Mitigating Bias in Facial Recognition with FairGAN

Sasha Harrison
Stanford University
aharris6@stanford.edu

Boxiao Pan
Stanford University
bxpan@stanford.edu

Abstract

Algorithmic bias has long been a well established phenomenon in computer vision. Previous work by Buolamwini and Gebru has shown that numerous popular facial recognition algorithms that perform gender classification demonstrate performance gaps among individuals of different races [1]. Building off of this work, we seek to neutralize these biases using a data preprocessing approach rooted in Generative Adversarial Networks. We plan to reproduce the approach described in FairGAN [2] to produce images that are debiased with respect to race. In our experiments, we analyze the generated images qualitatively, and compare gender classifiers trained on the real dataset, and the synthetic dataset produced by FairGAN, respectively. We hope to show empirically that the synthetic dataset greatly reduces racial bias in a downstream gender classification task by reducing the performance gap for gender classification between light-skinned and dark skinned individuals. Although our preliminary results are largely inconclusive, we cite specific future steps to move this project closer towards its goal.

1 Introduction

In recent years, there has been increasing interest in the propensity of decision making algorithms to reflect social biases with respect to race, gender, and other protected characteristics. Given the increasing ubiquity of these algorithms, it is necessary to better understand the causes of these biased outcomes and to mitigate their impact on society. In *Gender Shades* [1], the authors demonstrated that facial recognition algorithms sold by Microsoft, Face++ and IBM show significant disparities in the accuracy of gender classification between light-skinned and dark-skinned individuals. Their investigation revealed all three algorithms correctly classified male faces at a higher rate than female faces, and lighter faces at a higher rate than darker faces. A 2015 report released by the National Institute of Standards and Technology (NIST) found similar disparities in facial recognition models [3]. NIST investigated the gender classification accuracy of six facial recognition algorithms (5 proprietary and 1 academic) finding that overall, females were misclassified more often then males.

This paper investigates Generative Adversarial Networks (GANs) as a technique for eliminating racial biases from image datasets. To do so, we build on the FairGAN architecture proposed by Xu, Yuan, and Wu in *FairGAN: Fairness-aware Generative Adversarial Networks*. The FairGAN architecture adds an additional discriminator that penalizes correlation between the image and the protected characteristic (in our case, race) while maximizing the output's similarity to the training dataset. Thus, the goal is to generate a simulated dataset that is free of disparate impact.

In this project, we seek to apply the FairGAN architecture to the domain of image generation. We will apply FairGAN to the UTKFace dataset with the goal of producing a synthetic image dataset that removes correlation between generated face images and race. We hope to show that building a gender classifier based on this synthetic dataset will improve performance with respect to fairness metrics. We will measure the "fairness" of each classifier using the metrics detailed below. Our goals are twofold:

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning, Spring 2020.

(1) Demonstrate that FairGAN can generate detailed enough images to be applied as a data pre-processing method to image datasets.

(2) Show that this data pre-processing step can improve the fairness of a downstream facial recognition classifier.

If successful, this approach demonstrates that by eliminating correlations between race label and the visual properties of the images, we can eliminate the potential for racial bias in any downstream supervised learning task.

2 Related Work

The GAN architecture that serves as the basis for this paper was originally proposed in *FairGAN: Fairness-aware Generative Adversarial Networks*, which was one of the early works to use GANs as a method for fair data generation. Xu et. al. demonstrated the empirical effectiveness of FairGAN on the UCI Adult Income Dataset. They showed that a classifier predicting income level showed significant improvement in equality of opportunity, and only a small decrease in accuracy, when trained on the synthetic dataset generated by FairGAN. We expand on their work by demonstrating that the same problem formulation can be used on images, a data format that is more complex and rich in latent structure. Their 3 player adversarial architecture is the basis of our final image generating architecture.

GANs constitute a particularly impactful class of models in the field of algorithmic fairness because they are compatible with several different statistical fairness definitions. In [4], authors Beutel et. al. demonstrate not only that adversarial training can remove sensitive information from latent representations learned by neural networks, but also that the choice of data for adversarial training can affect the fairness properties of the result, and that GANs can be used to satisfy both demographic parity and equality of odds. All and all, this work demonstrates that the possible applications of adversarial networks are far reaching. Although we use the same fairness regularization strategy as in Xu et. al. we are excited about this approach because of its ability to adhere to numerous fairness definitions applicable to a wide range of problems.

We are not the first researchers to attempt to generate a fair image dataset using GANs. In *Fairness GAN*, authors Sattigeri et. al. use GANs to generate image datasets to satisfy demographic parity and equality of opportunity. Instead of using two binary adversaries as in [2], they instead use a structure inspired by AC-GAN in which they incorporate conditioned class prediction and the fairness constraint as secondary tasks for the existing discriminator. This single discriminator has four outputs, including a probability distribution over the source of the image (real or fake) and the probability distribution over the protected attribute given the image/label pair. Thus, although this paper's goal is similar to that of FairGAN, the architectural approach is quite different. We implemented both approaches, finding that qualitatively, they produced similar results.

3 Dataset and Evaluation Metrics

3.1 Dataset

We plan to use UTKFace dataset [5], which consists of approximately 25,000 face images labeled with age, gender, and racial identity of the subject. Specifically for race, the racial categories in the dataset are White, Black, Asian, Indian, and Others, which includes Hispanic, Latino, and Middle Eastern. We show several data samples in Fig. 1.

Figure 1: Data samples from UTKFace[5] with individuals having the same gender and age, but with different races. From left to right: White, Black, Asian, Indian, and Others.

In order to use the same analytical categories as in [1], we approximate a binary racial label by grouping images under white and non-white groups. We acknowledge that this is an inherently flawed approach; not only is race not commonly understood to be binary, but also is a social category that shifts over time and geographic region. Given more resources, we believe a more sound approach would be to reduce race to a fitzpatrick skin type (as in [1] and [6]), therefore working with an objective characteristic of face image. Despite these drawbacks, we still believe this approach has value as a proof of concept that can later be applied to more complex settings.

3.2 Fairness Metrics

Assuming a binary protected attribute S and an allocative decision variable Y (where $Y = 1$ represents a positive outcome and $Y = 0$ represents a negative outcome), a common metric to measure algorithmic fairness is equality of opportunity [4]. Equality of opportunity is given by

$$P(\hat{Y} = 1|S = 1, Y = 1) = P(\hat{Y} = 1|S = 0, Y = 1)$$

which enforces that qualified people should have equal probability of being granted the allocative decision regardless of demographic group.

The pattern recognized by NIST, as well as Buolamwini and Gebru, was that gender classification accuracy worsened for marginalized groups. In this case, the predicted variable y represents gender, which is not an allocative decision: the categories $y = 1$ and $y = 0$ should not be viewed as good and bad. Instead, we want a generalized notion of performance for each intersectional category. To do so, we will use PPV (Positive Predictive Value) as an indication of classifier performance for each of the demographic groups, which is consistent with the methodology presented by Buolamwini and Gebru [1]. It is important to note that positive predictive value is closely related to the equality of opportunity metric defined above.

In general, PPV is given by $\frac{TP}{TP+FN}$ where TP is true positives and FN is false negatives. For a given analytical category, determined by a unique pair ($Y = y$, $S = s$), we define the intra-group PPV as

$$PPV_{s,y} = P(\hat{Y} = y|S = s, Y = y) = \frac{TP}{TP + FN}$$

Ultimately, we will measure the bias of the classifier by the **performance gap**, meaning the absolute difference between the best intra-group PPV and the worst intra-group PPV.

$$pgap = |\max_{s,y} PPV_{s,y} - \min_{s,y} PPV_{s,y}|$$

The smaller the performance gap, the more fair the classifier becomes.

Then, we will use a second fairness metric which introduces the notion of statistical parity in a labeled dataset. In the generated dataset, the gender labels $\hat{y}$ are outputs of the generator network. We want to ensure that the generator G creates a balanced dataset with respect to the generated label y and it's interactions with the protected attribute s. Inspired by [2], we introduce the following metric to measure fairness in a synthetic dataset:

$$P(y = 1|s = 1) = P(y = 1|s = 0)$$

Problem Statement

In this paper, we seek to expand the formulation in [2] to image generation. Following their methods, our problem formulation is as follows:

Given a dataset $\{X, Y, S\} \sim P_{data}$, we aim to generate a synthetic dataset $\{\hat{X}, \hat{Y}, \hat{S}\} \sim P_G$ which achieves equality of opportunity with respect to the protected attribute $\hat{S}$. In order to do so, we use a GAN problem formulation, consisting of a generator network to create fake images, and

two discriminator networks, the first aimed at detecting whether input images are real or fake, and the second to penalize correlation between the synthetic images generated by G and the protected attribute S.

4 Methods

4.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) constitute the state of the art method for image generation. They are composed of two parts: a generator G and a discriminator D that are both deep neural networks. G generates fake samples given a prior distribution P_z and learns the distribution P_G with the goal of matching the data distribution P_{data}. The discriminator network D is a binary classifier that predicts whether a given sample is real or fake.

The objective function for D is the following:

$$\max_D \mathbb{E}_{X\sim P_{data}}\Big(\log D(x)\Big) + \mathbb{E}_{Z\sim P_z}\Big(\log(1-D(G(z)))\Big)$$

The discriminator provides the objective for the generator in the sense that G is trained to fool the discriminator, and aims to maximize the probability that D predicts that its output, $G(z)$, is real data. Thus, the objective function for G is as follows:

$$\min_G \mathbb{E}_{Z\sim P_z}\Big(\log(1-D(G(z)))\Big)$$

Overall, a GAN is formulated as a minimax game $\min_G \max_D V(D,G)$ where V is the objective function:

$$V(G,D) = \mathbb{E}_{X\sim P_{data}}\Big(\log D(x)\Big) + \mathbb{E}_{z\sim P_z}\Big(\log(1-D(G(z)))\Big)$$

This loss function is minimized when $P_G = P_{data}$. In general, generative adversarial networks minimize the Jensen-Shannon divergence between the distributions P_G and P_{data}. Qualitatively, we expect that over the course of training, the G and D networks get progressively better at their jobs such that generated images grow more and more to resemble input images from the dataset.

4.2 FairGAN for Images

Model

The architecture of FairGAN builds off of the GAN framework presented above, but incorporates both the protected attribute and discrete class labels into the generator and discriminator interfaces. FairGAN consists of one generator G and two discriminators D_1 and D_2. G generates a fake pair $(\hat{x},\hat{y})$ according to the distribution $P_G(x,y|s)$. Formally, the input/output behavior of G is $(\hat{x},\hat{y}) = G(z,s), z\sim P_z(z)$, where z is a noise vector generated from a prior distribution. As in a regular GAN, the discriminator D_1 is trained to distinguish real data $P_{data}(x,y,s)$ from fake, generated data $P_G(x,y,s)$.

Simultaneously, in order to generate fake data that is fair, we want to rid $P_G(x,y,s)$ of correlation between $(\hat{x},\hat{y})$. To do this, the second discriminator $D_2$is trained to distinguish between the two demographic groups $P_G(x,y|s=1)$ and $P_G(x,y|s=0)$. This enforces the constraint that $P_G(x,y|s=1) = P_G(x,y|s=0)$, or in other words, that the pair $P_G(x,y) \perp s$

Overall, the objective function of this three-player game formulation is:

$$\min_G \max_{D_1,D_2} V(G,D_1,D_2) = V_1(G,D_1) + \lambda V_2(G,D_2)$$

where V_1 resembles the classic GAN objective function:

$$V_1(G, D_1) = \mathbb{E}_{s\sim P_{data}(S),(x,y)\sim P_{data}(x,y|s)}\Big(\log D_1(x, y, s)\Big)+$$

$$\mathbb{E}_{\hat{s}\sim P_{data}(S),(\hat{x},\hat{y})\sim P_G(x,y|s)}\Big(\log(1 - D_1(\hat{x}, \hat{y}, \hat{s}))\Big)$$

and V_2 corresponds to the minimax game between the generator G and fairness discriminator, D_2:

$$V_2(G, D_2) = \mathbb{E}_{(\hat{x},\hat{y})\sim P_G(x,y|s=1)}\Big(\log D_2(\hat{x}, \hat{y})\Big)+$$

$$\mathbb{E}_{(\hat{x},\hat{y})\sim P_G(x,y|s=0)}\Big(\log(1 - D_2(\hat{x}, \hat{y}))\Big)$$

A high level diagram of this model can be seen in Figure 6.

4.3 Fairness GAN

In addition to using the FairGAN architecture to produce a debiased dataset, we also tried a second debiasing approach based on an AC-GAN formulation, similar to the approach used in *Fairness GAN* [6]. In an AC-GAN architecture, the task of producing generated images conditioned on a class label is treated differently than in the FairGAN formulation. Since we found the methodology of incorporating class labels was somewhat subjective, we were interested in comparing and contrasting these two approaches.

The generator outputs both a fake image and the associated gender label, $G(z, s) = (\hat{X}, \hat{y})$. The objective function for the generator combines two log-likelihoods for the correct source of the sample.

$$L^R_{S_J} = E\Big(\log P(S_j = 1|X, y)\Big), L^F_{S_J} = E\Big(\log P(S_j = 0|\hat{X}, \hat{y})\Big)$$

Corresponding to the joint distribution (X,Y), where 1 is the label for real images and 0 is the label for fake images. The second log likelihoods

$$L^R_{S_x} = E\Big(\log P(S_j = 1|X)\Big), L^F_{S_x} = E\Big(\log P(S_j = 0|\hat{X})\Big)$$

Pertain only to the images $X, \hat{X}$ without the class label y.

Then, as in an AC-GAN formulation, we include class-conditioned losses. According to the authors of AC-GAN, these class-conditioned losses add structure to the GAN optimization and improve the generation of plausible images:

$$L^R_C = E\Big(\log P(S = s|X)\Big), L^F_C = E\Big(\log P(S = s|\hat{X})\Big)$$

Finally, for fairness, we include a pair of losses to satisfy demographic parity:

$$L^R_{DP} = E\Big(\log P(C = c|y)\Big), L^F_{DP} = E\Big(\log P(C = c|\hat{y})\Big)$$

Overall, the discriminator maximizes:

$$L^D = L^R_{S_J} + L^F_{S_J} + L^R_{S_x} + L^F_{S_x} + L^R_C + L^R_{DP}$$

and the generator minimizes:

$$L^G = L^F_{S_J} + L^R_{S_x} - L^F_C + L^F_{DP}$$

5 Experiments

Baseline Gender Classifier

Gender Shades [1] establishes that commercial classifiers exhibit classification accuracy of 8.1% - 20.6% worse on female faces and 11.8% - 19.2% worse on darker faces [1]. As a baseline, we will fit a simple CNN-based classifier to take images from UTKFace as input and produce a gender label $\hat{y}$ as output. We expect from previous experiments done in this domain that this classifier will exhibit similar biases as those observed in [1].

As a baseline classifier, we trained a ResNet-18 model for 50 epochs, and the best model achieved a 88.5% classification accuracy on the validation set. We used a batch size of 128 and a learning rate of 1e-4. The PPV for each analytical group is as follows:

	PPV real
Non-white Female	0.853
White Female	0.816
Non-white Male	0.711
White Male	0.847

Table 1: PPV of RESNET-based baseline gender classifier

The worst performing grow is non-white males, while the best performing group is non-white females. Overall, these results are surprising in that we don't see a reliable trend that white, male examples have better results than dark or female examples.

The parity gap metric for these results is 0.142

3 Player FairGAN

To generate the fake image dataset, we implemented the final three player FairGAN structure described by Figure 6. We took inspiration from the opensource code at `https://github.com/znxlwm/pytorch-MNIST-CelebA-cGAN-cDCGAN`. Both the generator architectures and discriminator architectures are based on DCGAN. In the generator network, the fake images are generated from a noise vector z by a series of transposed convolutional layers that gradually enlarge the output. Similarly, the discriminator uses convolutional layers to downsample to image to output a probability distribution over the classes {real, fake}. To condition the images on the race label s, we include s as an input to both G and D. The label s is incorporated into the latent representations of these networks by passing the noise vector and binary label s through respective convolutional layers, then concatenating the result. The predicted gender label $\hat{y}$ is produced by passing this concatenated feature representation through two fully-connected layers.

Through experimentation, we found that the stability of training was highly influenced by the precise mechanisms used to generate $\hat{y}$ in the generator, and to incorporate $\hat{y}$ into the discriminator. Eventually, we settled on an architecture where $\hat{X}, \hat{y} = G(s, z)$: G takes the noise vector z and the race label s as input, and as output produces a fake image $\hat{X}$ and a gender label $\hat{y}$. The discriminator (called D1) predicting the source of the image takes the inputs $(\hat{x}, \hat{y}, \hat{s})$ and predicts whether the image is real or fake. The fairness discriminator (called D2) takes $(\hat{x}, \hat{y})$ and predicts $\hat{s}$, such that we penalize the weights of the generator if we can predict the racial label from the image.

Consistent with other GAN literature, we found it difficult to find a configuration that yielded a stable training procedure; we often encountered the problem that the discriminator D1 overpowered the generator network. D1's loss would quickly converge to zero, which in turn would result in neither of the networks being updated meaningfully. To combat training instability, we added noise to the groundtruth labels corresponding to real images such that the labels were in the range (0.9, 1.1). This prevents the discriminator loss from reaching zero, a failure mode in which the weights cease to be meaningfully updated since gradients vanish with zero loss. We also updated the discriminator less frequently than the generator by accumulating the gradients over 10 iterations before applying the optimizer. This resulted in less variance in losses over the training procedure. Lastly, we treated the predicted gender labels $\hat{y}$ as continuous instead of discrete in order to ensure better backpropagation through the weights of the generator network. We trained this network for 70 epochs with a learning

rate of 0.0002. We added the Fairness discriminator D2 at epoch 20, once the G and D1 networks were more stable. Overall, this experiment yielded face images that were comparable to the output of the conditional GAN baseline, and were visually reasonable faces. Sample output and graph of training losses can be seen in figures 2 and 3.

p(s = 1)	0.503
p(y = 1)	0.514
p(y = 1 | s = 0)	0.510
p(y = 1 | s = 1)	0.519

Table 2: Properties of generated dataset.

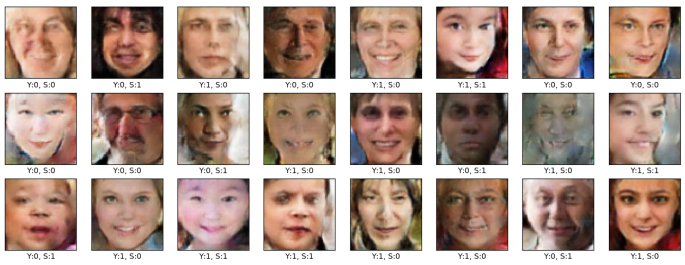

Figure 2: Generated faces from the final epoch of the FairGAN experiment

Over the course of training, we see a slight downward trend, and decreased variance, in the generator and discriminator losses. Table 2 shows that the generated balance is approximately balanced with respect to race and gender. For our four analytical categories, we have a 1:1:1:1 ratio.

Gender Classifier: FairGAN Synthetic Dataset

Using the synthetic dataset generated by the FairGAN experiment described above, we then trained a gender classifier to compare the results to the baseline. The training dataset consists entirely of generated images, similar to those in Figure 2. We evaluate the resulting classifier both on a held out test set of generated images (PPV Fake) and on the real test set from UTKFace (PPV real).

	PPV real	**PPV fake**
Non-white Female	0.743	0.900
White Female	0.779	0.923
Non-white Male	0.675	0.984
White Male	0.878	0.941

Table 3: PPV of the RESNET gender classifier on the four intersectional groups. Training data from FairGAN

On the real test set, the parity gap is 0.203. On the fake dataset, the parity gap is 0.084.

Fairness GAN

Next, we experimented with a Fairness GAN model. Unlike the FairGAN approach, the fairness GAN consists of the generator network and a single discriminator D, where D outputs three different binary class predictions. Similar to the experiment above, we trained this network for 70 epochs with a learning rate of 0.0002. The resulting images were similar in quality to those produced by FairGAN. Sample output and graph of training losses can be seen in figures 3 and 8.

Visually, the images have roughly the same quality as those produced by FairGAN.

Gender Classifier: Fairness GAN Synthetic Dataset

Using the same procedure as the experiments above, we generated a synthetic dataset of 25,000 training examples suing the generator from Fairness GAN. The generated dataset has the following composition:

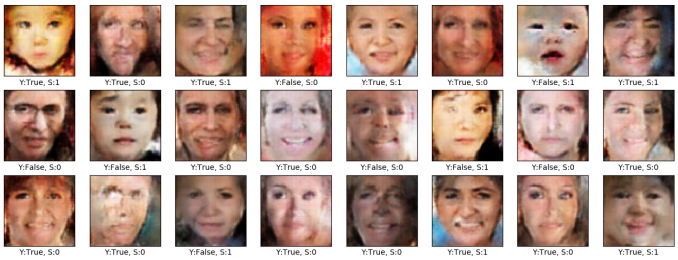

Figure 3: Generated faces from the final epoch of the Fairness GAN experiment

p(s = 1)	0.50
p(y = 1)	0.60
p(y = 1 | s = 0)	0.598
p(y = 1 | s = 1)	0.609

Table 4: Properties of Fairness GAN generated dataset.

There are a couple of interesting findings here. The proportion of white examples to non-white examples is exactly 50/50, which is consistent with drawing that label from a binomial(0.5) distribution. In both racial groups, about 60% of the examples are female, such that the training dataset overall contains an over-representation of women. In comparison to the dataset from FairGAN, this dataset is less balances with respect to gender.

We then trained a gender classifier for 50 epochs using the generated dataset from Fairness GAN, evaluating performance on both real and fake hold out sets.

	PPV real	**PPV fake**
Non-white Female	0.784	0.893
White Female	0.742	0.938
Non-white Male	0.691	0.834
White Male	0.874	0.874

Table 5: PPV of the resnet gender classifier on the four intersectional groups.

For the real data, the parity gap is 0.183. For the fake dataset, the parity gap is 0.104. The results on the fake dataset are lower than the comparable metrics from FairGAN, while the results on the real data from the two synthetic datasets seem comparable. With respect to parity gap, neither algorithm decreased the gap. on the real dataset in comparison to the baseline classifier. However, the parity gap on the fake test data is smaller than in the baseline gender classifier.

6 Discussion

Overall, the results of the FairGAN and Fairness GAN architectures both looked visually plausible. However, the numerical results from the gender classifiers built on the generated datasets showed that there is some room for improvement both with respect to overall PPV and with closing the parity gap (our measure of fairness). For the classifiers trained on synthetic data, the difference between overall PPV for the real and fake test datasets indicate that the generated distribution P_G to some extent approximates the real distribution P_{data} but that there is still room for improvement. This problem can be attributed to a number of factors, including low image resolution (64 x 64), the generator architecture not being deep enough to fully express P_{data}, or the discriminator network being too strong, resulting in less meaningful updates to the generator's parameters.

The parity gap did not improve on the real test dataset, but did improve on the fake test dataset, which matches the distribution of the synthetic training data. This is a hopeful result; if the synthetic data better approximated the real data, perhaps we would see this narrowed gap generalize to the real test dataset.

7 Future Work

In the future, we would hope to improve on the results reported in the experiments section by focusing on generator architectures that better approximate the real data distribution P_{data}.

We would also hope to produce higher resolution images (256 x 256), and introduce more numerical metrics to measure generator output. In particular, we would introduce the MS-SSIM metric to quantify the diversity of generated images. Since one well-documented failure mode of GANs is producing a low-diversity of output images, we hypothesize that measuring and improving $\hat{X}$ diversity would alleviate the problem of 100% training accuracy on synthetic datasets.

8 Appendix

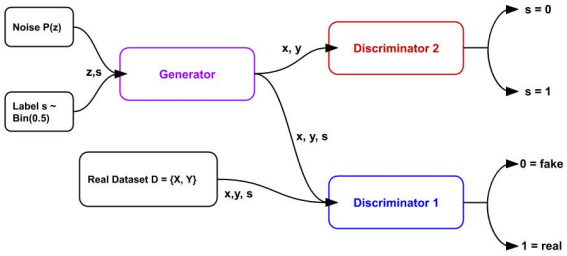

Figure 4: High level structure of fairGAN model

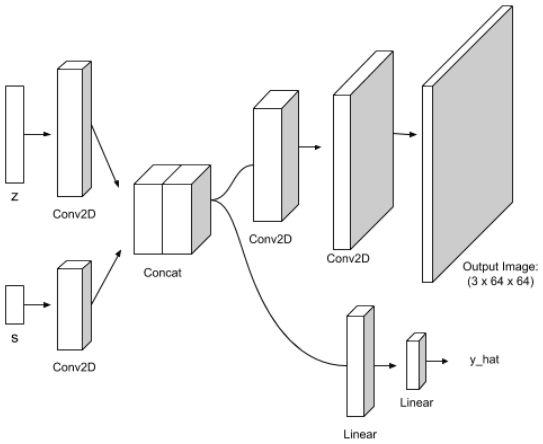

Figure 5: Architecture for the Generator network for the FairGAN experiment

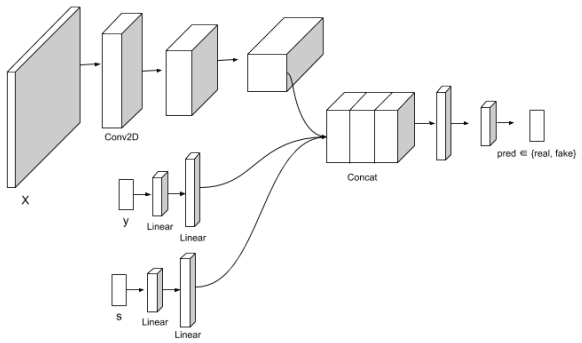

Figure 6: Architecture for the Discriminator network for FairGAN experiment

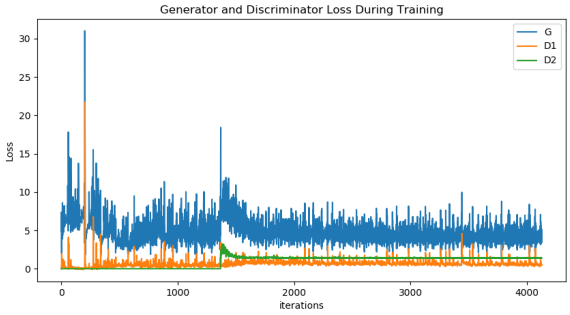

Figure 7: FairGAN experiment graph of training loss

	PPV real	PPV fake
Non-white Female	0.731	0.878
White Female	0.721	0.891
Non-white Male	0.628	0.862
White Male	0.772	0.9

Table 6: PPV of the Linear gender classifier on the four intersectional groups, trained on FAIRGAN synthetic dataset.

	PPV real	PPV fake
Non-white Female	0.732	0.878
White Female	0.722	0.892
Non-white Male	0.629	0.862
White Male	0.772	0.900

Table 7: PPV of the Linear gender classifier on the four intersectional groups, trained on Fairness-GAN produced synthetic dataset

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[2] Lu Zhang Depeng Xu, Shuhan Yuan and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *CoRR*, 2018.

[3] Mei Ngan, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

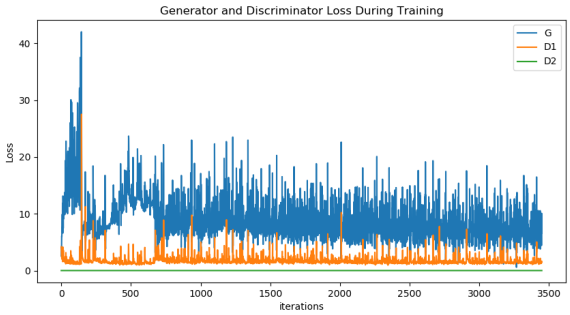

Figure 8: FairGAN experiment graph of training loss

[4] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, 2017.

[5] J Geralds. Utkface large scale face dataset. *github.com*.

[6] Vijil Chenthamarakshan Kush R. Varshney Prasanna Sattigeri, Samuel C. Hoffman. Fairness gan. 2018.