing face recognition algorithms. To do this, we first computed face representation vectors before and after applying the perturbations by PrivacyNet. Following Sixue *et al.* [51], we then group the samples into different categories based on their demographic attributes. For example, in case of perturbing gender, two groups are formed: a group of male samples vs. a group of female samples. Then, we measured the verification accuracy for all subjects in each group separately. The results of verification accuracy of each demographic group before and after applying perturbations made by PrivacyNet is shown in Figure 6.17.

As we can see from the results shown in Figure 6.17, the performances of different demographic groups before and after applying perturbations are very similar to each other, and we do not observe a noticeable change towards debiasing face recognition algorithms. We argue that this observation is due to the following reasons:

1. PrivacyNet does not completely remove the demographic information from face images, but rather, applies perturbation towards making the prediction of demographic attributes less reliable.
2. As PrivacyNet works in the image-level, the perturbed face images still need to be processed

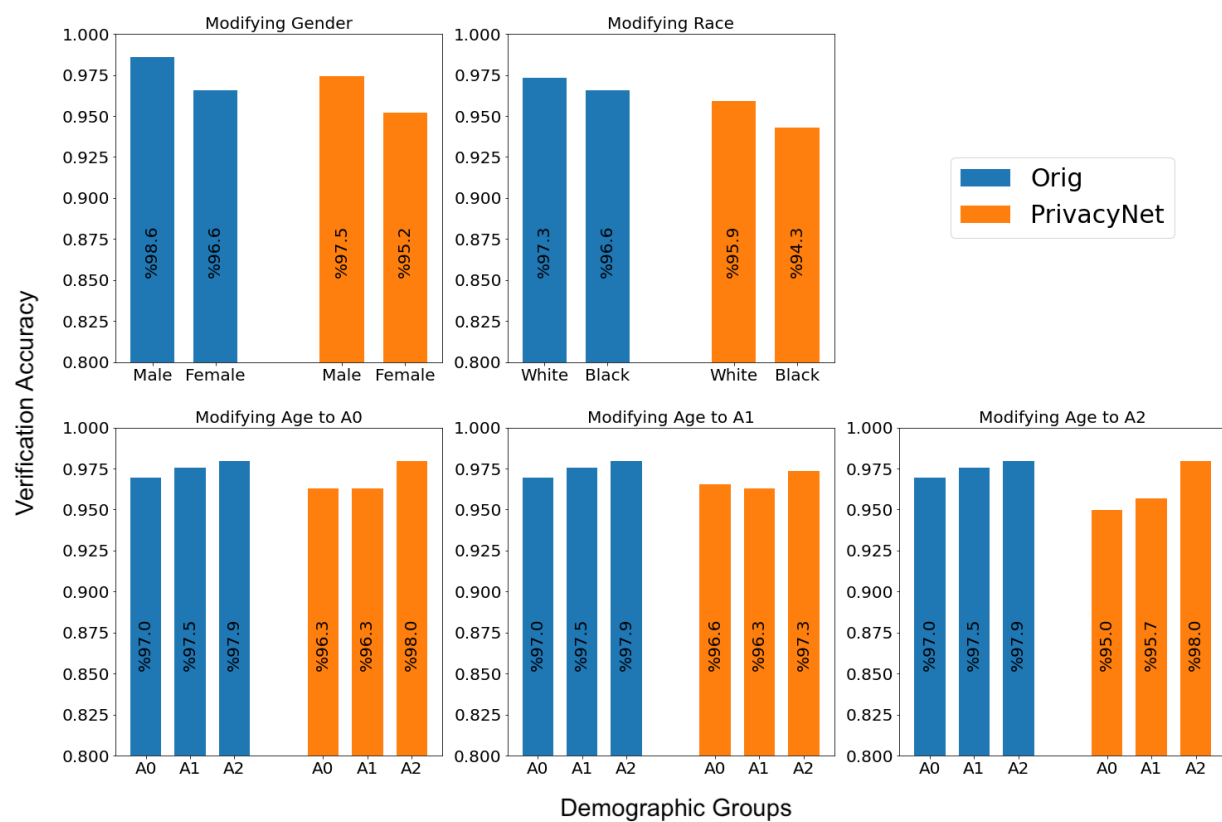


Figure 6.17: Comparison of verification accuracy for different demographic groups before (original) and after applying PrivacyNet perturbations, to investigate algorithmic bias.

by face matchers in order to calculate their face representation vectors, and this is the step that introduces the algorithmic bias. While other debiasing algorithms (such as Sixue *et al.* [51]) work on the extracted face representation vectors.

6.6 Summary and Future Work

In this work, we designed a special neural network model coined PrivacyNet for imparting multi-attribute privacy to face images including age, gender and race attributes. PrivacyNet utilizes the Semi-Adversarial Network (SAN) empowered by Generative Adversarial Networks (GAN) to synthesize a new face image from an input face image, where certain attributes are perturbed selectively, while other face attributes are preserved. Most importantly, the matching utility of face images from this transformation is preserved. Experimental results using three unseen face matchers as well as three unseen attribute classifiers show the efficacy of our proposed model in perturbing such attributes, while the matching utility of face images is not adversely impacted.

In the next chapter, we discuss two open problems for imparting soft-biometric privacy using the SAN models. First, it is important to see if the perturbations can be reliably detected using a machine-learning system. Secondly, the perturbations made by SAN from the human perspective needs to be investigated. We then describe our proposed work to study these issues.

Chapter 7

A deeper study on perturbations

7.1 Motivation

The problem of face recognition entails the comparison of face images for biometric recognition [91, 70]. A facial recognition system can operate in two modes. In the verification mode, the input face image is compared against only those images in the gallery belonging to the claimed identity in order to determine the veracity of the claim. In the identification mode, no identity is claimed, and so the input face image is compared against all the gallery images in order to determine its identity. With the unprecedented advancements in deep learning and computer vision [52], face recognition is now being widely deployed in a number of applications ranging from smartphones to border control to security and surveillance [46, 73, 81, 18]. However, recent research has raised several concerns regarding the fairness of facial recognition systems across different demographic groups [138, 137, 26, 35, 40]. Furthermore, a face image divulges rich auxiliary information, , since advancements in machine learning has enabled the extraction of age, gender¹, race², and health status [36]. However, extracting such sensitive information from a person's face image requires permission from the users, and extracting such information without permission violates users' privacy [100, 101].

¹Throughout this work, we assume binary labels for gender; however, it must be noted that societal and personal interpretation of gender can result in many more classes.

²While ethnicity is related to cultural identity of individuals in a group, race can inform physical characteristics.

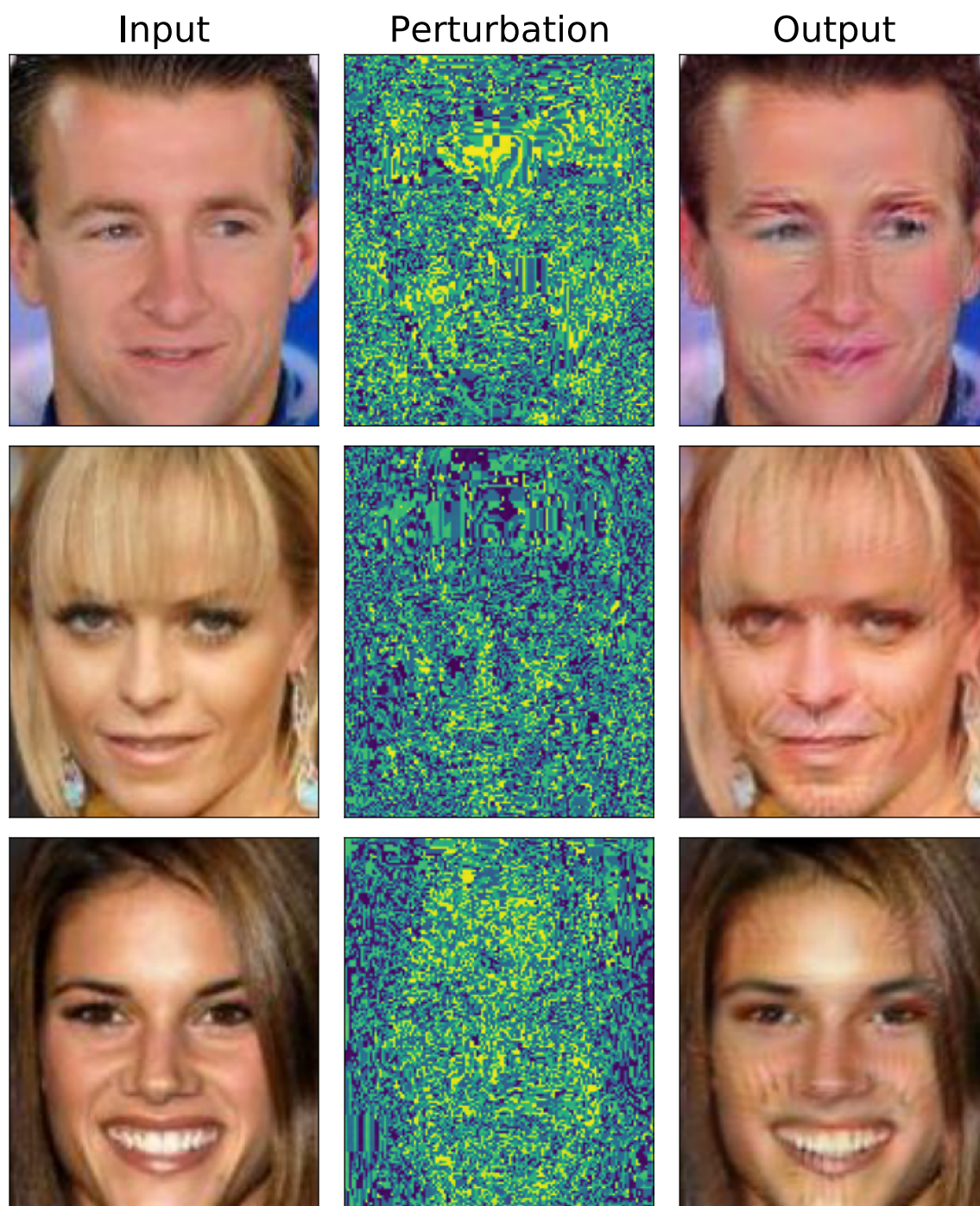


Figure 7.1: Example perturbations added automatically to input face images by the PrivacyNet [97] model for confounding gender information with respect to automated gender classifiers. The perturbations target gender-related features, with minimal adverse effects on recognition performance.

In order to enhance privacy, a new line of research has emerged in order to preempt the extraction of such auxiliary information without adversely affecting the recognition utility of face images [31, 103, 101, 100]. Early work toward this goal was based on adversarial mechanisms for confounding face attributes by perturbing the face images or their templates [31, 100]. However, generalizability of such methods is an issue. In other words, perturbations that confounded one attribute (say, gender) classifier, may not have the desired effect on another classifier. Morales *et al.* [103] developed SensitiveNet, which adds perturbations to face representation vectors based on a current attribute classifier, followed by training a new classifier using the perturbed representation vectors. This process is repeated until the performance of a newly trained classifier on perturbed representation vectors does not reach a threshold. While this technique has shown its efficacy in removing auxiliary information in face representation vectors, in some applications, it is necessary rely on using face images and apply such perturbations to the face images directly. To address this limitation, Mirjalili *et al.* [101] developed Semi-Adversarial Networks (SANs) for confounding arbitrary face attribute classifiers, and its extended version called PrivacyNet [97] for confounding multiple attributes (viz., age, gender, race). Some examples of input face images, their perturbations and the resulting outputs for confounding gender classifiers are shown in Fig. 7.1.

While previous work [103, 98, 97] have shown their efficacy in imparting demographic privacy to face images, not much study has been conducted towards *interpretability* of the perturbation and *understanding* how such models work. While the machine learning community has investigated the process of generating adversarial examples and interpreting their effect on the classifier [8, 54], the effect of adversarial examples in confounding human observers is negligible as adversarial examples are mostly created as imperceptible perturbations to input images [54, 108].

However, we do not know how SAN-based models add perturbations to input face images [97, 98]. Furthermore, the effect of SAN perturbations on human observers is still not clear. In the

domain of machine learning, interpretability of a model means understanding why a ML model has given certain output. Therefore, in this paper, we investigate the perturbations introduced by SAN, and study the interpretability of SAN results. Furthermore, the effect of the perturbations introduced by SAN models on human observers is also studied.

The contributions of this work are as follows:

- Studying how the SAN model finds the perturbations for confounding an attribute classifier
- Designing experiments for studying the effect of SAN perturbations on human observers

The outline of the paper is as follows. We first provide an overview of SAN model as proposed by Mirjalili *et al.* [101] in Section 7.2. Next, in Section 7.3, we study the interpretability of the perturbations added by SAN to input images using two methods, namely, Grad-CAM [128] and CNN-fixations [102]. Finally, in Section 7.3.1, we describe our designed experiments for studying the effects of perturbations applied to face images on human observers.

7.2 An Overview of Semi-Adversarial Networks

Semi-Adversarial Networks (SAN) was first proposed by Mirjalili *et al.* [101] for imparting demographic privacy to face images. The general idea of SAN (see Fig. 7.2) relies on adding perturbations to an input face image in order to confound the utility of one classifier (*e.g.*, a gender classifier), while retaining the utility of another classifier (*e.g.*, a face matcher). To do this, they leveraged auxiliary classifiers where the SAN model learns to synthesize perturbed outputs that can confound one classifier, while minimizing the effect of such perturbations on the other classifier. In the context of demographic privacy, the utility of face recognition is retained, while confounding the extraction of demographic attributes. The schematic representation of a typical SAN model is

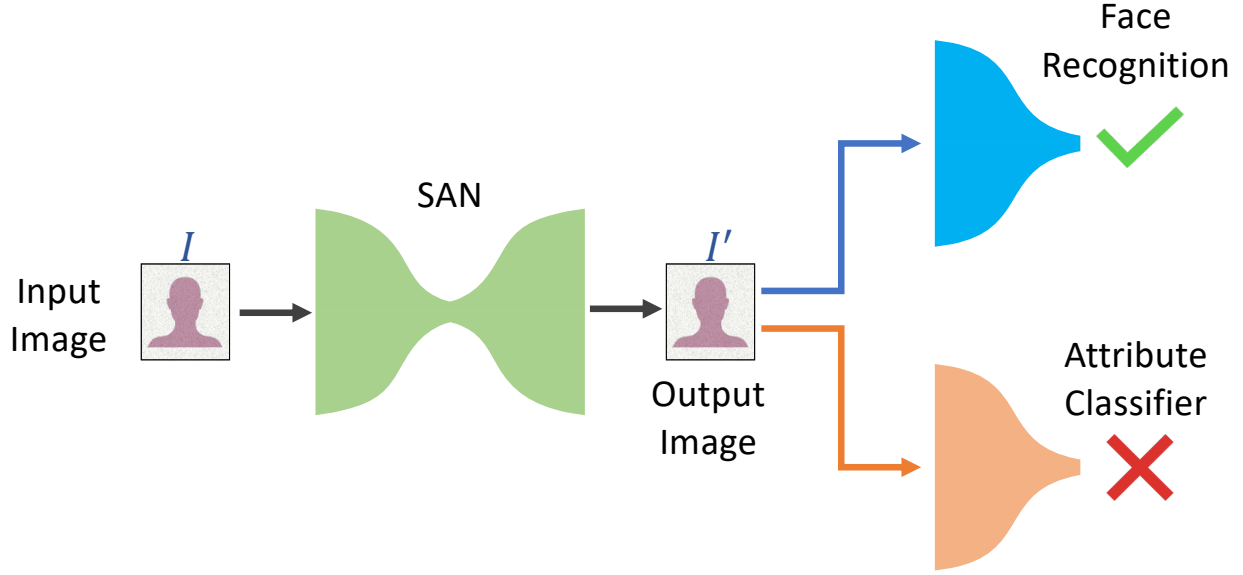


Figure 7.2: The general idea of semi-adversarial networks originally proposed by Mirjalili *et al.* [101] for imparting demographic privacy to face images. The SAN model adds perturbations to an input face image such that demographic attribute(s) cannot be reliably extracted from the output, while the output can still be used for face recognition.

shown in Fig. 7.2.

Mirjalili *et al.* [97] later extended the SAN model to multi-attribute face privacy, where a combination of demographic attributes (e.g., age and gender) could be selectively confounded. To do this, they used conditional Generative Adversarial Networks (GANs) [155, 33], composed of a generator, a discriminator and one or more auxiliary attribute classifiers. The generator tries to synthesize realistic-looking outputs while at the same, the outputs can selectively confound or retain the utility of auxiliary attribute classifiers based on the given conditional variable [97].

In contrast to the original SAN model [101], PrivacyNet [97] does not rely on a pre-trained auxiliary attribute classifier. Instead, an auxiliary attribute classifier is trained simultaneously with the generator. Furthermore, the auxiliary attribute classifiers and the discriminator sub-networks are shared, which helps the model learn the distribution of face attributes based on the training data. This results in improving the generalizability of PrivacyNet with respect to unseen attribute

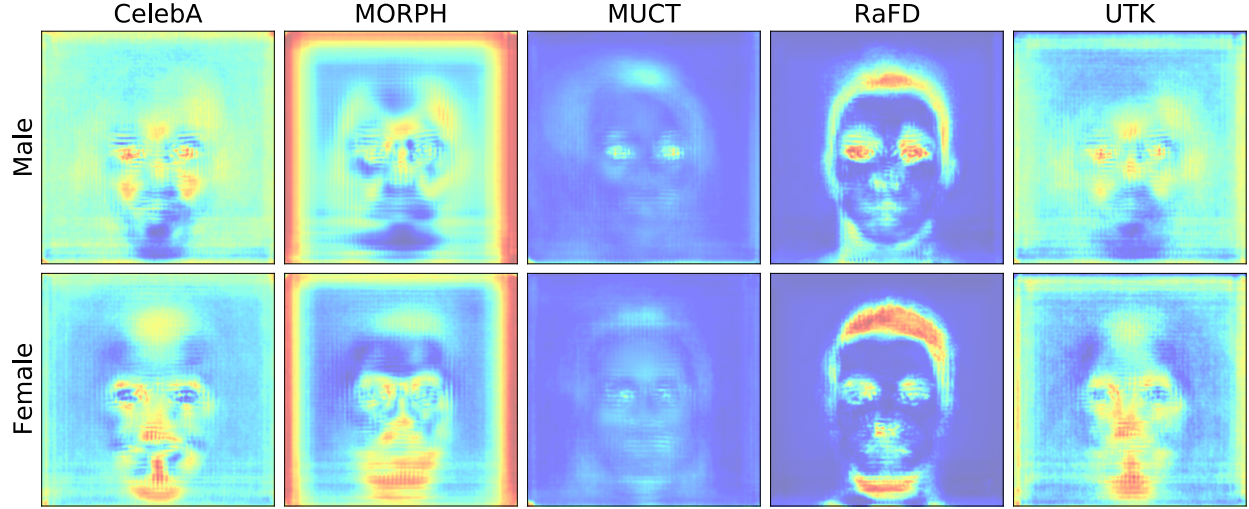


Figure 7.3: Heatmaps of average perturbations for Female and Male samples obtained using a privacy preserving method (PrivacyNet [97]). The results indicate that in most datasets, PrivacyNet focuses on hair and eyes for input female face images, while for input male face images. Furthermore, the area related to facial hair (beard and moustache) have opposite effects for male and female face images.

classifiers.

Figure 7.3 shows the average heatmaps of the intensity of perturbations introduced by PrivacyNet to input face images for confounding gender information on CelebA [86], MORPH [122], MUCT [94], RaFD [83], and UTK-face [154]. It can be observed that in most datasets, the pixels corresponding to facial hair (moustache and beard) are smoothed-out for male samples, while the opposite effect has occurred for female samples.

7.3 Understanding SAN perturbations

While Mirjalili *et al.* [97] have shown the efficacy of their proposed SAN model in imparting demographic privacy to face images, in this section, we delve deeper to interpret the perturbations produced by SAN, which is important for deploying a machine learning model in a real-world application. To accomplish this, we have used two models that allow for identifying which pixels

in the input image are important for the prediction made by a deep neural network classifier.

The first model for interpreting DNN results is called CNN Fixations [102], which starting from the network output, selects the activations of the previous layer that lead to positive activations at the current layer. The second model for interpreting the results is another visualization technique for DNN classifiers called Grad-CAM [128]. Note that while the CNN fixations rely on positive activations at each layer of a DNN, Grad-CAM is a gradient-based technique which focuses on features with the highest gradient values.

The CelebA dataset [86], which contains 202,599 face images of celebrities, was split into 90% train and 10% test partition in a subject-disjoint manner, according to [97]. The PrivacyNet model was trained on CelebA-train dataset. Furthermore, a CNN-based gender classifier with 4 convolutional layers followed by 2 fully-connected layers was further trained independent of the SAN training, and then evaluated on the examples in CelebA-test dataset.

In the following sub-sections, we will see the results of both CNN-fixations and Grad-CAM models.

7.3.1 Studying the interpretability of SAN perturbations using CNN-fixations

We have applied CNN-fixations to the results of the gender classifier using two versions of the examples in CelebA-test dataset as input to the classifier. First we applied CNN-fixations to the original test examples before applying PrivacyNet. Second, we obtained the outputs of PrivacyNet using CelebA-test examples, and then applied CNN-fixations on outputs of PrivacyNet. The results of CNN-fixations before and after adding PrivacyNet perturbations are compared in Fig. 7.4.

Based on the results of CNN-fixations on samples from CelebA-test dataset, we can observe that fixations on the original images correspond to the most important features for gender classification. After applying SAN perturbations to these examples, the position of the fixations change

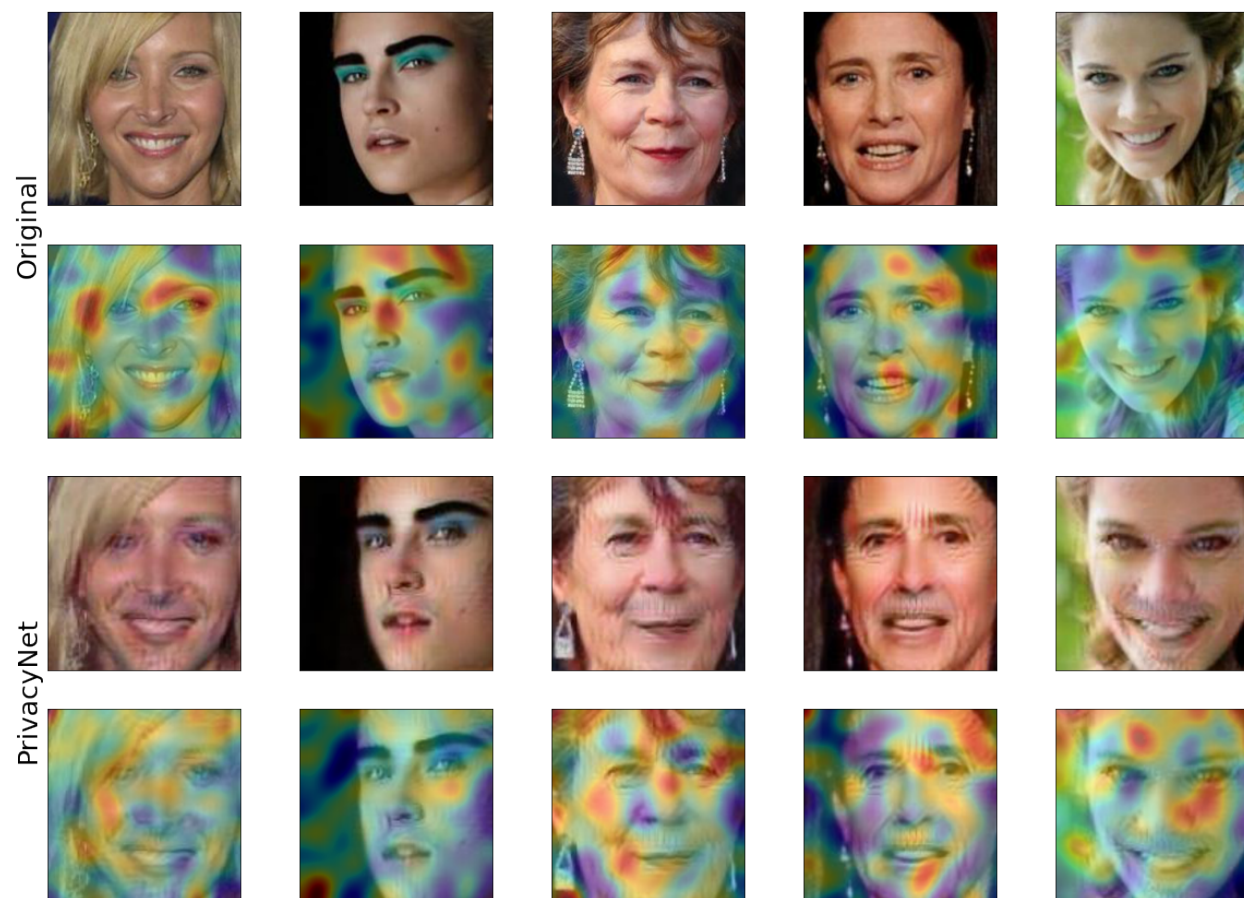


Figure 7.4: Detecting important features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet [97] outputs).

significantly. This suggests that SAN perturbations mostly target the features related to the texture of such regions, and as a result, after applying such perturbations, the important features for gender classification will change in order to make the image become a member of the opposite class. Furthermore, we can also see that in a majority of cases, SAN introduces perturbations in the same location as identified by the CNN fixations. This is an important observation given that the SAN models were trained using auxiliary gender classifiers, while these fixations are obtained from an unseen gender classifier.

Results for other datasets including MUCT [95], MORPH [122], RaFD [83] and UTK-face [154] dataset is provided in the supplementary materials.

7.3.2 Studying the interpretability of SAN perturbations using Grad-CAM

Next, we look at the detected features obtained by Grad-CAM [128]. Similar to the previous subsection, Grad-CAM was applied to the results of the attribute classifier using both original as well as outputs of the SAN model. Figure 7.5 shows the results on samples from CelebA test dataset, before (original images) and after SAN perturbations (PrivacyNet outputs).

While the change in detected features using Grad-CAM from original images to PrivacyNet outputs is consistent with the results of CNN-fixations, we note an important observation. After introducing perturbations, the regions highlighted with blue in Grad-CAM visualization correspond to areas that have not been modified by CNN perturbations. As a result, these blue regions still correspond to the important features for the class label of the original image, and as a result, Grad-CAM gives lower weights to such regions, indicating that these regions are not important for gender classification. In contrast, the regions that have been perturbed to a greater extent are the ones that are also highlighted in red via Grad-CAM visualization (see Figure 7.4 vs. Figure 7.5).

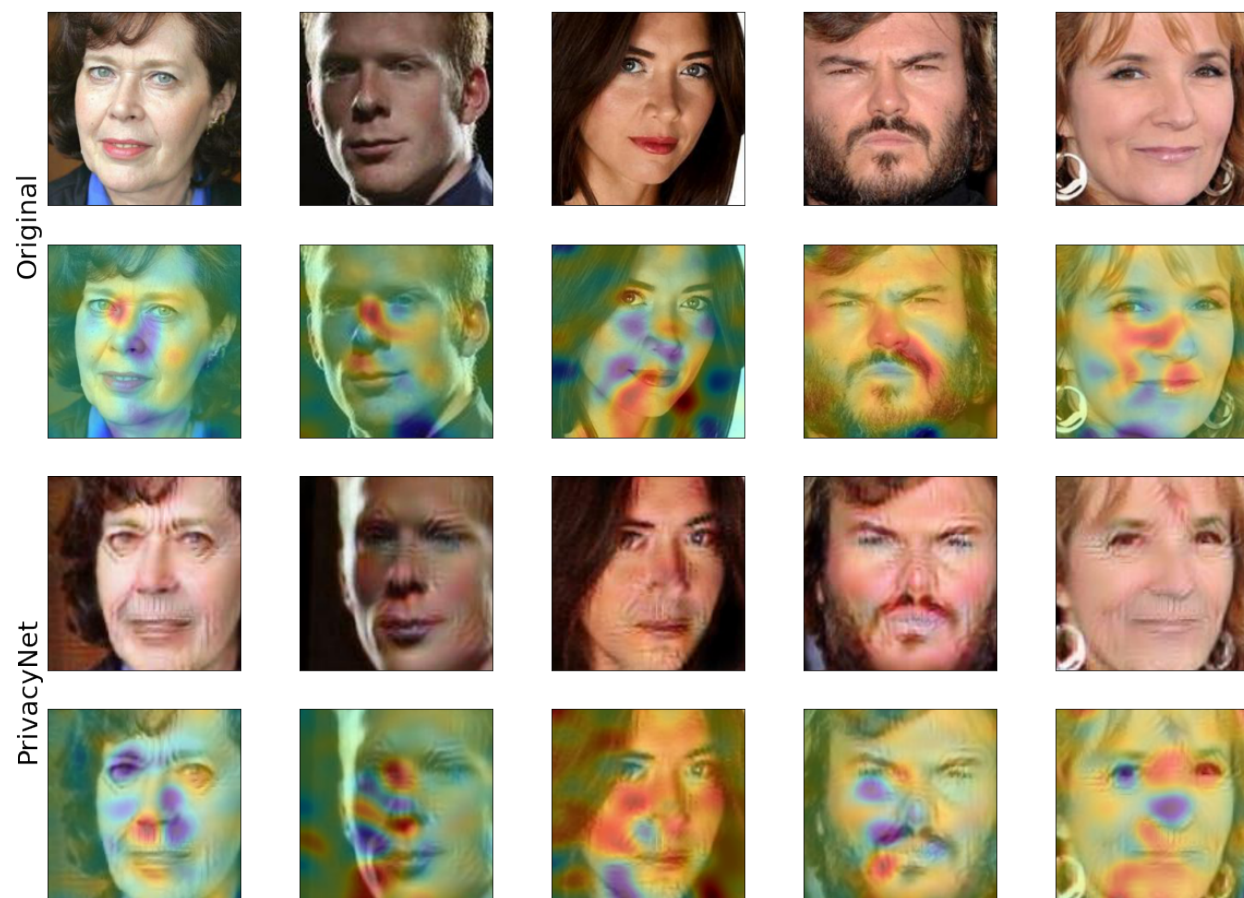


Figure 7.5: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs).

7.3.3 Summary of interpretability studies on SAN perturbations

Based on the results described in the previous subsections (as well as the results shown for other datasets in the supplementary figures), in summary we can make the following conclusion:

- For female faces, SAN focuses mostly on perturbations in the locations containing facial hair such as moustache, beard, as well as removing the lipstick and face blush (if applicable).
- For male faces, SAN focuses on removing beard or moustache or the wrinkles on the forehead. In addition, in some cases the perturbations are towards adding lipstick, blush, or even slightly changing the shape of nose and eyes.

In the next section, we further study the effect of such perturbations on human perception.

7.4 Studying the effect of perturbations from the human perspective

In the following two sub-sections, we present the results of two different experiments. First, we perform a preliminary experiment by presenting the original face images and their cropped versions to (voluntary) human evaluators. In the second experiment, we extend the previous experiment and utilize Amazon MTurk to study human perception of SAN perturbations at a broader scale.

7.4.1 Preliminary experiment with volunteer assessors

For a preliminary experiment on the performance of human observers on gender detection, we presented a few perturbed face images to some volunteer participants. Our preliminary analysis shows that peripheral information that are present in the background of an image has a substantial

impact on the performance of humans in detecting the correct gender attribute from the perturbed face images. Such peripheral information includes wearing jewelries, make-up, type of clothing, as well as hairstyle and the existence of facial hair. Given that such information are distinctive for gender classification, and that human brains have naturally learned to use these discriminatory information for gender classification, then the existence of such information in the perturbed samples help humans correctly classify gender of the perturbed face images.

While humans can use the peripheral information present in a face portrait image, our proposed SAN model does not perturb these information such as clothing or hairstyle. One reason for this is that our model relies on auxiliary gender classifiers, which are trained on a small dataset, compared to the amount of data that humans have observed in their lifetime. Furthermore, humans have the ability to utilize information from other domains. For example, they can easily assign a certain type of clothing that may not have seen before to a particular gender, while gender classifiers trained on only face images would have difficulties doing that. Recall that the SAN model relies on auxiliary attribute classifiers, and these attribute classifiers are trained on a finite dataset. As a result of this, the auxiliary gender classifier used for training the SAN model only focuses on features present in the face area, as opposed to peripheral information. Therefore, our SAN model cannot modify the clothing of an individual in order to confound the gender information present in the image. Instead, we have observed that our model is able to modify some other gender-related features that appear on the face area, for example, facial hair, facial make-up, *etc.*

We note that face recognition matchers generally do not utilize such peripheral information. Information about the clothes, or hair style may not be reliably used for face matching, as this information have high intra-class variability, and therefore, are not useful for face matchers. Most face matchers receive cropped face images as input, and modern face matchers do that internally and only focus on the face area, discarding the peripheral components. Therefore, given that face

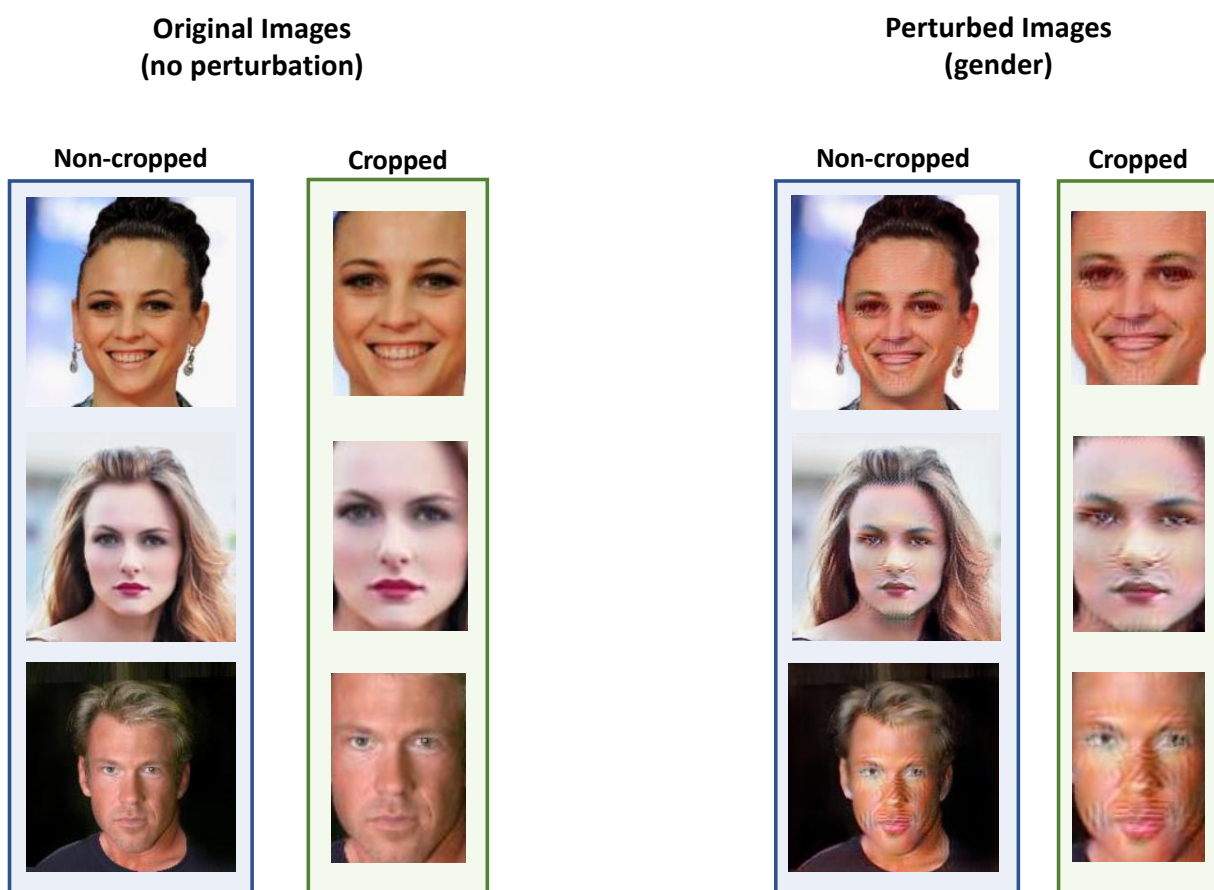


Figure 7.6: Cropped face images vs. full-face portrait images. Humans can easily detect gender from full-face portraits based on the peripheral information such as hair and clothing, paying minimal attention to the details of the facial texture. Therefore, predicting gender from the non-cropped original, cropped original, and non-cropped face images is trivial for humans while detecting gender from cropped face images after applying perturbations (right-most column) is more challenging.

matchers are fairly robust with regard to cropped face images as opposed to full portrait images, we presented the cropped face images to volunteer participants.

The results on measuring the performance of human participants on gender detection from the modified face images are shown in Fig. 7.7. In this experiment, eight face images were modified via PrivacyNet model for confounding gender. The modified face images were then cropped to include only the face areas and remove the peripheral information. The cropped face images were then presented to 11 participants, and their responses were recorded. The number of times each label is provided by participants for these face images are provided in Fig. 7.7. The consensus of the 11 responses for each face image is compared with the ground-truth label of the original images, and when the consensus results match with the true label, the results are highlighted in green. These results show that only 4 out of 8 cases the consensus of 11 participants were able to correctly predict the gender. Such significant change in the consensus results of the 11 participants supports our hypothesis.

The individual responses from each participant are also shown in Fig. 7.8. The responses from participants demonstrate that the highest performance among participant is 6/8, while the lowest performance is 2/8, and the average among all 11 participants is 41%. Furthermore, the performance of participants varies among different face images. The highest performance is observed for face images 2 and 3, which is 8/11 and the lowest value is recorded to be 2/11. Given that humans are considered experts in gender detection due to evolutionary reasons, these results indicate the success of the SAN model in confounding gender attribute from face images.

7.4.2 Human perception study with Amazon MTurk

In the next experiment, in order to get statistically reliable results on the performance of humans, we extended the previous study to Amazon MTurk which have much larger number of participants.

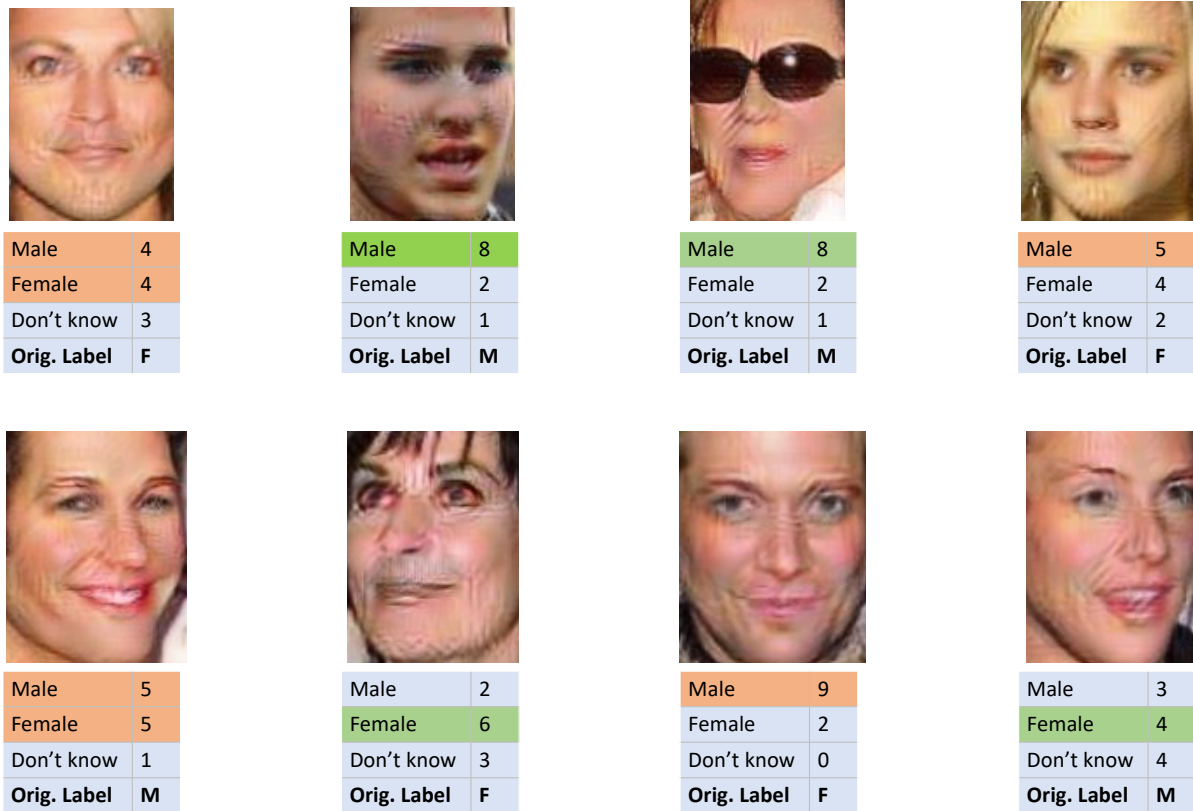


Figure 7.7: Our preliminary experiments on measuring the performance of human observers in gender classification of perturbed images using PrivacyNet. The responses from 11 participants are acquired for each image. The cases where the majority of participants have correctly predicted the gender (matches with the original label) are highlighted with green, and those in which the majority of predictions do not match with the original label are highlighted with orange.



Participants									Correct Prediction Rate
1	M	M	--	M	M	--	M	--	2/8
2	--	M	M	F	F	M	M	F	3/8
3	M	F	M	M	F	F	M	F	2/8
4	--	M	M	--	F	F	M	--	3/8
5	F	M	--	F	M	M	M	--	4/8
6	F	M	F	F	M	F	M	M	6/8
7	M	M	M	M	F	F	F	F	4/8
8	F	F	M	F	--	F	M	M	5/8
9	F	M	M	M	M	F	M	M	6/8
10	--	--	--	--	M	M	F	M	3/8
11	M	M	M	M	M	--	M	F	3/8
Accuracy	4/11	8/11	7/11	4/11	6/11	6/11	2/11	4/11	

Figure 7.8: Our preliminary experiments on measuring the performance of human observers in gender classification of perturbed images using PrivacyNet. The responses from 11 participants are acquired for each image. The cases where the majority of participants have correctly predicted the gender (matches with the original label) are highlighted with green, and those in which the majority of predictions do not match with the original label are highlighted with orange.

In this experiment, each participant is asked to classify the gender in face images. For cropping face images, we used an automatic face detection software, called DLib [77]. Three different versions of each face image in our study were generated: 1) cropped original image, 2) non-cropped output of SAN, and 3) cropped outputs of SAN. A randomized subset of images in each category were given to the participants. We have made sure that a single participant does not see more than one version of one face image because that could have introduced bias to their classification. Figure 7.9 shows snapshots of two example queries displayed to Amazon MTurk participants.

Analyzing the collected data After obtaining the classification results from the participants, we analyzed the data to study the impact of the perturbations on human observers. In this section, we have used a commercial-off-the-shelf gender classifier (G-COTS) that has shown state-of-the-art performance in gender classification. The resulting bar-plots before and after perturbations are shown in Fig. 7.10.

The results from G-COTS shows a decline in performance for both SAN outputs, cropped and non-cropped. However, from the human results, we see that the performance on non-cropped SAN outputs is similar to that of the original images, which shows that humans observers are still able to predict the original gender label from outputs of SAN. However, once SAN outputs are cropped, the performance drops significantly.

The decline in gender detection performance by human observers on the cropped perturbed images confirms our hypothesis that the perturbations are effective for confusing humans in gender classification. Furthermore, the difference in the performance of gender classification by human participants on the non-cropped and cropped output images gives further insight on how much human brain focuses on peripheral attributes (clothing and hair) for gender classification.

Figure 7.11 also shows the confusion matrix of the human results on original samples, outputs of SAN without cropping, and cropped SAN outputs.


Instructions

[View full instructions](#)
[View tool guide](#)

Inspect the displayed face image and identify the gender (sex: male vs female) to the best of your knowledge.

Choose the appropriate label that best suits the image.

Choose the correct category



Select an option

Male	1
Female	2

Zoom in

Zoom out

Move

Fit image

Submit


Instructions

[View full instructions](#)
[View tool guide](#)

Inspect the displayed face image and identify the gender (sex: male vs female) to the best of your knowledge.

Choose the appropriate label that best suits the image.

Choose the correct category



Select an option

Male	1
Female	2

Zoom in

Zoom out

Move

Fit image

Submit

Figure 7.9: Snapshots of two example queries as were displayed to Amazon MTurk participants: Given the displayed image to the participants, they were asked to choose the gender based on their best judgment.

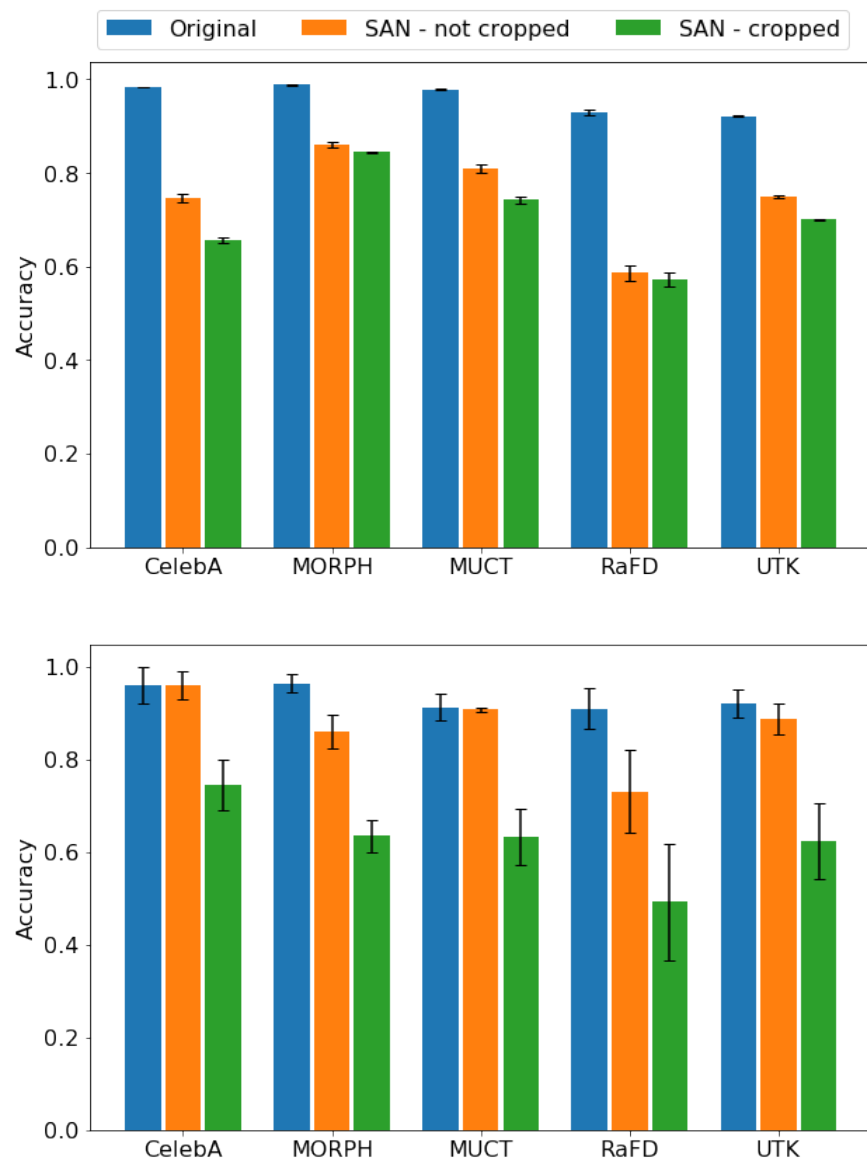


Figure 7.10: Accuracy of gender prediction on original images, SAN outputs without cropping, and cropped SAN outputs: performance of ML-based gender predictor, G-COTS (top), and human performance using Amazon MTurk (bottom).

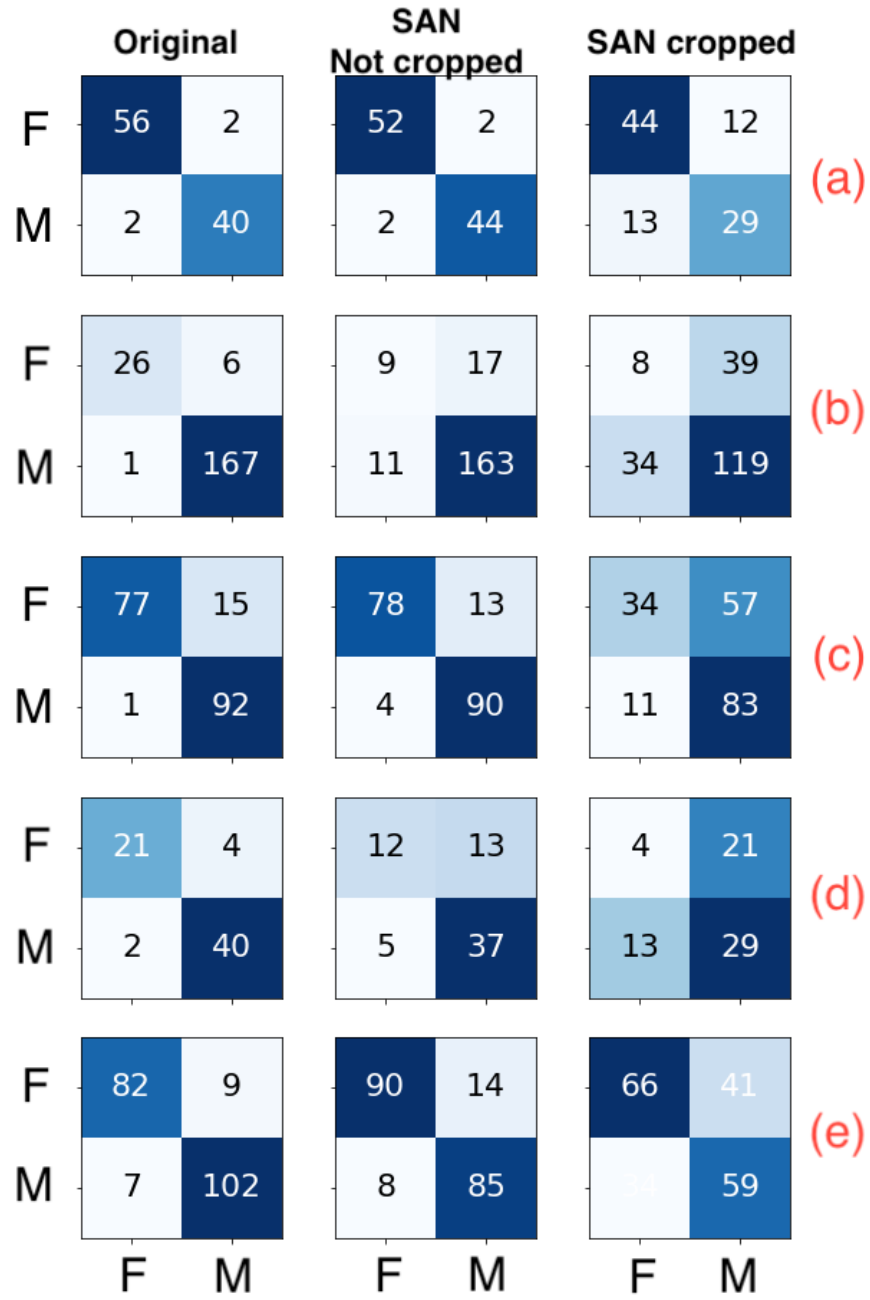


Figure 7.11: Confusion matrices of human performance in gender classification on samples from the original datasets, as well as outputs of SAN without cropping, and cropped SAN outputs: (a) CelebA, (b) MORPH, (c) MUCT, (d) RaFD, (e) UTK-face.

7.4.3 Controlling the trade-off between the degree of privacy vs. matching performance

Based on the results from previous chapters, there is a trade-off between the degree of demographic privacy and the matching performance. Therefore, in order to explore the trade-off between matching performance and the degree of privacy, in this section we perform a preliminary experiment to explore such trade-off while considering only the gender attribute. The purpose of this experiment is to show the feasibility of controlling such trade-off. In this experiment, we explore the operating curve using the hyper-parameter λ_M which is the coefficient of the loss term corresponding to the matching term for training the generator. It is expected that decreasing λ_M results in higher degree of privacy while lowering the matching performance. Figure 7.12 shows the trade-off between the demographic privacy and matching accuracy. Along the x-axis, we have reported the EER for gender prediction, while along the y-axis, we have reported the TMR of the face matcher at an FMR of 0.1%.

This particular analysis was undertaken on CelebA dataset which was split into a 196K training images and 2500 test images. Such curve can be used to select the operating point while considering the trade-off between matching accuracy and degree of demographic privacy. For example, in an application where matching performance is more important than demographic privacy, the operating point can be selected towards preserving the matching performance.

Based on such operating curve, a vendor of face recognition who is deploying this application while concerned with demographic privacy can tune the trade-off between privacy and matching accuracy.

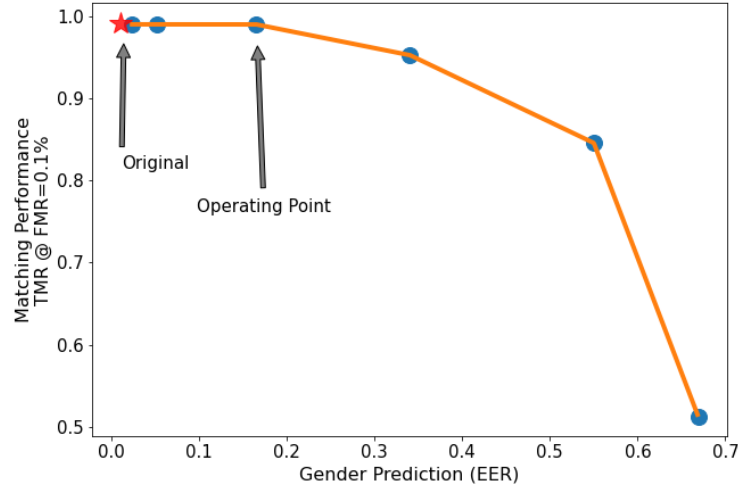


Figure 7.12: Operating curve showing trade-off between matching accuracy and the degree of privacy that can be obtained using PrivacyNet model. This shows the feasibility of controlling the trade-off by carefully selecting the operating point for an application.

7.5 Summary

In this work, we delved deeper into understanding perturbations with two studies. First, we investigated how the SAN model perturbs a face image and the reason for perturbing specific regions of an input face image by the SAN model. We compared the important regions for gender classification before and after applying SAN perturbations, which showed that SAN model tries to target the important regions in the original input face image. In the second study, we investigated the effect of perturbations on gender classification from the human perspective. For this purpose, we analyzed the results of human evaluators which confirmed that the presence of peripheral information such as hair-style, types of clothing and jewelries have significant impact on the performance of human evaluators. Therefore, cropping the face areas on the outputs of SAN resulted in significant decrease in performance of human evaluators.

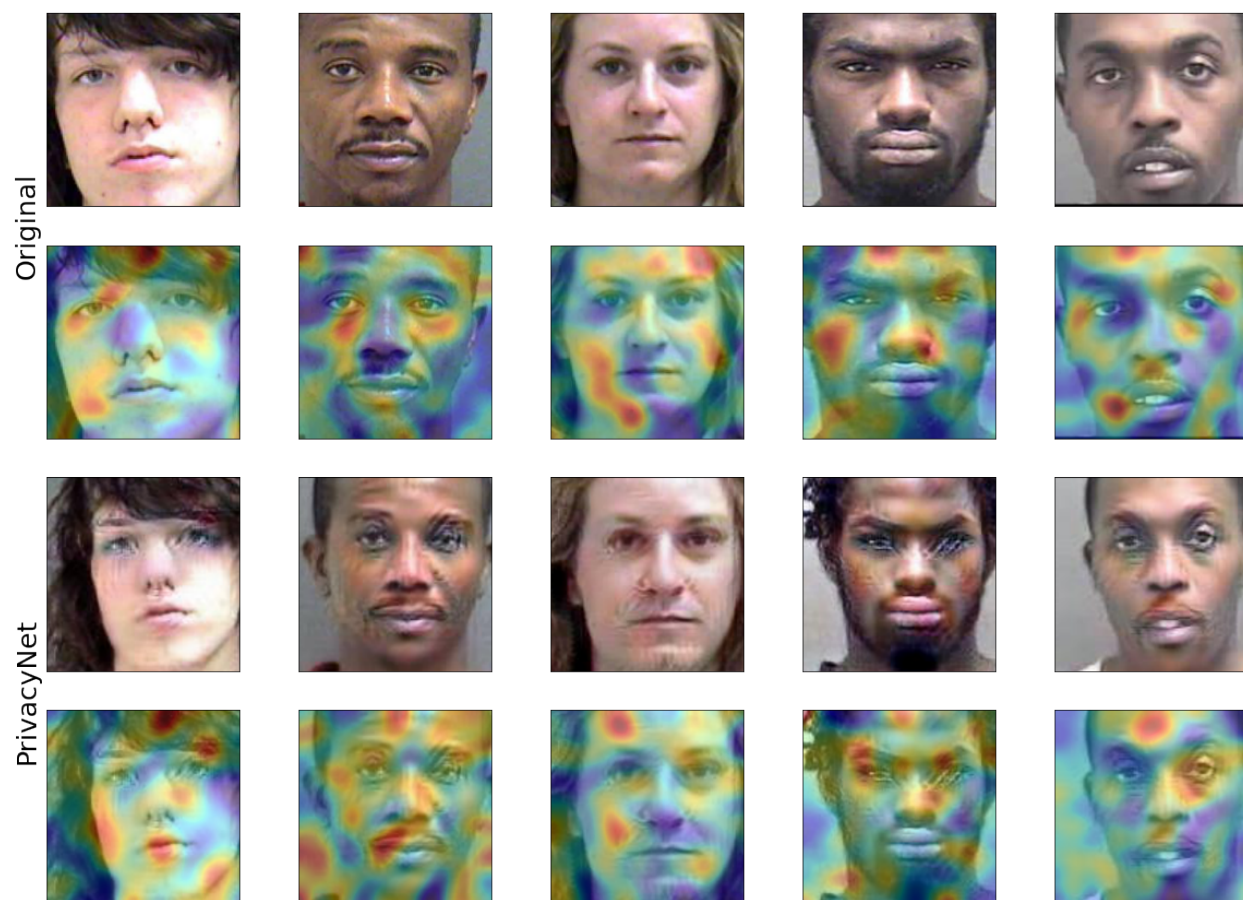


Figure 7.13: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MORPH dataset.

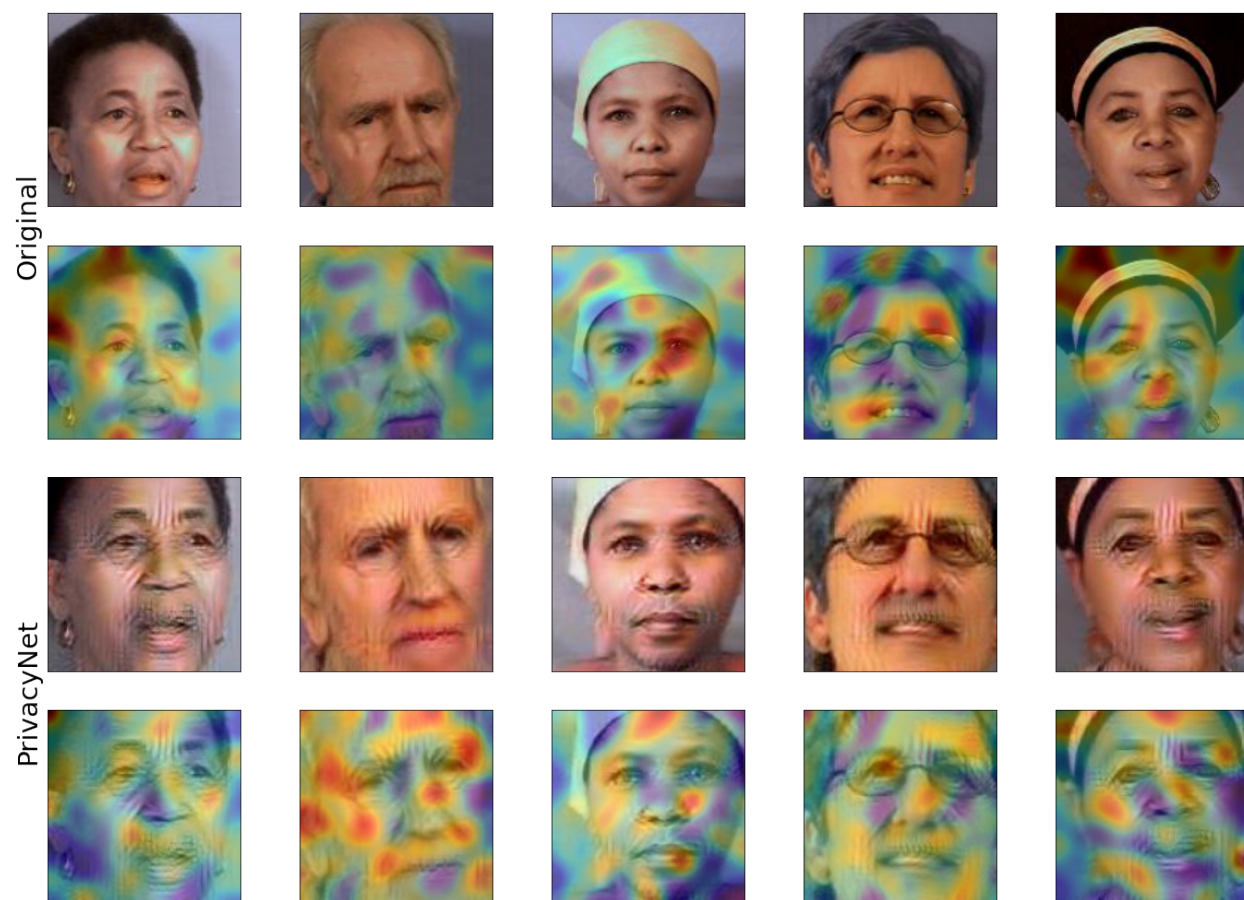


Figure 7.14: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MUCT dataset.

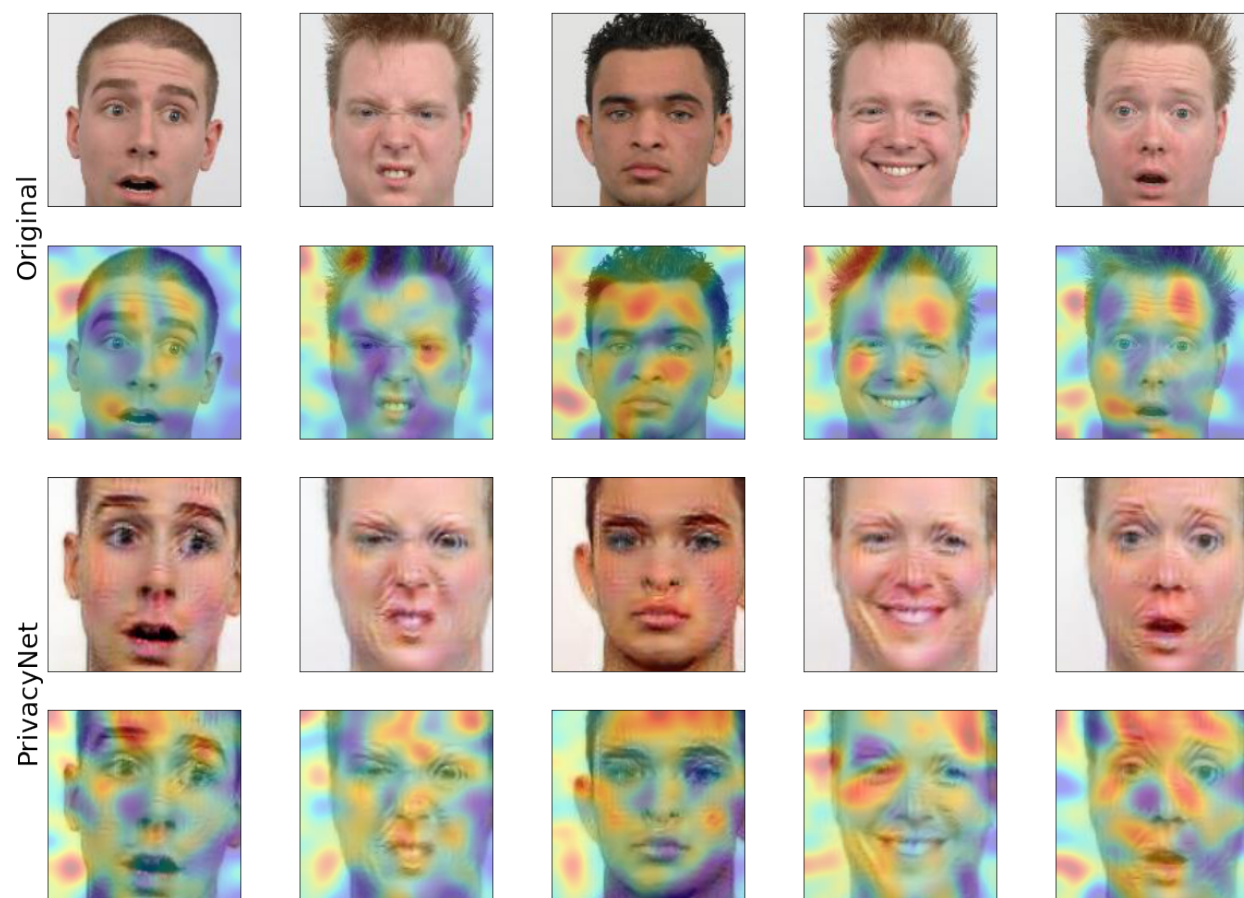


Figure 7.15: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in RaFD dataset.

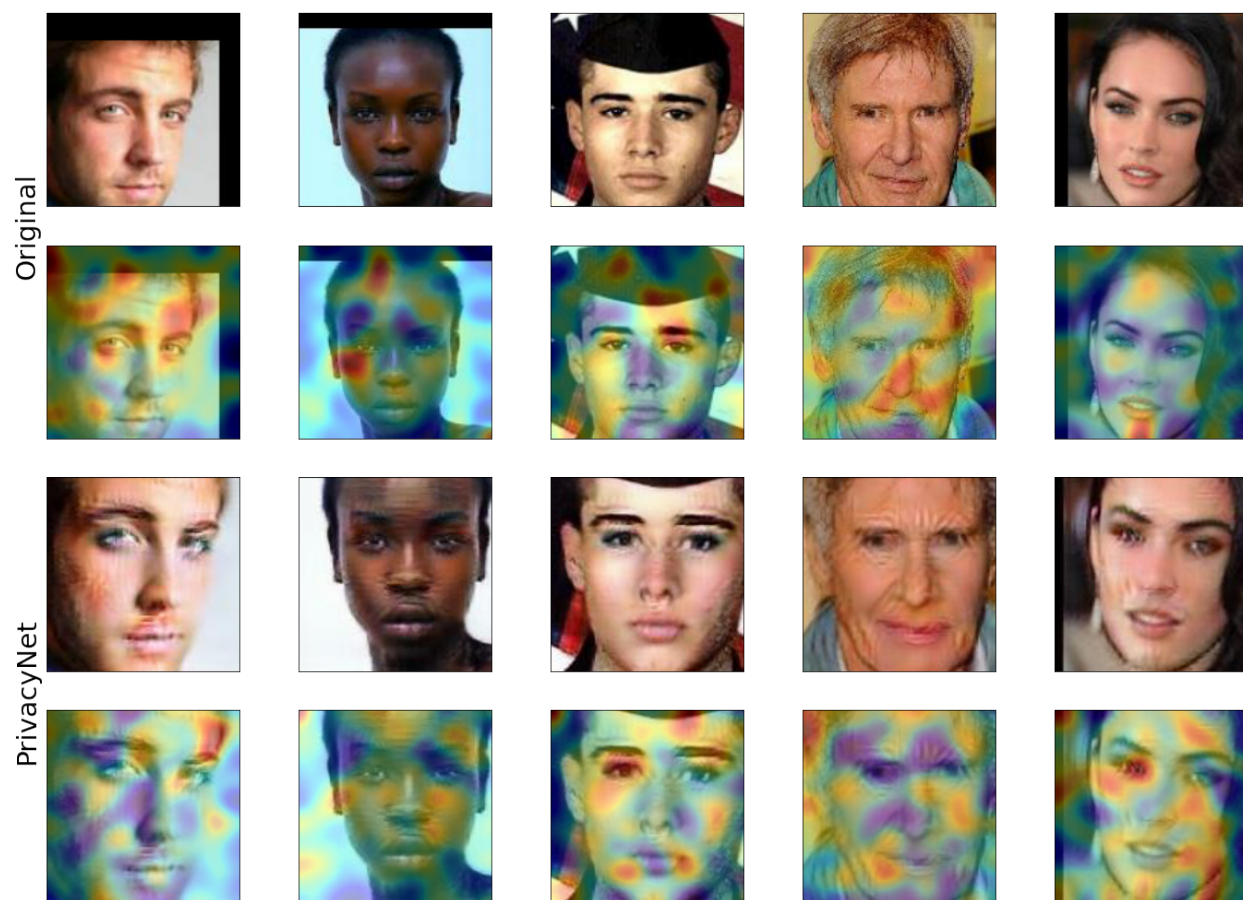


Figure 7.16: Visualizing detected features for gender classification using CNN-fixations [102] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in UTK-face dataset.

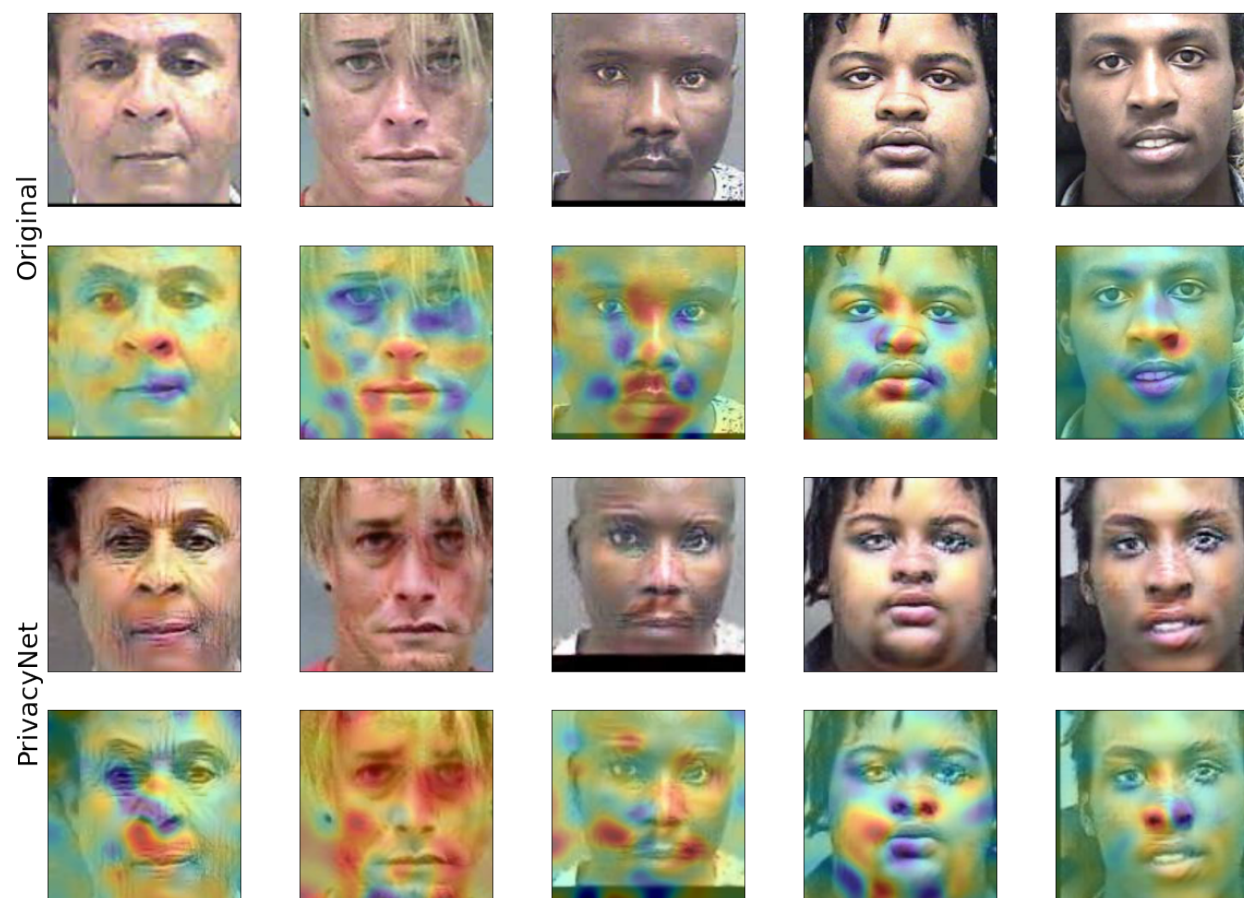


Figure 7.17: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MORPH dataset.

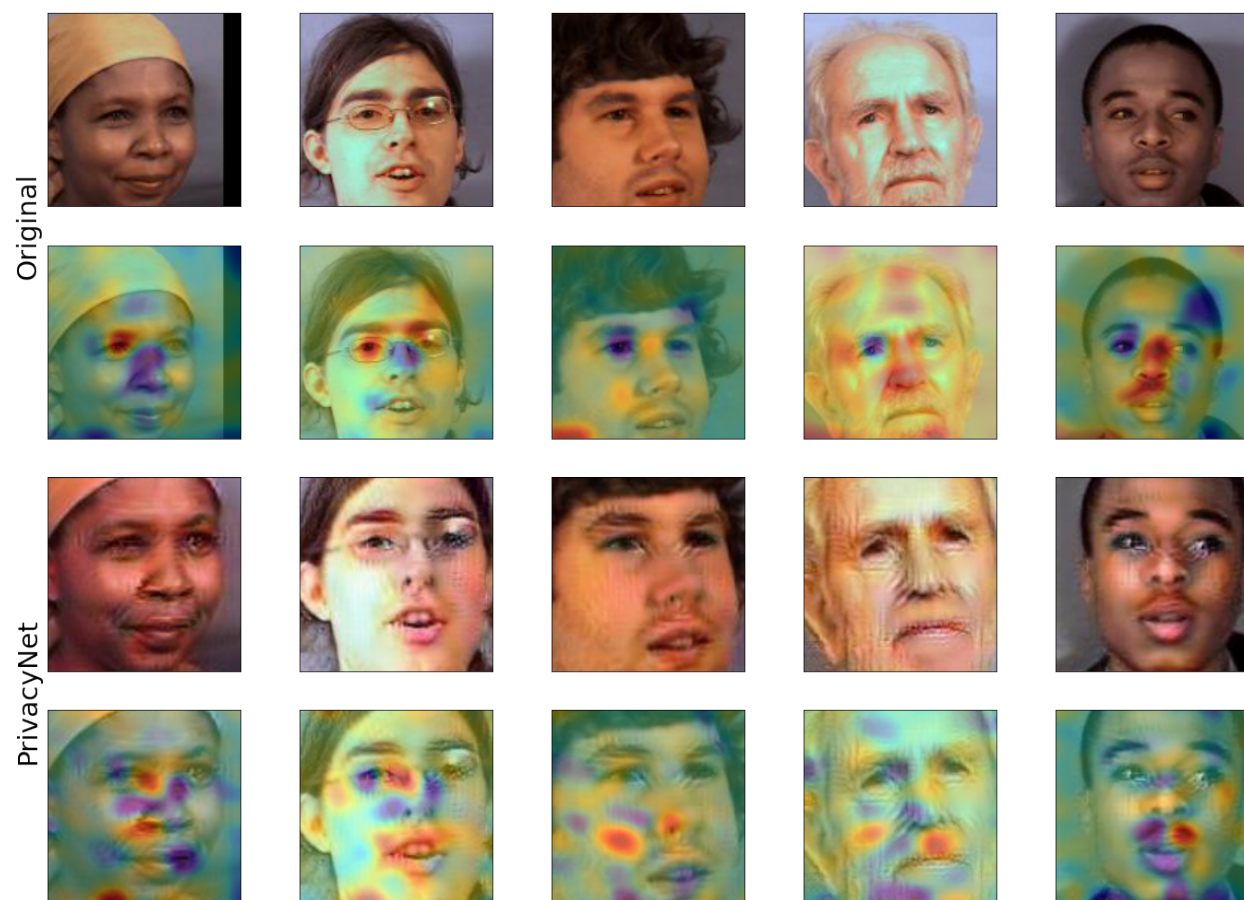


Figure 7.18: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in MUCT dataset.

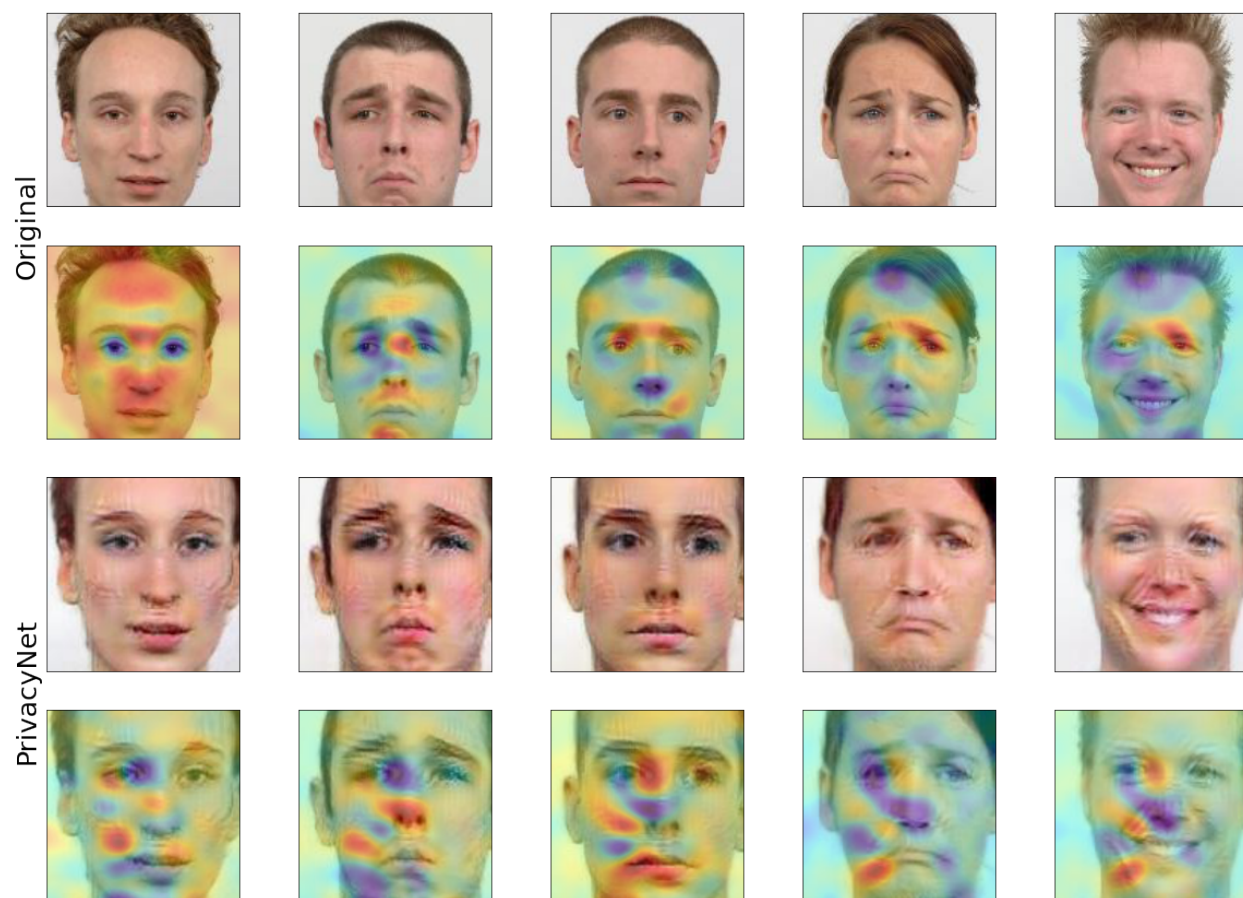


Figure 7.19: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in RaFD dataset.

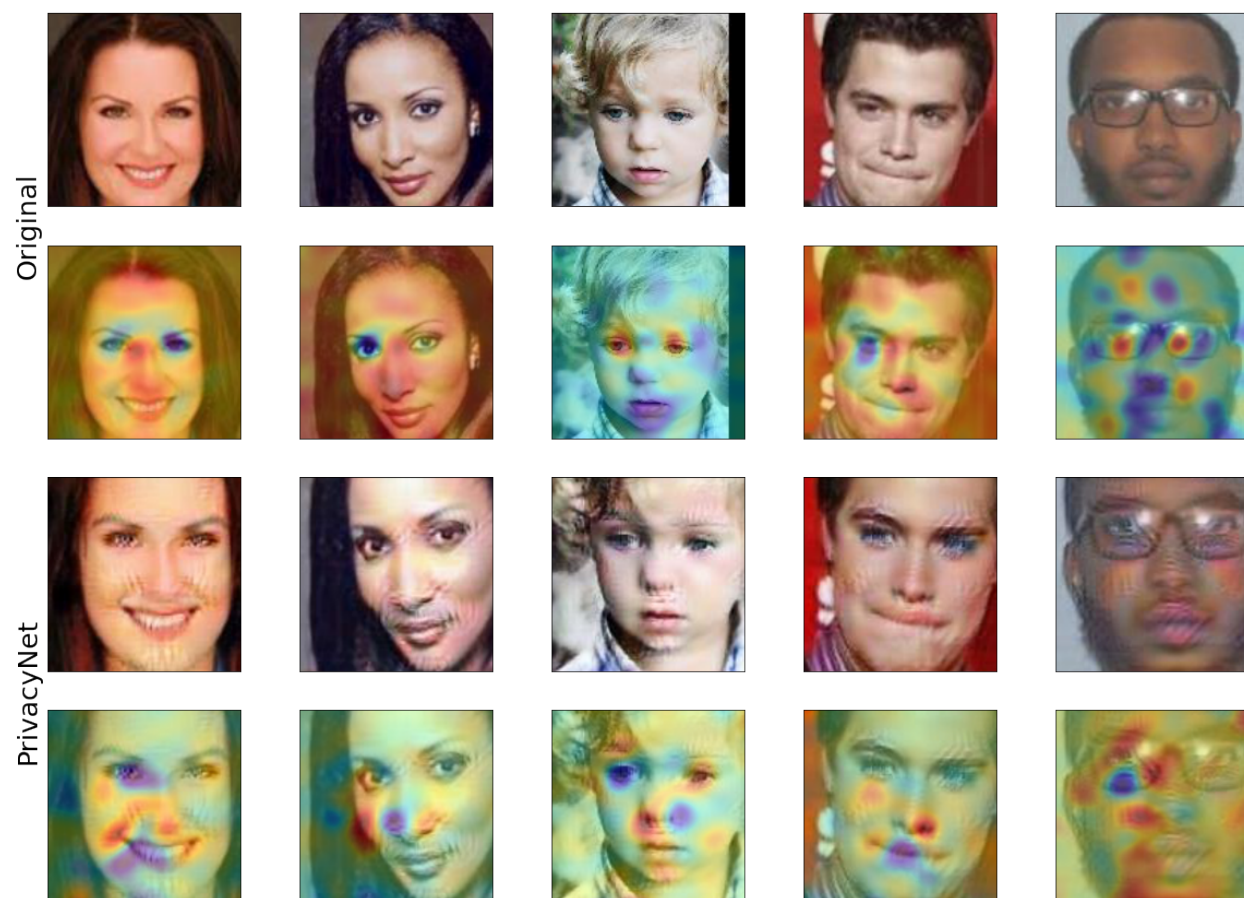


Figure 7.20: Visualizing detected features for gender classification using Grad-CAM [128] before (original images) and after SAN perturbations (PrivacyNet outputs) on some samples in UTK-face dataset.

Chapter 8

Summary and Conclusions

Face images have been widely used for recognition purposes, which is to recognize the person in the image. However, besides the recognition, face images can be used to extract demographic information. On the other hand, recent privacy regulations such as European Union General Data Protection and Regulation (EU-GDPR) restricts the use of personal data of data subjects for purposes beyond the specific purpose for which the data is collected and the consent from data subjects is acquired. As a result, it is important to provide means to avoid the extraction of demographic information from face images, when such images are only collected for recognition purposes. This study was focused on imparting multi-attribute demographic privacy to face images, such that gender, age and race cannot be reliably extracted from face images, while not adversely affecting the recognition utility of such data.

8.1 Research contributions and main findings of this work

In Chapter 2, we consider the problem of deriving additive perturbations based on a face-based given gender classifier, and provide the following contributions:

- Designed an efficient gradient descent-based technique that iteratively adds perturbations to a face image in order to confound gender information. At the same, the matching performance of the perturbed face images was retained.

- Performed extensive experiments to show that the proposed method is applicable to any gender classifier in a black-box scenario.

Chapter 3 presents addresses the problem related to transferability of perturbation to unseen gender classifiers. The contributions of this chapter are

- Designed a neural network model called Semi-Adversarial Networks (SAN) that is trained with an auxiliary gender classifier and an auxiliary face matcher for imparting gender privacy to face images.
- The SAN model learns to derive perturbations that are transferable to unseen gender classifiers.

In Chapter 4, we focused on the generalizability issue of the SAN model, which is to ensure that the perturbed face images are able to confound *arbitrary unseen* gender classifiers. The contributions of this chapter are as follows:

- Designed an ensemble SAN model that trains multiple SAN models for addressing the generalizability issue, where each SAN model is trained using an auxiliary gender classifier.
- Enhancing diversity among the SAN models in the ensemble by oversampling randomly selected samples, which leads to enhancing diversity in auxiliary gender classifiers, and thereby, improving the generalizability of the ensemble SAN model.

Chapter 5 introduces Stack-SAN for enhancing the generalizability of the SAN model, with the following contributions:

- Designed a model that sequentially trains and stacks SAN models such that the output from the SAN model i is fed as input to the SAN model $i + 1$.

- Performed extensive experiments that showed gender classification performance decreases progressively as the input goes through the stack of SAN models.

Chapter 6 extends the SAN model to multi-attribute including gender, age, and race. The contributions are the following:

- Designed a multi-attribute face privacy model called PrivacyNet using Generative Adversarial Networks (GAN) that is able to modify input face images such that gender, age, and race can be selectively confounded, while the matching utility is retained.
- Conducted extensive experiments using multiple attribute classifiers and face matchers, and showed the efficacy of the proposed model on multiple datasets.

Chapter 7 addresses two research problem related to how the perturbations made by SAN are perceived by human observers.

- Designed an experiment where a subset of face images were shown to human observers using Amazon MTurk platform, where the participants were asked to classify the gender of the person shown in the image.
- The designed experiment confirmed that the perturbations made by the SAN model were able to fool human observers.

8.2 Limitations of this study

PrivacyNet model was shown to be able to successfully confound gender and race from face images in a selective manner; however, its performance with regard to age showed it to be less satisfactory than the other two attributes. In this regard, we observed the following:

- The performance of modifying age was observed to be biased, as the performance of modifying the age-group of a face image from $A0$ to $A2$ was higher than changing it from $A0$ to $A1$.
- In some cases where the age-group of a subject was not meant to be modified, we observed certain cases appear younger by visual inspection.

We hypothesize that the first issue is caused since we have used an age-group classifier, while age-group is an ordinal attribute, i.e., there is a natural order between the values: $A0 < A1 < A2$. This issue can be addressed by replacing the auxiliary age-group classifier with an ordinal regression model.

The second issue is due to the way the boundaries of the age-groups are defined. Face aging has shown non-linear properties, causing more variations in certain age-ranges than others. In this study, we did not optimize the definition of age-groups; however, finding optimal age-groups can address this issue.

8.3 Recommendations and future work

Based on the limitations of the work mentioned in the previous section regarding the performance of PrivacyNet on modifying or preserving age-group of subjects, I recommend using an ordinal regression model that can reliably determine the age or age-group of a subject, and use that as an auxiliary network for training the PrivacyNet model.

In addition, while this study has focused on face images, extending demographic attribute privacy to other biometric modalities, such as finger print and iris is necessary. Further, a controllable model for imparting various degrees of privacy to *individual* subjects is needed.

BIBLIOGRAPHY

BIBLIOGRAPHY

- [1] “AB-375 Privacy: personal information: businesses”. In: California Legislative Information (2018).
- [2] Mart’ın Abadi et al. “Tensorflow: Large-scale machine learning on heterogeneous distributed systems”. In: arXiv preprint arXiv:1603.04467 (2016).
- [3] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. “Privacy and human behavior in the age of information”. In: *Science* 347.6221 (2015), pp. 509–514.
- [4] Alessandro Acquisti and Ralph Gross. “Predicting Social Security numbers from public data”. In: *Proceedings of the National Academy of Sciences* 106.27 (2009), pp. 10975–10980.
- [5] Alessandro Acquisti, Leslie K. John, and George Loewenstein. “What is privacy worth?” In: *The Journal of Legal Studies* 42.2 (2013), pp. 249–274.
- [6] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. “The economics of privacy”. In: *Journal of Economic Literature* 54.2 (2016), pp. 442–92.
- [7] Prachi Agrawal and PJ Narayanan. “Person de-identification in videos”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 21.3 (2011), pp. 299–310.
- [8] Naveed Akhtar and Ajmal Mian. “Threat of adversarial attacks on deep learning in computer vision: A survey”. In: *IEEE Access* 6 (2018), pp. 14410–14430.
- [9] Alexander Amini et al. “Uncovering and mitigating algorithmic bias through learned latent structure”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 289–295.
- [10] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. “OpenFace: A general purpose face recognition library with mobile applications”. In: *CMU School of Computer Science* (2016).
- [11] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. “Face aging with conditional generative adversarial networks”. In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 2089–2093.
- [12] Martin Arjovsky, Soumith Chintala, and L’eon Bottou. “Wasserstein gan”. In: arXiv preprint arXiv:1701.07875 (2017).
- [13] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Pl’oger. “Real-time Convolutional Neural Networks for Emotion and Gender Classification”. In: arXiv preprint arXiv:1710.07557 (2017).

- [14] Suryanti Awang et al. “Feature level fusion of face and signature using a modified feature selection technique”. In: 2013 International Conference on Signal-Image Technology & Internet-Based Systems. IEEE. 2013, pp. 706–713.
- [15] Christopher Beckham and Christopher Pal. “A simple squared-error reformulation for ordinal classification”. In: arXiv preprint arXiv:1612.00775 (2016).
- [16] Yoshua Bengio. “Learning deep architectures for AI”. In: Foundations and trends® in Machine Learning 2.1 (2009), pp. 1–127.
- [17] Dmitri Bitouk et al. “Face swapping: automatically replacing faces in photographs”. In: ACM Transactions on Graphics (TOG) 27.3 (2008), p. 39.
- [18] Anna Katarzyna Bobak, Andrew James Dowsett, and Sarah Bate. “Solving the border control problem: Evidence of enhanced face matching in individuals with extraordinary face recognition skills”. In: PloS one 11.2 (2016), e0148148.
- [19] Denton Bobeldyk and Arun Ross. “Predicting Eye Color from Near Infrared Iris Images”. In: International Conference on Biometrics (ICB). IEEE. 2018, pp. 104–110.
- [20] Denton Bobeldyk and Arun Ross. “Predicting Soft Biometric Attributes from 30 Pixels: A Case Study in NIR Ocular Images”. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE. 2019, pp. 116–124.
- [21] Michael Boyle, Christopher Edwards, and Saul Greenberg. “The effects of filtered video on awareness and privacy”. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work. 2000, pp. 1–10. ISBN: 1-58113-222-0.
- [22] Joy Buolamwini and Timnit Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. In: Conference on fairness, accountability and transparency. 2018, pp. 77–91.
- [23] Q. Cao et al. “VGGFace2: A dataset for recognising faces across pose and age”. In: International Conference on Automatic Face and Gesture Recognition. 2018.
- [24] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. “Consistent Rank Logits for Ordinal Regression with Convolutional Neural Networks”. In: arXiv preprint arXiv:1901.07884 (2019).
- [25] Zhangjie Cao et al. “Partial Transfer Learning with Selective Adversarial Networks”. In: arXiv preprint arXiv:1707.07901 (2017).
- [26] Aniello Castiglione et al. “Biometrics in the cloud: challenges and research opportunities”. In: IEEE Cloud Computing 4.4 (2017), pp. 12–17.
- [27] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. “Cross-age reference coding for age-invariant face recognition and retrieval”. In: European Conference on Computer Vision. Springer. 2014, pp. 768–783.

- [28] Irene Chen, Fredrik D Johansson, and David Sontag. “Why is my classifier discriminatory?” In: *Advances in Neural Information Processing Systems*. 2018, pp. 3539–3550.
- [29] Jun-Cheng Chen et al. “A cascaded convolutional neural network for age estimation of unconstrained faces”. In: *Proceedings of the IEEE Conference on Biometrics Theory, Applications and Systems*. 2016, pp. 1–8.
- [30] Shixing Chen et al. “Using Ranking-CNN for age estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5183–5192.
- [31] Saheb Chhabra et al. “Anonymizing k-Facial Attributes via Adversarial Perturbations”. In: *arXiv preprint arXiv:1805.09380* (2018).
- [32] Donald G Childers and Ke Wu. “Gender recognition from speech. Part II: Fine analysis”. In: *The Journal of the Acoustical society of America* 90.4 (1991), pp. 1841–1856.
- [33] Y. Choi et al. “StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8789–8797. DOI: 10.1109/CVPR.2018.00916.
- [34] Timothy F. Cootes et al. “Active shape models-their training and application”. In: *Computer Vision and Image Understanding* 61.1 (1995), pp. 38–59.
- [35] Naser Damer et al. “MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network”. In: *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE. 2018, pp. 1–10.
- [36] Antitza Dantcheva, Petros Elia, and Arun Ross. “What else does your biometric data reveal? A survey on soft biometrics”. In: *IEEE Transactions on Information Forensics and Security* 11.3 (2016), pp. 441–467.
- [37] Antitza Dantcheva et al. “Bag of soft biometrics for person identification”. In: *Multimedia Tools and Applications* 51.2 (2011), pp. 739–777.
- [38] Brian DeCann and Arun Ross. “Relating ROC and CMC curves via the biometric menagerie”. In: *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on*. IEEE. 2013, pp. 1–8.
- [39] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *International Workshop on Multiple Classifier Systems*. Springer. 2000, pp. 1–15.
- [40] Pawel Drozdowski et al. “Demographic bias in biometrics: A survey on an emerging challenge”. In: *IEEE Transactions on Technology and Society* (2020).

- [41] Liang Du et al. “GARP-face: Balancing privacy protection and utility preservation in face de-identification”. In: IEEE International Joint Conference on Biometrics. 2014, pp. 1–8.
- [42] Cynthia Dwork et al. “Exposed! a survey of attacks on private data”. In: Annual Reviews (2017).
- [43] Sarah Evans, Nick Neave, and Delia Wakelin. “Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice”. In: Biological Psychology 72.2 (2006), pp. 160 –163. ISSN: 0301-0511.
- [44] Pedro F. Felzenszwalb, Ross B. Girshick, and David McAllester. “Cascade object detection with deformable part models”. In: Computer vision and pattern recognition (CVPR). 2010, pp. 2241–2248. ISBN: 1-4244-6985-6.
- [45] Ferdinando Fioretto and Pascal Van Hentenryck. “Privacy-Preserving Federated Data Sharing”. In: Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS). 2019, pp. 638–646.
- [46] Gian Luca Foresti et al. “Face detection for visual surveillance”. In: 12th International Conference on Image Analysis and Processing, 2003. Proceedings. IEEE. 2003, pp. 115–120.
- [47] Xin Geng, Chao Yin, and Zhi-Hua Zhou. “Facial Age Estimation by Learning from Label Distributions”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 35.10 (2013), pp. 2401–2412.
- [48] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. “Automatic age estimation based on facial aging patterns”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 29.12 (2007), pp. 2234–2240.
- [49] P Gnanasivam and Dr S Muttan. “Fingerprint gender classification using wavelet transform and singular value decomposition”. In: arXiv preprint arXiv:1205.6745 (2012).
- [50] Beatrice A Golomb, David T Lawrence, and Terrence J Sejnowski. “Sexnet: A neural network identifies sex from human faces.” In: NIPS. Vol. 1. 1990, p. 2.
- [51] Sixue Gong, Xiaoming Liu, and A Jain. “Jointly de-biasing face recognition and demographic attribute estimation”. In: European Conference on Computer Vision (Virtual). 2020.
- [52] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
- [53] Ian Goodfellow et al. “Generative Adversarial Nets”. In: Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2014, pp. 2672–2680. (Visited on 04/27/2017).

- [54] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: arXiv preprint arXiv:1412.6572 (2014).
- [55] Shivanand Gornale, Abhijit Patil, and C Veersheety. “Fingerprint Based Gender Identification using Discrete Wavelet Transform and Gabor Filters”. In: International Journal of Computer Applications 975 (2016), p. 8887.
- [56] Ralph Gross. “Face de-identification using multi-factor active appearance models”. 2008.
- [57] Ralph Gross et al. “Face de-identification”. In: Protecting privacy in video surveillance. Springer, 2009, pp. 129–146.
- [58] Ralph Gross et al. “Integrating utility into face de-identification”. In: International Workshop on Privacy Enhancing Technologies. Springer, 2005, pp. 227–242.
- [59] Ralph Gross et al. “Model-based face de-identification”. In: Computer Vision and Pattern Recognition Workshop (CVPRW). 2006. ISBN: 0-7695-2646-2.
- [60] Manuel G˘unther, Andras Rozsa, and Terranee E Boulton. “AFFACT: Alignment-free facial attribute classification technique”. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). 2017, pp. 90–99.
- [61] Samta Gupta and A Prabhakar Rao. “Fingerprint Based Gender Classification using Discrete Wavelet Transform & Artificial Neural Network”. In: International Journal of Computer Science and Mobile Computing 3.4 (2014), pp. 1289–1296.
- [62] S. Gutta, H. Wechsler, and P. J. Phillips. “Gender and Ethnic Classification of Face Images”. In: International Conference on Automatic Face and Gesture Recognition. 1998, pp. 194–199. DOI: 10.1109/AFGR.1998.670948.
- [63] Lars Kai Hansen and Peter Salamon. “Neural network ensembles”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 12.10 (1990), pp. 993–1001.
- [64] Kaiming He et al. “Deep residual learning for image recognition”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016, pp. 770–778.
- [65] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: Science 313.5786 (2006), pp. 504–507.
- [66] Guosheng Hu et al. “Attribute-enhanced Face Recognition with Neural Tensor Fusion Networks”. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017, pp. 3744–3753.
- [67] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 7132–7141.

- [68] Gary Huang et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Tech. rep. 07-49. University of Massachusetts, Amherst, 2007.
- [69] P. Isola et al. “Image-to-Image Translation with Conditional Adversarial Networks”. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632.
- [70] Anil Jain, Arun A. Ross, and Karthik Nandakumar. Introduction to biometrics. Springer Science & Business Media, 2011. ISBN: 0-387-77326-6.
- [71] Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. “Right for the Right Reason: Training Agnostic Networks”. In: International Symposium on Intelligent Data Analysis. Springer. 2018, pp. 164–174.
- [72] Amin Jourabloo, Xi Yin, and Xiaoming Liu. “Attribute preserved face de-identification”. In: International Conference on Biometrics (ICB). 2015, pp. 278–285. ISBN: 2376-4201.
- [73] Behrooz Kamgar-Parsi, Wallace Lawson, and Behzad Kamgar-Parsi. “Toward development of a face recognition system for watchlist surveillance”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 33.10 (2011), pp. 1925–1937.
- [74] Jonathan Katz et al. Handbook of applied cryptography. CRC press, 1996.
- [75] Michael Kim, Omer Reingold, and Guy Rothblum. “Fairness through computationally bounded awareness”. In: Advances in Neural Information Processing Systems. 2018, pp. 4842–4852.
- [76] Taeksoo Kim et al. “Learning to discover cross-domain relations with generative adversarial networks”. In: arXiv preprint arXiv:1703.05192 (2017).
- [77] Davis E. King. “Dlib-ml: A Machine Learning Toolkit”. In: Journal of Machine Learning Research 10 (2009), pp. 1755–1758.
- [78] Günter Klambauer et al. “Self-normalizing neural networks”. In: Advances in Neural Information Processing Systems. 2017, pp. 972–981.
- [79] Pavel Korshunov and Touradj Ebrahimi. “Using face morphing to protect privacy”. In: 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance. IEEE. 2013, pp. 208–213.
- [80] Ludmila I Kuncheva. Combining pattern classifiers: methods and algorithms. John Wiley & Sons, 2004.
- [81] Taekyoung Kwon and Hyeonjoon Moon. “Biometric authentication for border control applications”. In: IEEE Transactions on Knowledge and Data Engineering 20.8 (2008), pp. 1091–1096.

- [82] Stephen Lagree and Kevin W Bowyer. “Predicting Ethnicity and Gender from Iris Texture”. In: International Conference on Technologies for Homeland Security (HST). IEEE. 2011, pp. 440–445.
- [83] Oliver Langner et al. “Presentation and validation of the Radboud Faces Database”. In: Cognition and Emotion 24.8 (2010), pp. 1377–1388.
- [84] Gil Levi and Tal Hassner. “Age and gender classification using convolutional neural networks”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2015, pp. 34–42.
- [85] Ling Li and Hsuan-Tien Lin. “Ordinal regression by extended binary classification”. In: Advances in neural information processing systems. 2007, pp. 865–872.
- [86] Ziwei Liu et al. “Deep learning face attributes in the wild”. In: Proceedings of the IEEE International Conference on Computer Vision. 2015, pp. 3730–3738.
- [87] Xiaoguang Lu and Anil K Jain. “Ethnicity Identification from Face Images”. In: Proceedings of SPIE. Vol. 5404. 2004, pp. 114–123.
- [88] Z. Lu et al. “Recent Progress of Face Image Synthesis”. In: 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR). 2017, pp. 7–12. DOI: 10.1109/ACPR.2017.2.
- [89] Erno Mäkinen and Roope Raisamo. “Evaluation of gender classification methods with automatically detected and aligned faces”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 30.3 (2008), pp. 541–547.
- [90] AM Martinez and R Benavente. “The AR face database”. In: (1998).
- [91] Iacopo Masi et al. “Deep face recognition: A survey”. In: 31st SIBGRAPI Conference on Graphics, Patterns and Images. 2018, pp. 471–478.
- [92] Blaz Meden, Peter Peer, and Vitomir Struc. “Selective Face Deidentification with End-to-End Perceptual Loss Learning”. In: 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOB). IEEE. 2018, pp. 1–7.
- [93] Blaz Meden et al. “k-Same-Net: k-Anonymity with Generative Deep Neural Networks for Face Deidentification”. In: Entropy 20.1 (2018), p. 60.
- [94] Stephen Milborrow, John Morkel, and Fred Nicolls. “The MUCT landmarked face database”. In: PRASA (2010).
- [95] Stephen Milborrow, John Morkel, and Fred Nicolls. “The MUCT landmarked face database”. In: Pattern Recognition Association of South Africa (2010).
- [96] Stephen Milborrow and Fred Nicolls. “Active shape models with SIFT descriptors and MARS”. In: International Conference on Computer Vision Theory and Applications (VISAPP). Vol. 2. IEEE, 2014, pp. 380–387. ISBN: 989-758-133-2.

- [97] V. Mirjalili, S. Raschka, and A. Ross. “PrivacyNet: Semi-Adversarial Networks for Multi-Attribute Face Privacy”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 9400–9412.
- [98] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. “FlowSAN: privacy-enhancing semi adversarial networks to confound arbitrary face-based gender classifiers”. In: *IEEE Access* 7 (2019), pp. 99735–99745.
- [99] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. “Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers”. In: *Proc. of 9th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. Los Angeles, CA, 2018.
- [100] Vahid Mirjalili and Arun Ross. “Soft Biometric Privacy: Retaining Biometric Utility of Face Images while Perturbing Gender”. In: *Proc. of International Joint Conference on Biometrics (IJCB)*. 2017.
- [101] Vahid Mirjalili et al. “Semi-Adversarial Networks: Convolutional autoencoders for imparting privacy to face images”. In: *Proc. of 11th IAPR International Conference on Biometrics (ICB 2018)*. IEEE. Gold Coast, Australia, 2018.
- [102] Konda Reddy Mopuri, Utsav Garg, and R Venkatesh Babu. “CNN fixations: an unraveling approach to visualize the discriminative image regions”. In: *IEEE Transactions on Image Processing* 28.5 (2018), pp. 2116–2125.
- [103] Aythami Morales et al. “SensitiveNets: Learning agnostic representations with application to face images”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- [104] Iynkaran Natgunanathan et al. “Protection of Privacy in Biometric Data”. In: *IEEE Access* 4 (2016), pp. 880–892.
- [105] Elaine M. Newton, Latanya Sweeney, and Bradley Malin. “Preserving privacy by deidentifying face images”. In: *IEEE Transactions on Knowledge and Data Engineering* 17.2 (2005), pp. 232–243.
- [106] Zhenxing Niu et al. “Ordinal regression with multiple output CNN for age estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 4920–4928.
- [107] Asem Othman and Arun Ross. “Privacy of facial soft biometrics: Suppressing gender but retaining identity”. In: *European Conference on Computer Vision Workshop*. Springer, 2014, pp. 682–696.
- [108] Nicolas Papernot et al. “The limitations of deep learning in adversarial settings”. In: *IEEE European Symposium on Security and Privacy (EuroS&P)*. 2016, pp. 372–387.

- [109] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. “Deep face recognition”. In: British Machine Vision Conference. Vol. 1. 2015.
- [110] Adam Paszke et al. “Automatic differentiation in PyTorch”. In: NIPS-W. 2017.
- [111] Padma Polash Paul and Marina Gavrilova. “Multimodal cancelable biometrics”. In: 11th International Conference on Cognitive Informatics and Cognitive Computing. IEEE. 2012, pp. 43–49.
- [112] Padma Polash Paul and Marina Gavrilova. “Rank level fusion of multimodal cancelable biometrics”. In: 2014 IEEE 13th International Conference on Cognitive Informatics and Cognitive Computing. IEEE. 2014, pp. 80–87.
- [113] Claudio Perez et al. “Gender classification from face images using mutual information and feature fusion”. In: International Journal of Optomechatronics 6.1 (2012), pp. 92–119.
- [114] Marianna Pronobis et al. Analysis of F0 and cepstral features for robust automatic gender recognition. Tech. rep. Idiap, 2009.
- [115] PROPUBLICA. Facebook Is Letting Job Advertisers Target Only Men. Accessed: 2019-05-16. 2018. (Visited on 09/18/2019).
- [116] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. “HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2017).
- [117] Sebastian Raschka and Vahid Mirjalili. Python Machine Learning, 2nd Ed. Birmingham, UK: Packt Publishing, 2017, p. 602. ISBN: 978-1787125933.
- [118] Nalini K. Ratha, Jonathan H. Connell, and Ruud M. Bolle. “Enhancing security and privacy in biometrics-based authentication systems”. In: IBM Systems Journal 40.3 (2001), pp. 614–634.
- [119] Ajita Rattani, Cunjian Chen, and Arun Ross. “Evaluation of Texture Descriptors for Automated Gender Estimation from Fingerprints”. In: European Conference on Computer Vision (ECCV). Springer. 2014, pp. 764–777.
- [120] “Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016”. In: Official Journal of the European Union L 119 (2016).
- [121] DA Reid et al. “Soft biometrics for surveillance: an overview”. In: Handbook of statistics. Vol. 31. Elsevier, 2013, pp. 327–352.
- [122] Karl Ricanek and Tamirat Tesafaye. “MORPH: A longitudinal image database of normal adult age-progression”. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR). IEEE. 2006, pp. 341–345.

- [123] Arun Ross et al. “Some Research Problems in Biometrics: The Future Beckons”. In: 2019 International Conference on Biometrics (ICB), Crete, Greece, 2019, pp. 1-8.
- [124] Rasmus Rothe, Radu Timofte, and Luc Van Gool. “Deep expectation of real and apparent age from a single image without facial landmarks”. In: International Journal of Computer Vision 126.2-4 (2018), pp. 144–157.
- [125] Duncan A. Rowland and David I. Perrett. “Manipulating facial appearance through shape and color”. In: IEEE Computer Graphics and Applications 15.5 (1995), pp. 70–76.
- [126] Andras Rozsa et al. “Are facial attributes adversarially robust?” In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE. 2016, pp. 3121–3127.
- [127] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 815–823.
- [128] Ramprasaath R Selvaraju et al. “Grad-CAM: Visual explanations from deep networks via gradient-based localization”. In: Proceedings of the IEEE International Conference on Computer Vision. 2017, pp. 618–626.
- [129] S. R. Shinde and S. D. Thepade. “Gender Classification with KNN by Extraction of Haar Wavelet Features from Canny Shape Fingerprints”. In: International Conference on Information Processing. 2015, pp. 702–707. DOI: 10.1109/INFOP.2015.7489473.
- [130] Reza Shokri et al. “Membership inference attacks against machine learning models”. In: IEEE Symposium on Security and Privacy (SP). 2017, pp. 3–18.
- [131] Terence Sim and Li Zhang. “Controllable face privacy”. In: 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG). Vol. 4. 2015, pp. 1–8. ISBN: 1-4799-6026-8.
- [132] Thilo Strauss et al. “Ensemble methods as a defense to adversarial perturbations against deep neural networks”. In: arXiv preprint arXiv:1709.03423 (2017).
- [133] Jinli Suo et al. “High-resolution face fusion for gender conversion”. In: IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41.2 (2011), pp. 226–237.
- [134] Latanya Sweeney. “k-anonymity: A model for protecting privacy”. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.5 (2002), pp. 557–570.
- [135] Christian Szegedy et al. “Intriguing properties of neural networks”. In: arXiv preprint arXiv:1312.6199 (2013).
- [136] Juan Tapia and Carlos Aravena. “Gender Classification from NIR Iris Images using Deep Learning”. In: Deep Learning for Biometrics. Springer, 2017, pp. 219–239.

- [137] Philipp Terhörst et al. “PE-MIU: A Training-Free Privacy-Enhancing Face Recognition Approach Based on Minimum Information Units”. In: IEEE Access (2020).
- [138] Philipp Terhörst et al. “Unsupervised Enhancement of Soft-biometric Privacy with Negative Face Recognition”. In: arXiv preprint arXiv:2002.09181 (2020).
- [139] Philipp Terhörst et al. “Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations”. In: Applied Intelligence (2019).
- [140] Vince Thomas et al. “Learning to Predict Gender from Iris Images”. In: Proceedings of First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS). IEEE. 2007, pp. 1–5.
- [141] B. P. Tiddeman, M. R. Stirrat, and D. I. Perrett. “Towards realism in facial prototyping: results of a wavelet MRF method”. In: Proc. Theory and Practice of Computer Graphics. Vol. 1. 2006, pp. 20–30.
- [142] Fernando De la Torre et al. “Intraface”. In: 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG). Vol. 1. 2015, pp. 1–8. ISBN: 1-4799-6026-8.
- [143] Luan Tran, Xi Yin, and Xiaoming Liu. “Disentangled representation learning GAN for pose-invariant face recognition”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
- [144] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. “Suspicious human activity recognition: a review”. In: Artificial Intelligence Review 50.2 (2018), pp. 283–339.
- [145] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. “Instance Normalization: The Missing Ingredient for Fast Stylization”. In: CoRR abs/1607.08022 (2016).
- [146] S Vasuhi et al. “An efficient multi-modal biometric person authentication system using fuzzy logic”. In: ICoAC 2010, pp. 74–81.
- [147] L Walavalkar et al. “Support vector learning for gender classification using audio and visual cues”. In: International Journal of Pattern Recognition and Artificial Intelligence 17.03 (2003), pp. 417–439.
- [148] Mei Wang and Weihong Deng. “Deep face recognition: a survey”. In: arXiv preprint arXiv:1804.06655 (2018).
- [149] Qizheng Wang et al. “Face detection for privacy protected images”. In: IEEE Access 7 (2019), pp. 3918–3927.
- [150] Lingyun Wen and Guodong Guo. “A computational approach to body mass index prediction from face images”. In: Image and Vision Computing 31.5 (2013), pp. 392–400.

- [151] Ke Wu and Donald G Childers. “Gender recognition from speech. Part I: Coarse analysis”. In: The journal of the Acoustical Society of America 90.4 (1991), pp. 1828–1840.
- [152] Yifan Wu et al. “Privacy-Protective-GAN for Privacy Preserving Face De-Identification”. In: Journal of Computer Science and Technology 34.1 (2019), pp. 47–60.
- [153] Hao Zhang et al. “On the Effectiveness of Soft Biometrics for Increasing Face Verification Rates”. In: Computer Vision and Image Understanding 137 (2015), pp. 50–62.
- [154] Song Yang Zhang Zhifei and Hairong Qi. “Age Progression/Regression by Conditional Adversarial Autoencoder”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
- [155] J. Zhu et al. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: 2017 IEEE International Conference on Computer Vision (ICCV). 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.
- [156] James Zou and Londa Schiebinger. “AI can be sexist and racist—it’s time to make it fair”. (2018).