The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^{2}HCAI)

Through a Fair Looking-Glass: Mitigating Bias in Image Datasets

Amirarsalan Rajabi[1] **Mehdi Yazdani-Jahromi**[1] **Ozlem Ozmen Garibay**[1,2] **Gita Sukthankar**[1]

[1] Department of Computer Science, University of Central Florida [2] Department of Industrial Engineering and Management Systems, University of Central Florida
{amirarsalan,yazdani}@knights.ucf.edu, ozlem@ucf.edu, gitars@eecs.ucf.edu

Abstract

With the recent growth in computer vision applications, the question of how fair and unbiased they are has yet to be explored. There is abundant evidence that the bias present in training data is reflected in the models, or even amplified. Many previous methods for image dataset de-biasing, including models based on augmenting datasets, are computationally expensive to implement. In this study, we present a fast and effective model to de-bias an image dataset through reconstruction and minimizing the statistical dependence between intended variables. Our architecture includes a U-net to reconstruct images, combined with a pre-trained classifier which penalizes the statistical dependence between target attribute and the protected attribute. We evaluate our proposed model on CelebA dataset, compare the results with two state-of-the-art de-biasing method, and show that the model achieves a promising fairness-accuracy combination.

Introduction

Due to their increased usage within myriad software applications, artificial intelligence algorithms now influence many aspects of people's lives, particularly when they are embedded into decision-support tools used by educators, government agencies, and various industry sectors. Thus, it is crucial to make sure that these algorithms are scrutinized to ensure fairness and remove unjust biases. Bias has been shown to exist in several deployed AI systems, including the well known Correlational Offender Management Profiling for Alternative Sanctions (COMPAS). COMPAS is an automated decision making system used by the US criminal justice system for assessing a criminal defendant's likelihood of re-offending. By exploring the risk scores assigned to individuals, this system has been shown to be biased against African Americans (Chouldechova 2017). Other examples include a version of Google's targeted advertising system in which highly paid jobs were advertised more frequently to men vs. women (Lambrecht and Tucker 2019).

Bias in computer vision is a major problem, often stemming from the training datasets used for computer vision models (Tommasi et al. 2017). There is evidence suggesting the existence of multiple types of bias, including capture and selection bias, in popular image datasets (Torralba and Efros 2011). The problems arising from bias in computer vision can manifest in different ways. For instance, it is observed that in activity recognition models, when the datasets contain gender bias, the bias is further amplified by the models trained on those datasets (Zhao et al. 2017). Face recognition models may exhibit lower accuracy for some classes of race or gender (Buolamwini and Gebru 2018).

This paper addresses the issue of a decision-making process being dependent on *protected attributes*, where this dependence should ideally be avoided. From a legal perspective, a protected attribute is an attribute upon which discrimination is illegal (Pessach and Shmueli 2020), e.g. gender or race. Let $D = (\mathcal{X}, \mathcal{S}, \mathcal{Y})$ be a dataset, where $\mathcal{X}$ represents unprotected attributes, $\mathcal{S}$ is the protected attribute, and $\mathcal{Y}$ be the target attribute. If in the dataset D, the target attribute is not independent of the protected attribute ($\mathcal{Y} \not\perp \mathcal{S}$), then it is very likely that the decisions $\hat{\mathcal{Y}}$ made by a decision-making system which is trained on D, is also not independent of the protected attribute ($\hat{\mathcal{Y}} \not\perp \mathcal{S}$).

We propose a model to reconstruct an image dataset to reduce statistical dependency between a protected attribute and target attribute. We modify a U-net (Ronneberger, Fischer, and Brox 2015) to reconstruct the image dataset and apply the Hilbert-Schmidt norm of the cross-covariance operator (Gretton et al. 2005a) between reproducing kernel Hilbert spaces of the target attribute and the protected attribute, as a measure of statistical dependence. Unlike many previous algorithms, our proposed method doesn't require training new classifiers on the unbiased data, but instead reconstructing images in a way that reduces the bias entailed by using the same classifiers.

In Section Methodology we present the problem, the notion of independence, and our proposed methodology. In Section Experiments we describe the CelebA dataset and the choice of feature categorization, introduce the baseline model with which we compare our results (Ramaswamy, Kim, and Russakovsky 2021), our model's implementation details, and finally present the experiments and results.

Background

Bias mitigation methods can be divided into three general categories of *pre-process*, *in-process*, and *post-process*. Preprocess methods include modifying the training dataset be-

fore feeding it to the machine learning model. In-process methods include adding regularizing terms to penalize some representation of bias during the training process. Finally, post-process methods include modifying the final decisions of the classifiers (Hardt, Price, and Srebro 2016). Kamiran and Calders (Kamiran and Calders 2012) propose methods such as suppression which includes removing attributes highly correlated with the protected attribute, reweighing, i.e. assigning weights to different instances in the data, and massaging the data to change labels of some objects. Bias mitigation methods often come at the expense of losing some accuracy, and these preliminary methods usually entail higher fairness-utility cost. More sophisticated methods with better results include using generative models to augment the biased training dataset with unbiased data (Ramaswamy, Kim, and Russakovsky 2021), or training the models on entirely synthetic unbiased data (Rajabi and Garibay 2021). (Wang et al. 2020) provide a set of analyses and a benchmark to evaluate and compare bias mitigation techniques in visual recognition models.

Works such as (Wang, Narayanan, and Russakovsky 2020; Yang et al. 2020) suggest methods to mitigate bias in visual datasets. Several studies have deployed GANs for bias mitigation in image datasets. For example, (Sattigeri et al. 2019) modified the value function of GAN to generate fair image datasets. FairFaceGAN (Hwang et al. 2020) implements a facial image-to-image translation, preventing unwanted translation in protected attributes. Ramaswamy et al. propose a model to produce training data that is balanced for each protected attribute, by perturbing the latent vector of a GAN (Ramaswamy, Kim, and Russakovsky 2021). Other studies employing GANs for fair data generation include (Choi et al. 2020; Sharmanska et al. 2020).

A variety of techniques beyond GANs have been applied to the problems of fairness in AI. A deep information maximization adaptation network was used to reduce racial bias in face image datasets (Wang et al. 2019a), and reinforcement learning was used to learn a race-balanced network in (Wang and Deng 2019). Wang et al. propose a generative few-shot cross-domain adaptation algorithm to perform fair cross-domain adaption and improve performance on minority category (Wang et al. 2021). The work in (Xu et al. 2021) proposes adding a penalty term into the softmax loss function to mitigate bias and improve fairness performance in face recognition. (Quadrianto, Sharmanska, and Thomas 2019) propose a method to discover fair representations of data with the same semantic meaning of the input data. Adversarial learning has also successfully been deployed for this task (Zhang, Lemoine, and Mitchell 2018; Wang et al. 2019b).

Methodology

Consider a dataset $D = (\mathcal{X}, \mathcal{S}, \mathcal{Y})$, where $\mathcal{X}$ is the set of images, $\mathcal{Y} = \{+1, -1\}$ is the target attribute such as attractiveness, and $\mathcal{S} = \{A, B, C, ...\}$ is the protected attribute such as gender. Assume there exists a classifier $f : (\mathcal{X}) \rightarrow \mathcal{Y}$, such that the classifier's prediction for target attribute is not independent from the protected attribute, i.e. $f(\mathcal{X}) \not\perp \mathcal{S}$. Our objective is to design a transformation $g : \mathcal{X} \rightarrow \widetilde{\mathcal{X}}$, such that 1) $f(\widetilde{\mathcal{X}}) \perp \mathcal{S}$, i.e. the classifier's predictions for target attribute is independent of the protected attribute , and 2) $f(\widetilde{\mathcal{X}}) \approx f(\mathcal{X})$, i.e. the classifier still achieves high accuracy.

In other words we want to train a network to transform our original images, such that if the classifiers that are trained on the original and unmodified images, are used to predict the target attribute (attractiveness in our example) from the transformed version of an image, they still achieve high accuracy, while the predictions of those classifiers are independent of the protected attribute (gender in our example). It should be noted that we are not seeking to train new classifiers, but rather only aim to modify the input images. This is a main distinction between our methodology and most of other techniques (e.g. (Quadrianto, Sharmanska, and Thomas 2019) and (Ramaswamy, Kim, and Russakovsky 2021)), in which the process includes training new classifiers on modified new image datasets and achieving *fair classifiers*.

Our proposed model consists of a U-net (Ronneberger, Fischer, and Brox 2015) as the neural network that transforms the original images. This type of network was originally proposed for medical image segmentation, and has been widely used since its introduction. The encoder-decoder network consists of two paths, a contracting path consisting of convolution and max pooling layers, and a consecutive expansive path consisting of upsampling of the feature map and convolutions. Contrary to (Ronneberger, Fischer, and Brox 2015) where each image is provided with a segmented image label, we provide our U-net with the exact same image as the label, and alter the loss function from cross-entropy to mean squared error, so that the network gets trained to produce an image as close to the original image as possible, in a pixel-wise manner.

While some previous fairness studies consider *decorrelating* the target attribute from the protected attributes, what must be ultimately sought however, is independence between the protected attribute and the target attribute. Dealing with two random variables which are uncorrelated is easier than independence, as two random variables might have a zero correlation, and still be dependent (e.g. two random variables A and B with recordings $A = [-2, -1, 0, 1, 2]$ and $B = [4, 1, 0, 1, 4]$ have zero covariance, but are apparently not independent). Given a Borel probability distribution $\mathbf{P}_{ab}$ defined on a domain $\mathcal{A} \times \mathcal{B}$, and respective marginal distributions $\mathbf{P}_a$ and $\mathbf{P}_b$ on $\mathcal{A}$ and $\mathcal{B}$, independence of a and b ($a \perp\!\!\!\perp b$) is equal to $\mathbf{P}_{ab}$ factorizing as $\mathbf{P}_a$ and $\mathbf{P}_b$. Furthermore, two random variables a and b are independent, if and only if any bounded continuous function of the two random variables are uncorrelated (Gretton et al. 2005b).

Let $\mathcal{F}$ and $\mathcal{G}$ denote all real-value functions defined on domains $\mathcal{A}$ and $\mathcal{B}$ respectively. In their paper (Gretton et al. 2005a) define the Hilbert-Schmidt norm of the cross-covariance operator:

$$HSIC(\mathbf{P}_{ab}, \mathcal{F}, \mathcal{G}) := ||C_{ab}||^2_{HS} \quad (1)$$

where C_{ab} is the cross-covariance operator. They show that if $||C_{ab}||^2_{HS}$ is zero, then $cov(f, g)$ will be zero for any

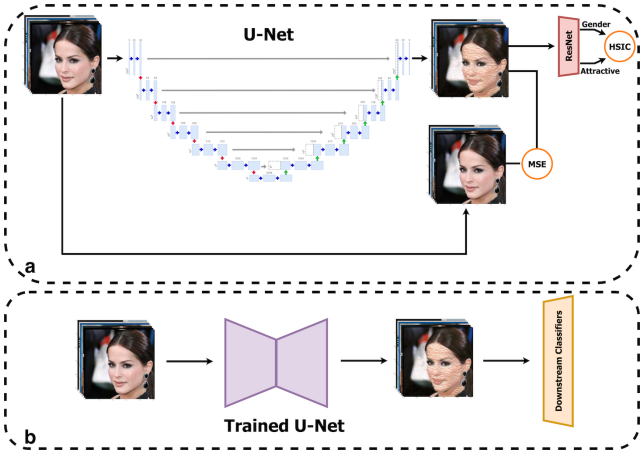

Figure 1: Our model consists of an encoder-decoder (U-net) and a double-output pre-trained ResNet classifier. First, the output batch of the U-net (reconstructed images) is compared with the original batch of images by calculating MSE loss. Then, the output batch of the U-net passes through the ResNet and statistical dependency of the two vectors is calculated by HSIC. Detailed architecture of the U-net is described in the supplementary material.

$f \in \mathcal{F}$ and $g \in \mathcal{G}$, and therefore the random variables a and b will be independent. Furthermore, they show if $\mathcal{Z} := (a_1, b_1), ..., (a_n, b_n) \in \mathcal{A} \times \mathcal{B}$ are a series of n independent observations drawn from $\mathbf{P}_{ab}$, then a (biased) estimator of **HSIC** is (Gretton et al. 2005a):

$$HSIC(\mathcal{Z}, \mathcal{F}, \mathcal{G}) := (n-1)^{-2}\mathbf{tr}(KHLH) \tag{2}$$

where $H, K, L \in \mathbb{R}^{n \times n}$, K and L are Gram matrices (Horn and Johnson 2012), $K_{ij} := k(a_i, a_j)$, $L_{ij} := l(b_i, b_j)$, k and l are universal kernels, and $H_{ij} := \delta_{ij} - n^{-1}$ centers the observations in feature space. We use Hilbert-Schmidt independence criteria to penalize the model for dependence between the target attribute and the protected attribute.

Training Loss Function

We seek to modify a set of images, such that 1) the produced images are close to the original images, and 2) the predicted target attribute is independent from the predicted protected attribute. In the optimization problem, image quality (1) is measured by pixel-wise MSE loss. For independence (2), consider our U-net network as a mapping from original image to the transformed image, i.e. $U_w(\mathbf{x}) = \widetilde{\mathbf{x}}$. Consider also a function $h : \mathcal{X} \to [0,1] \times [0,1]$, where $h(\mathbf{x}_i) = (h_1(\mathbf{x}_i), h_2(\mathbf{x}_i)) = (\mathrm{P}(y_i = 1|\mathbf{x}_i), \mathrm{P}(s_i = 1|\mathbf{x}_i))$. Our objective is to train the parameters of U_w such that $h_1(U_w(\mathbf{x})) \perp\!\!\!\perp h_2(U_w(\mathbf{x}))$, i.e. $h_1(U_w(\mathbf{x}))$ is independent of $h_2(U_w(\mathbf{x}))$.

Given X representing a batch of N training images and $\widetilde{X}$ representing the transformed batch, our formal optimization problem is as follows:

$$\underset{U_w}{\text{minimize}} \underbrace{\frac{1}{NCWH}\sum_{n=1}^{N}\sum_{i,j,k}(\mathbf{x}_{ijk}^n - \widetilde{\mathbf{x}}_{ijk}^n)^2}_{\text{image accuracy}} + \lambda \times \underbrace{HSIC(h_1(\widetilde{X}), h_2(\widetilde{X}))}_{\text{independence}} \tag{3}$$

where N is the number of samples, C is the number of channels of an image, W is the width of an image, H is the height of an image, and λ is the parameter that controls the trade-off between accuracy of the transformed images and independence (fairness). In practice, the mapping function U_w that we use is a U-net, the function $h(\cdot)$ is a pre-trained classifier with two outputs h_1 and h_2, each being the output of a Sigmoid function within the range of $[0, 1]$, where $h_1 = \mathrm{P}(Y = 1|X)$ (a vector of size N), and $h_2 = \mathrm{P}(S = 1|X)$ (also a vector of size N), and $HSIC(\cdot, \cdot)$ denotes Hilbert-Schmidt Independence Criteria.

Figure 1 shows the network architecture and a schematic of the training procedure. Consider a batch of original images X entering the U-net. The U-net then produces the reconstructed images $U_w(X) = \widetilde{X}$. To calculate the *image accuracy* part of the loss function, the original image batch X is provided as label and the Mean Squared Error is calculated to measure the accuracy of the reconstructed images. The ResNet component in Figure 1 is our $h(\cdot)$ function as described before, which is a pre-trained ResNet classifier that takes as input a batch of images and returns two probability vectors. The second part of the loss function, *independence*, is calculated by entering the reconstructed images $\widetilde{X}$ into this ResNet classifier, and calculating the HSIC between the two vectors.

As noted before, the image dataset is reconstructed in a way that using them on the original biased classifiers, will result in an improvement in classifications. This is dissimilar to some previous works such as (Ramaswamy, Kim, and Russakovsky 2021) and (Quadrianto, Sharmanska, and Thomas 2019), in which the model training process includes augmenting the original dataset with generated images and training new fair classifiers (Ramaswamy, Kim, and Russakovsky 2021), or discovering fair representations of images and subsequently training new classifiers (Quadrianto, Sharmanska, and Thomas 2019).

Experiments

In this section, we test the methodology described in Section Methodology on CelebA dataset (Liu et al. 2015). We first introduce the CelebA dataset and the attribute categories in CelebA. We then describe the implementation details of our model. Subsequently, the method described in the work of (Ramaswamy, Kim, and Russakovsky 2021) and the two versions of it that we use as baseline models to compare our results with are introduced. Finally, we introduce evaluation metrics and present the results.

Figure 2: Examples of CelebA dataset original images. Images in the first row are labeled `not Male` and images in the second row are labeled `Male`. In each row, the first three images are labeled `Attractive` and the last three images are labeled `not Attractive`.

CelebA dataset

CelebA is a popular dataset that is widely used for training and testing models for face detection, particularly recognising facial attributes. It consists of 202,599 face images of celebrities, with 10,177 identities. Each image is annotated with 40 different binary attributes describing the image, including attributes such as `Black_Hair`, `Pale_Skin`, `Wavy_Hair`, `Oval_Face`, `Pointy_Nose`, and other attributes such as `Male`, `Attractive`, `Smiling`, etc. The CelebA dataset is reported to be biased (Zhang, Wang, and Zhu 2018). In this experiment, we consider `Male` attribute as the protected attribute (with `Male` $= 0$ showing the image does not belong to a man and `Male` $= 1$ showing the image belongs to a man), and `Attractive` to be the target attribute. We divide the dataset into train and test sets, with train set containing 182,599 and test set containing 20,000 images. In the training set, 67.91% of images with `Male` $=$ 0 are annotated to be attractive (`Attractive` $= 1$), while only 27.93% of images with `Male` $= 1$ are annotated as being attractive (`Attractive` $= 1$). This shows bias exists against images with `Male` $= 1$.

In order to compare our results with (Ramaswamy, Kim, and Russakovsky 2021), we follow their categorization of CelebA attributes. Leaving out `Male` as the protected attribute, among the rest 39 attributes in CelebA dataset, (Ramaswamy, Kim, and Russakovsky 2021) eliminates some attributes such as `Blurry` and `Bald` as they contain less than 5% positive images. The remaining 26 attributes is subsequently categorized into three groups. *inconsistently-labeled* attributes are the ones that by visually examining sets of examples, the authors often disagree with the labeling and could not distinguish between positive and negative examples (Ramaswamy, Kim, and Russakovsky 2021). This group includes attributes such as `Straight_Hair`, and `Big_Hair`. The second group of attributes are the ones that are called *gender-dependent* and the images are labeled to have (or not have) attributes based on the perceived gender (Ramaswamy, Kim, and Russakovsky 2021). These include attributes such as `Young`, `Arched_Eyebrows` and `Receding_Hairline`. Finally, the last group of attributes are called *gender-independent*. These attributes are fairly consistently labeled and are not much dependent on gender expression. This group includes attributes such as `Black_Hair`, `Bangs`, and `Wearing_Hat`. The list of all attributes is provided in supplementary material.

Attribute classifiers

For attribute classifiers, we use ResNet-18 pre-trained on ImageNet, in which the last layer is replaced with a layer of size one, along with a Sigmoid activation for binary classification. We train all models for 5 epochs with batch sizes of 128. We use the Stochastic Gradient Descent optimizer with a learning rate of 1e-3 and momentum of 0.9. We use a step learning rate decay with step size of 7 and factor of 0.1. After training, we will have 26 classifiers that receive an image and perform a binary classification on their respective attribute.

Implementation details

As shown in Figure 1, a ResNet-18 network is used to accompany the U-net to produce predictions for `Male` and `Attractive`. Prior to training the U-net, the ResNet-18 (Russakovsky et al. 2015) which is pre-trained on ImageNet, is modified by replacing its output layer with a layer of size two, outputing the probability of attractiveness and gender. The ResNet-18 is then trained for 5 epochs on the train set, with a batch size of 128. We use the Stochastic Gradient Descent optimizer with a learning rate of 1e-3 and momentum of 0.9. We use a step learning rate decay with step size of 7 and factor of 0.1. After the ResNet is trained and prepared, we train the U-net as described in Section Methodology on the train set. The detailed architecture of the U-net is described in Supplementary Material. In our implementation of biased estimator of HSIC estimator in Equation 2,

Figure 3: Examples of CelebA dataset images and how the model reconstructs them. The first row shows a set of images from the original testing set, and the second row shows the reconstructed images.

we use Gaussian RBF kernel function for $k(\cdot,\cdot)$ and $l(\cdot,\cdot)$. The training was conducted on a machine with two NVIDIA GeForce RTX 3090, and each training of the U-Net took 1 hour. When the training is complete, the U-net is ready to reconstruct images. Figure 3 shows six examples of how the U-net modifies the original images. We train our model for 5 epochs with an $\lambda = 0.07$.

Comparison with baseline models

We compare our results with Ramaswamy et al.'s method, described in their paper 'Fair Attribute Classification through Latent Space De-biasing' (Ramaswamy, Kim, and Russakovsky 2021). Building on work by (Denton et al. 2019) which demonstrates a method to learn interpretable image modification directions, they develop an improved method by perturbing latent vector of a GAN, to produce training data that is balanced for each protected attribute. By augmenting the original dataset with the generated data, they train target classifiers on the augmented dataset, and show that these classifiers will be fair, with high accuracy. The second model that we compare our results with is explicit removal of biases from neural network embeddings, presented in (Alvi, Zisserman, and Nellåker 2018). The authors provide an algorithm to remove multiple sources of variation from the feature representation of a network. This is achieved by including secondary branches in a neural network with the aim to minimize a confusion loss, which in turn seeks to change the feature representation of data such that it becomes invariant to the spurious variations that are desired to be removed.

We implement Ramaswamy et al.'s method as follows: As mentioned in their paper, we used progressive GAN with 512-D latent space trained on the CelebA training set from the PyTorch GAN Zoo. We use 10,000 synthetic images and label the synthetic images with a ResNet-18 (modified by adding a fully connected layer with 1,000 neurons). Then we trained a linear SVM to learn the hyper-planes in the latent space as proposed in the original paper. We generate $\mathcal{X}_{syn}$ (160,000 images) to generate a synthetic dataset which aims to de-bias `Male` from all 26 attributes one by one. Next, we train ResNet-18 classifiers on the new datasets consisting of augmenting $\mathcal{X}$ and $\mathcal{X}_{syn}$. We call this model as *GANDeb*. We use the implementation of (Alvi, Zisserman, and Nellåker 2018) with the uniform confusion loss $-(1/|D|)\sum_d \log q_d$ provided in (Wang et al. 2020).

Evaluation metrics

In evaluating the results of our model with the baseline models, three metrics are used. To capture the accuracy of the classifiers, we measure the *average precision*. This metric combines precision and recall at every position and computes the average. A higher average precision (**AP**) is desired. To measure fairness, there are multiple metrics proposed in the literature (Mehrabi et al. 2021). Among the most commonly used metrics is *demographic parity* (**DP**). This metric captures the disparity of receiving a positive decision among different protected groups ($|P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)|$). A smaller **DP** shows a fairer classification and is desired. Finally for our last fairness measure, we follow (Lokhande et al. 2020) and (Ramaswamy, Kim, and Russakovsky 2021) and use *difference in equality of opportunity* (**DEO**), i.e. the absolute difference between the true positive rates for both gender expressions ($|TPR(S = 0) - TPR(S = 1)|$). A smaller **DEO** is desired.

Results

All the values reported in this section, are evaluated on the same test set. Prior to comparing the results of our method with the comparison models, to assess the original training data, the performance of baseline, i.e. classifiers being trained on the original train set, and tested on the test set is presented. The AP, DP, and DEO values of classifiers trained on the original training set is shown in Table 1 under *Baseline*. Looking into Baseline values, the AP of classifiers for gender-independent category of attributes is higher

than gender-dependent category, and the AP of inconsistent category is less than the other two categories. As expected, DP and DEO for gender-dependent category of attributes is higher than the other two categories.

In Table 1, we compare our model with GAN Debiasing (GanDeb) (Ramaswamy, Kim, and Russakovsky 2021), Adversarial debiasing (AdvDb) presented in (Alvi, Zisserman, and Nellåker 2018), and the Baseline on the original data. Looking into the average precision scores, the results show that GanDeb is slightly performing better than Ours. This is anticipated, since half of the training data for GanDeb consists of the original images, and therefore a higher average precision is expected. AdvDb on the other hand is performing poorly in terms of average precision, with average precision scores far away from other models.

Looking into demographic parity scores, the results show that GanDeb falls behind the other two models in two out of three attribute categories. While Ours is performing better for gender dependent and gender independent attribute categories. Looking into the third fairness measure, difference in equality of opportunity, AdvDb and ours are performing better than GanDeb in all three categories of attributes. Ours beats AdvDb for inconsistent attributes category, AdvDb beats Ours in gender dependent category, and AdvDb slightly beats Ours for gender independent category of attributes. In summary, Ours is close to GanDeb in terms of maintaining high average precision scores, which means higher accuracy of prediction, while beating GanDeb in terms of fairness metrics. Also, while AdvDb performance in terms of fairness enforcement is better than ours in 3 out of 6 cases, it falls behind significantly in terms of average precision.

To explore the trade-off between fairness and precision, we perform the following experiment: λ was increased between $[0.01, 0.15]$ in steps of 0.01, and for each value of λ, the model was trained three times, each time for 1 epoch. Figure 4 shows how AP, DEO, and DP change. The results show that by increasing λ, precision decreases while fairness measures improve.

Interpretation and the effect on other attributes

In this section, we aim to display the correspondence between an attribute's relationship with `Attractive` attribute, and the extent to which the model modifies that attribute. To do so, for each attribute, we record two values, namely HSIC value between that attribute and the `Attractive` attribute, and the change in demographic parity. To calculate the change in demographic parity, we first calculate the demographic parity of the classifier for that specific attribute, when the classifier classifies the original testing set images (similar to *Baseline* in previous tables, but for each attribute separately). We then calculate the demographic parity of the classifier for that specific attribute, when the classifier receives the modified training images **Ours(5,0.07)**. We then subtract the two values, to get the change in demographic parity for that specific attribute. Figure 5 presents the results, with the red bars showing the change in demographic parity for each attribute, and the blue bars showing the statistical dependence measured by HSIC, between each attribute with `Attractive` attribute, in the original training data. The results show that the absolute change in demographic parity is positively correlated with that attribute's statistical dependence with the attribute `Attractive`, with a Pearson correlation coefficient of 0.757. For instance, we observe large changes in demographic parity for attributes such as `Young`, `Big_Nose`, `Pointy_Nose`, `Oval_Face`, and `Arched_Eyebrows`, as they are typically associated with being attractive, and therefore reflected in the CelebA dataset labels.

Conclusions

We proposed an image reconstruction process to mitigate bias against a protected attribute. The model's performance was evaluated on CelebA dataset and compared with an augmentation based method developed by (Ramaswamy, Kim, and Russakovsky 2021). The proposed model showed promising results in mitigating bias while maintaining high precision for classifiers. An interesting aspect of the results is that although we only explicitly train the U-net to remove dependence between the target attribute (`Attractive`) and the protected attribute (`Male`), classifiers related to many other attributes, most of which have a statistical dependency with the target attribute, become 'fairer'. An advantage of the proposed model is that it does not rely on modifying downstream classifiers, and rather includes only modifying the input data, hence making it suitable to be deployed in an automated machine learning pipeline more easily and with lower cost. As a potential future direction, we intend to consider the problem in a situation where multiple protected attributes are present, and attributes are non-binary. We also intend to apply similar methodology on other data types such as tabular data.

References

Alvi, M.; Zisserman, A.; and Nellåker, C. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Choi, K.; Grover, A.; Singh, T.; Shu, R.; and Ermon, S. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, 1887–1898. PMLR.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Denton, E.; Hutchinson, B.; Mitchell, M.; Gebru, T.; and Zaldivar, A. 2019. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005a. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

	AP ↑			DP ↓			DEO ↓		
	Incons.	G-dep	G-indep	Incons.	G-dep	G-indep	Incons.	G-dep	G-indep
Baseline	0.667	0.79	0.843	0.147	0.255	0.137	0.186	0.243	0.163
GanDeb	0.641	0.763	0.831	0.106	0.233	0.119	0.158	0.24	0.142
AdvDb	0.243	0.333	0.218	0.091	0.169	0.121	0.136	0.149	0.098
Ours	0.618	0.732	0.839	0.097	0.146	0.118	0.124	0.172	0.114

Table 1: Comparing the results of our model with Baseline, GAN debiasing (GanDeb), and Adversarial debiasing (AdvDb). Showing AP (Average Precision, higher the better), DP (Demographic Parity, lower the better), and DEO (Difference in Equality of Opportunity, lower the better) values for each attribute category. Each number is the average over all attributes within that specific attribute category.

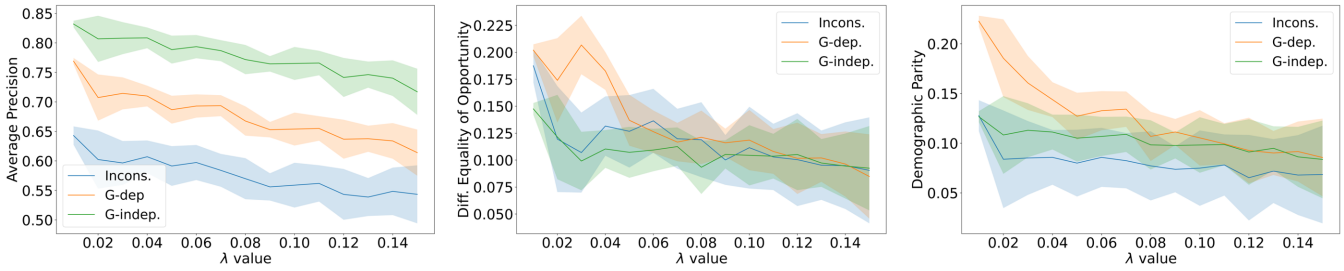

Figure 4: Exploring the trade-off between accuracy and fairness by incremental increasing of parameter λ. Each data point is the average over three trainings, with standard deviation of the three trainings shown as confidence intervals.

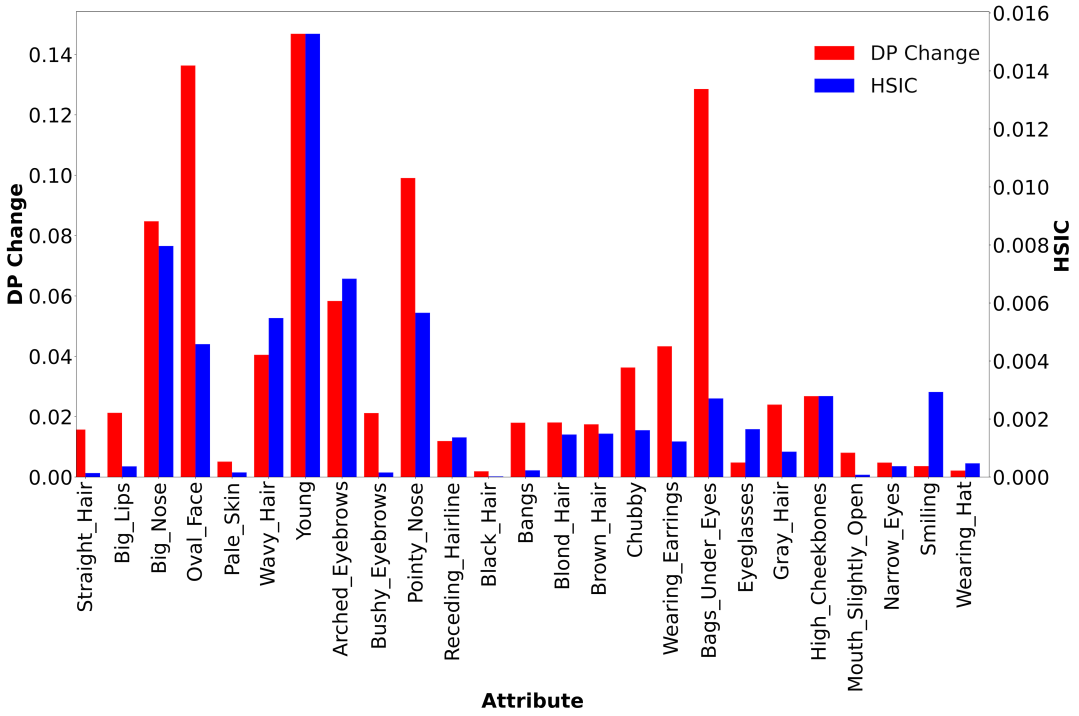

Figure 5: Displaying the relationship between an attribute's statistical dependence on `Attractive` attribute, and the extent to which the model modifies that attribute. Blue bars show the HSIC between each attribute with `Attractive` attribute in the original data. Red bars show the absolute difference in demographic parity of each attribute's classifier, acting on original images and transformed images, respectively.

Gretton, A.; Herbrich, R.; Smola, A.; Bousquet, O.; Schölkopf, B.; et al. 2005b. Kernel methods for measuring independence.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.

Horn, R. A.; and Johnson, C. R. 2012. *Matrix analysis*. Cambridge university press.

Hwang, S.; Park, S.; Kim, D.; Do, M.; and Byun, H. 2020. FairfaceGAN: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Lambrecht, A.; and Tucker, C. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management science*, 65(7): 2966–2981.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lokhande, V. S.; Akash, A. K.; Ravi, S. N.; and Singh, V. 2020. Fairalm: Augmented lagrangian method for training fair models with little regret. In *European Conference on Computer Vision*, 365–381. Springer.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6): 1–35.

Pessach, D.; and Shmueli, E. 2020. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8227–8236.

Rajabi, A.; and Garibay, O. O. 2021. TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks. *arXiv preprint arXiv:2109.00666*.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9301–9310.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5): 3–1.

Sharmanska, V.; Hendricks, L. A.; Darrell, T.; and Quadrianto, N. 2020. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*.

Tommasi, T.; Patricia, N.; Caputo, B.; and Tuytelaars, T. 2017. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, 37–55. Springer.

Torralba, A.; and Efros, A. A. 2011. Unbiased look at dataset bias. In *CVPR 2011*, 1521–1528. IEEE.

Wang, A.; Narayanan, A.; and Russakovsky, O. 2020. REVISE: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision*, 733–751. Springer.

Wang, M.; and Deng, W. 2019. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*.

Wang, M.; Deng, W.; Hu, J.; Tao, X.; and Huang, Y. 2019a. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 692–702.

Wang, T.; Ding, Z.; Shao, W.; Tang, H.; and Huang, K. 2021. Towards Fair Cross-Domain Adaptation via Generative Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 454–463.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019b. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5310–5319.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Xu, X.; Huang, Y.; Shen, P.; Li, S.; Li, J.; Huang, F.; Li, Y.; and Cui, Z. 2021. Consistent Instance False Positive Improves Fairness in Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 578–586.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–558.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, Q.; Wang, W.; and Zhu, S.-C. 2018. Examining cnn representations with respect to dataset bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.