Discovering Fair Representations in the Data Domain

Novi Quadrianto[‡*], Viktoriia Sharmanska[§], Oliver Thomas[‡]
‡Predictive Analytics Lab (PAL), University of Sussex, Brighton, United Kingdom
§Department of Computing, Imperial College London, United Kingdom

Abstract

Interpretability and fairness are critical in computer vision and machine learning applications, in particular when dealing with human outcomes, e.g. inviting or not inviting for a job interview based on application materials that may include photographs. One promising direction to achieve fairness is by learning data representations that remove the semantics of protected characteristics, and are therefore able to mitigate unfair outcomes. All available models however learn latent embeddings which comes at the cost of being uninterpretable. We propose to cast this problem as data-to-data translation, i.e. learning a mapping from an input domain to a fair target domain, where a fairness definition is being enforced. Here the data domain can be images, or any tabular data representation. This task would be straightforward if we had fair target data available, but this is not the case. To overcome this, we learn a highly unconstrained mapping by exploiting statistics of residuals – the difference between input data and its translated version – and the protected characteristics. When applied to the CelebA dataset of face images with gender attribute as the protected characteristic, our model enforces equality of opportunity by adjusting the eyes and lips regions. Intriguingly, on the same dataset we arrive at similar conclusions when using semantic attribute representations of images for translation. On face images of the recent DiF dataset, with the same gender attribute, our method adjusts nose regions. In the Adult income dataset, also with protected gender attribute, our model achieves equality of opportunity by, among others, obfuscating the wife and husband relationship. Analyzing those systematic changes will allow us to scrutinize the interplay of fairness criterion, chosen protected characteristics, and prediction performance.

1. Introduction

Machine learning systems are increasingly used by government agencies, businesses, and other organisations to assist in making life-changing decisions such as whether or not to invite a candidate to a job interview, or whether to give someone a loan. The question is how can we ensure that those systems are *fair*, i.e. they do not discriminate against individuals because of their gender, disability, or other personal ("protected") characteristics? For example, in building an automated system to review job applications, a photograph might be used in addition to other features to make an invite decision. By using the photograph as is, a discrimination issue might arise, as photographs with faces could reveal certain protected characteristics, such as gender, race, or age (e.g. [14, 5, 4, 29]). Therefore, any automated system that incorporates photographs into its decision process is at risk of indirectly conditioning on protected characteristics (indirect discrimination). Recent advances in learning fair representations suggest adversarial training as the means to hide the protected characteristics from the decision/prediction function [2, 49, 33]. All fair representation models, however, learn *latent embeddings*. Hence, the produced representations cannot be easily interpreted. They do not have the semantic meaning of the input that photographs, or education and training attainments, provide when we have job application data. If we want to encourage public conversations and productive public debates regarding fair machine learning systems [18], interpretability in how fairness is met is an integral yet overlooked ingredient.

In this paper we focus on representation learning models that can transform inputs to their fair representations and retain the semantics of the input domain in the transformed space. When we have image data, our method will make a semantic change to the appearance of an image to deliver a certain fairness criterion[1]. To achieve this, we perform *a data-to-data translation* by learning a mapping from data in a source domain to a target domain. Mapping from source to target domain is a standard procedure, and many methods are available. For example, in the image domain, if we have aligned source/target as training data, we can use the pix2pix method of [24], which is based on conditional generative adversarial networks (cGANs) [36]. Zhu et al.'s

[*] Also with Higher School of Economics, Moscow, Russia

[1] Examples of fairness criteria are equality of true positive rates (TPR), also called equality of opportunity [22, 47], between males and females.

CycleGAN [50] and Choi et al.'s StarGAN [7] solve a more challenging setting in which only *un*aligned training examples are available. However, we can not simply reuse existing methods for source-to-target mapping because we do *not have data in the target domain* (e.g. fair images are not available; images by themselves can not be fair or unfair, it is only when they are coupled with a particular task that the concern of fairness arises).

To illustrate the difficulty, consider our earlier example of an automated job review system that uses photographs as part of an input. For achieving fairness, it is tempting to simply use GAN-driven methods to *translate female face photos to male*. We would require training data of female faces (source domain) and male faces (target domain), and only unaligned training data would be needed. This solution is however fundamentally flawed; who gets to decide that we should translate in this direction? Is it fairer if we translate male faces to female instead? An ethically grounded approach would be to translate both male and female face photos (source domain) to appropriate middle ground face photos (target domain). This challenge is actually multi-dimensional, it contains at least *two sub-problems*: a) how to have a general approach that can handle image data as well as tabular data (e.g. work experience, education, or even semantic attribute representations of photographs), and b) how to find a middle-ground with a multi-value (e.g. race) or continuous value (e.g. age) protected characteristic or even multiple characteristics (e.g. race and age).

We propose a solution to the multi-dimensional challenge described above by exploiting statistical (in)dependence between translated images and protected characteristics. We use the Hilbert-Schmidt norm of the cross-covariance operator between reproducing kernel Hilbert spaces of image features and protected characteristics (Hilbert-Schmidt independence criterion [20]) as an empirical estimate of statistical independence. This flexible measure of independence allows us to take into account higher order independence, and handle a multi-/continuous value and multiple protected characteristics.

Related work We focus on expanding the related topic of learning fair, *albeit uninterpretable*, representations. The aim of fair representation learning is to learn an intermediate representation of the data that preserves as much information about the data as possible, while simultaneously removing protected characteristic information such as age and gender. Zemel et al. [48] learn a probabilistic mapping of the data point to a set of latent prototypes that is independent of protected characteristic (equality of acceptance rates, also called a statistical parity criterion), while retaining as much class label information as possible. Louizos et al. [32] extend this by employing a deep variational auto-encoder (VAE) framework for finding the fair latent representation. In recent years, we see increased adversarial learning methods for fair representations. Ganin et al. [15] propose adversarial representation learning for domain adaptation by requiring the learned representation to be indiscriminate with respect to differences in the domains. Multiple data domains can be translated into multiple demographic groups. Edwards and Storkey [12] make this connection and propose adversarial representation learning for the statistical parity criterion. To achieve other notions of fairness such as equality of opportunity, Beutel et al. [2] show that the adversarial learning algorithm of Edwards and Storkey [12] can be reused but we only supply training data with positive outcome to the adversarial component. Madras et al. [33] use a label-aware adversary to learn fair and transferable latent representations for the statistical parity as well as equality of opportunity criteria.

None of the above learn fair representations while simultaneously retaining the semantic meaning of the data. There is an orthogonal work on feature selection using human perception of fairness (e.g. [21]), while this approach undoubtedly retains the semantic meaning of tabular data, it has not been generalized to image data. In an independent work to ours, Sattigeri et al. [40] describe a similar motivation of producing fair representations in the input image domain; their focus is on creating a whole new image-like dataset, rather than conditioning on each input image. Hence it is not possible to visualise a fair version for a given image as provided by our method (refer to Figures 2 and 3).

2. Interpretability in Fairness by Residual Decomposition

We will use the illustrative example of an automated job application screening system. Given input data (photographs, work experience, education and training, personal skills, etc.) $\mathbf{x}^n \in \mathcal{X}$, output labels of performed well or not well $y^n \in \mathcal{Y} = \{+1, -1\}$, and protected characteristic values, such as *race* or *gender*, $s^n \in \{A, B, C, D, \ldots\}$, or *age*, $s^n \in \mathbb{R}$, we would like to train a classifier f that decides whether or not to invite a person for an interview. We want the classifier to predict outcomes that are accurate with respect to y^n but fair with respect to s^n.

2.1. Fairness definitions

Much work has been done on mathematical definitions of fairness (e.g. [28, 8]). It is widely accepted that no single definition of fairness applies in all cases, but will depend on the specific context and application of machine learning models [18]. In this paper, we focus on the *equality of opportunity* criterion that requires the classifier f and the protected characteristic s be independent, conditional on the label being positive [2], in shorthand notation $f \perp\!\!\!\perp s \mid y = +1$.

[2]With binary labels, it is assumed that positive label is a desirable/advantaged outcome, e.g. expected to perform well at the job.

Expressing the shorthand notation in terms of a conditional distribution, we have $\mathbb{P}(f(\mathbf{x})|s, y = +1) = \mathbb{P}(f(\mathbf{x})|y = +1)$. With binary protected characteristic, this reads as equal true positive rates across the two groups, $\mathbb{P}(f(\mathbf{x}) = +1|s = A, y = +1) = \mathbb{P}(f(\mathbf{x}) = +1|s = B, y = +1)$. Equivalently, the shorthand notation can also be expressed in terms of joint distributions, resulting in $\mathbb{P}(f(\mathbf{x}), s|y = +1) = \mathbb{P}(f(\mathbf{x})|y = +1)\mathbb{P}(s|y = +1)$. The advantage of using the joint distribution expression is that the variable s does not appear as a conditioning variable, making it straightforward to use the expression for a multi- or continuous value or even multiple protected characteristics.

2.2. Residual decomposition

We want to learn a data representation $\tilde{\mathbf{x}}^n$ for each input $\mathbf{x}^n$ such that: a) it is able to predict the output label y^n, b) it protects s^n according to a certain fairness criterion, c) it lies in the same space as $\mathbf{x}^n$, that is $\tilde{\mathbf{x}}^n \in \mathcal{X}$. The third requirement ensures the learned representation to have the same *semantic meaning* as the input. For example, for images of people faces, the goal is to modify facial appearance in order to remove the protected characteristic information. For tabular data, we desire systematic changes in values of categorical features such as education (bachelors, masters, doctorate, etc.). Visualizing those systematic changes will give evidence on how our algorithm enforces a certain fairness criterion. This will be a powerful tool, albeit all the powers hinge on *observational data*, to scrutinize the interplay between fairness criterion, protected characteristics, and classification accuracy. We proceed by making the following decomposition assumption on $\mathbf{x}$:

$$\phi(\mathbf{x}) = \phi(\tilde{\mathbf{x}}) + \phi(\hat{\mathbf{x}}), \tag{1}$$

with $\tilde{\mathbf{x}}$ to be the component that is independent of s, $\hat{\mathbf{x}}$ denoting the component of $\mathbf{x}$ that is dependent on s, and $\phi(\cdot)$ is some *pre-trained* feature map. We will discuss about the specific choice of this pre-trained feature map for both image and tabular data later in the section. What we want is to learn a mapping from a source domain (input features) to a target domain (fair features with the semantics of the input domain), i.e. $T : \mathbf{x} \to \tilde{\mathbf{x}}$, and we will parameterize this mapping $T = T_\omega$ where ω is a class of autoencoding transformer network. For our architectural choice of transformer network, please refer to Section 3.

To enforce the decomposition structure in (1), we need to satisfy two conditions: a) $\tilde{\mathbf{x}}$ to be independent of s, and b) $\hat{\mathbf{x}}$ to be dependent of s. Given a particular statistical dependence measure, the first condition can be achieved by *minimizing* the dependence measure between $P = \{\phi(\tilde{\mathbf{x}}^1), \ldots, \phi(\tilde{\mathbf{x}}^N)\} = \{\phi(T_\omega(\mathbf{x}^1)), \ldots, \phi(T_\omega(\mathbf{x}^N))\}$ and $S = \{s^1, \ldots, s^N\}$; N is the number of training data points. For the second condition, we first define a *residual*:

$$\phi(\mathbf{x}) - \phi(\tilde{\mathbf{x}}) = \phi(\mathbf{x}) - \phi(T_\omega(\mathbf{x})) = \phi(\hat{\mathbf{x}}), \tag{2}$$

where the last term is the data component that is *dependent* on a protected characteristic s. We can then enforce the second condition by *maximizing* the dependence measure between $R = \{\phi(\hat{\mathbf{x}}^1), \ldots, \phi(\hat{\mathbf{x}}^N)\} = \{\phi(\mathbf{x}^1) - \phi(T_\omega(\mathbf{x}^1)), \ldots, \phi(\mathbf{x}^N) - \phi(T_\omega(\mathbf{x}^N))\}$ and S. We use the decomposition property as a guiding mechanism to learn the parameters ω of the transformer network T_ω.

In the fair and interpretable representation learning task, we believe using residual is well-motivated because we know that our generated fair features should be somewhat similar to our input features. Residuals will make learning the transformer network easier. Taking into consideration that we do not have training data about the target fair features $\tilde{\mathbf{x}}$, we should not desire the transformer network to take the input feature $\mathbf{x}$ and *generate* a new output $\tilde{\mathbf{x}}$. Instead, it should just learn how to *adjust* our input $\mathbf{x}$ to produce the desired output $\tilde{\mathbf{x}}$. The concept of residuals is universal, for example, a residual block has been used to speed up and to prevent over-fitting of a very deep neural network [23], and a residual regression output has been used to perform causal inference in additive noise models [37].

Formally, given the N training triplets (X, S, Y), to find a fair and interpretable representation $\tilde{\mathbf{x}} = T_\omega(\mathbf{x})$, our optimization problem is given by:

$$\underset{T_\omega}{\text{minimize}} \underbrace{\sum_{n=1}^{N} \mathcal{L}(T_\omega(\mathbf{x}^n), y^n)}_{\text{prediction loss}} + \lambda_1 \underbrace{\sum_{n=1}^{N} \|\mathbf{x}^n - T_\omega(\mathbf{x}^n)\|_2^2}_{\text{reconstruction loss}} + \lambda_2 \left(\underbrace{-\text{HSIC}(R, S|Y = +1) + \text{HSIC}(P, S|Y = +1)}_{\text{decomposition loss}} \right) \tag{3}$$

where $\text{HSIC}(\cdot, \cdot)$ is the statistical dependence measure, and λ_i are trade-off parameters. HSIC is the Hilbert-Schmidt norm of the cross-covariance operator between reproducing kernel Hilbert spaces. This is equivalent to a non-parametric distance measure of a joint distribution and the product of two marginal distributions using the Maximum Mean Discrepancy (MMD) criterion[19]; MMD has been successfully used in fairnesss literature in it's own right [32, 38]. Section 2.1 discusses defining statistical independence based on a joint distribution, contrasting this with a conditional distribution. We use the biased estimator of HSIC [20, 42]: $\text{HSIC}_{\text{emp.}} = (N-1)^{-2} \operatorname{tr} HKHL$, where $K, L \in \mathbb{R}^{N \times N}$ are the kernel matrices for the *residual* set R and the protected characteristic set S respectively, i.e. $K_{ij} = k(r^i, r^j)$ and $L_{ij} = l(s^i, s^j)$ (similar definition for measuring independence between sets P and S). We use a Gaussian RBF kernel function for both $k(\cdot, \cdot)$ and $l(\cdot, \cdot)$. Moreover, $H_{ij} = \delta_{ij} - N^{-1}$ centres the observations of set R and set S in RKHS feature space. The prediction loss

is defined using a softmax layer on the output of the transformer network. While in image data we add the total variation (TV) penalty [34] on the fair representation to ensure spatial smoothness, we do not enforce any regularization term for tabular data. In summary, we learn a new representation $\tilde{\mathbf{x}}$ that removes statistical dependence on the protected characteristic s (by minimizing HSIC$(P, S|Y = +1)$) and enforces the dependence of the residual $\mathbf{x} - \tilde{\mathbf{x}}$ and s (by maximizing HSIC$(R, S|Y = +1)$). We can then train any classifier f using this new representation, and it will inherently satisfy the fairness criterion [33].

Neural style transfer and pre-trained feature space Neural style transfer (e.g. [17, 25]) is a popular approach to perform an image-to-image translation. Our decomposition loss in (3) is reminiscent of a style loss used in neural style transfer models. The style loss is defined as the distance between second-order statistics of a style image and the translated image. Excellent results [17, 25, 43, 44] on neural style transfer rely on pre-trained features. Following this spirit, we also use a "pre-trained" feature mapping $\phi(\cdot)$ in defining our decomposition loss. For image data, we take advantage of the powerful representation of deep convolutional neural networks (CNN) to define the mapping function [17]. The feature maps of $\mathbf{x}$ in the layer l of a CNN are denoted by $F^l_{\mathbf{x}} \in R^{N_l \times M_l}$ where N_l is the number of the feature maps in the layer l and M_l is the height times the width of the feature map. We use the vectorization of $F^l_{\mathbf{x}}$ as the required mapping $\phi(\mathbf{x}) = \text{vec}(F^l_{\mathbf{x}})$. Several layers of a CNN will be used to define the full mapping (see Section 3). For tabular data, we use the following random Fourier feature [39] mapping $\phi(\mathbf{x}) = \sqrt{2/D}\cos(\langle\theta, \mathbf{x}\rangle + b)$ with a bias vector $b \in \mathbb{R}^D$ that is uniformly sampled in $[0, 2\pi]$, and a matrix $\theta \in \mathbb{R}^{d\times D}$ where θ_{ij} is sampled from a Gaussian distribution. We have assumed the input data lies in a d-dimensional space, and we transform them to a D-dimensional space.

3. Experiments

We gave an illustrative example about screening job applications, however, no such data is publicly available. We will instead use publicly available data to simulate the setting. We conduct the experiments using three datasets: the CelebA image dataset[3] [30], the Diversity in Faces (DiF) dataset [4] [35], and the Adult income dataset[5] from the UCI repository [9]. The CelebA dataset has a total of $202,599$ celebrity images. The images are annotated with 40 attributes that reflect appearance (hair color and style, face shape, makeup, for example), emotional state (smiling), gender, attractiveness, and age. For this dataset, we use gender as a binary protected characteristic, and attractiveness as the proxy measure of getting invited for a job interview in the world of fame. We randomly select 20K images for testing and use the rest for training the model. The DiF dataset has only been introduced very recently and contains nearly a million human face images reflecting diversity in ethnicity, age and gender. We include preliminary results using 200K images for training and 200K images for testing our model on this dataset. The images are annotated with attributes such as race, gender and age (both continual and discretized into seven age groups) as well as facial landmarks and facial symmetry features. For this dataset, we use gender as a binary protected characteristic, and the discretized age groups as a predictive task. The Adult income dataset is frequently used to assess fairness methods. It comes from the Census bureau and the binary task is to predict whether or not an individual earns more than \$50K per year. It has a total of $45,222$ data instances, each with 14 features such as gender, marital status, educational level, number of work hours per week. For this dataset, we follow [48] and consider gender as a binary protected characteristic. We use $28,222$ instances for training, and $15,000$ instances for testing. We enforce equality of opportunity as the fairness criteria throughout for the three experiments.

3.1. The Adult Income dataset

The focus is to investigate whether (**Q1**) our proposed fair and interpretable learning method performs on a par with state-of-the-art fairness methods, and whether (**Q2**) performing a tabular-to-tabular translation brings us closer to achieving interpretability in how fairness is being satisfied. We compare our method against an unmodified $\mathbf{x}$ using the following classifiers: 1) logistic regression (`LR`) and 2) support vector machine with linear kernel (`SVM`), We select the regularization parameter of `LR` and `SVM` over 6 possible values (10^i for $i \in [0, 6]$) using 3-fold cross validation. We then train classifiers 1–2 with the learned representation $\tilde{\mathbf{x}}$ and with the latent embedding $\mathbf{z}$ of a state-of-the-art adversarial model described in Beutel et al. [2]. We also apply methods which reweigh the samples to simulate a balanced dataset with regard to the protected characteristic FairLearn [1] `Fair Reduction` 3-4 and Kamiran & Calders [26] `Kamiran & Calders` 5-6, optimized with both the cross-validated `LR` and `SVM` (1-2), giving (`Fair Reduction LR`), (`Fair Reduction SVM`), (`Kamiran & Calders LR`) and (`Kamiran & Calders SVM`) respectively. As a reference, we also compare with: 7) Zafar et al.'s [47] fair classification method (`Zafar et al.`) that adds equality of opportunity directly as a constraint to the learning objective function. It has been shown that applying fairness constraints in succession as 'fair pipelines' do not enforce fairness [11, 3], as

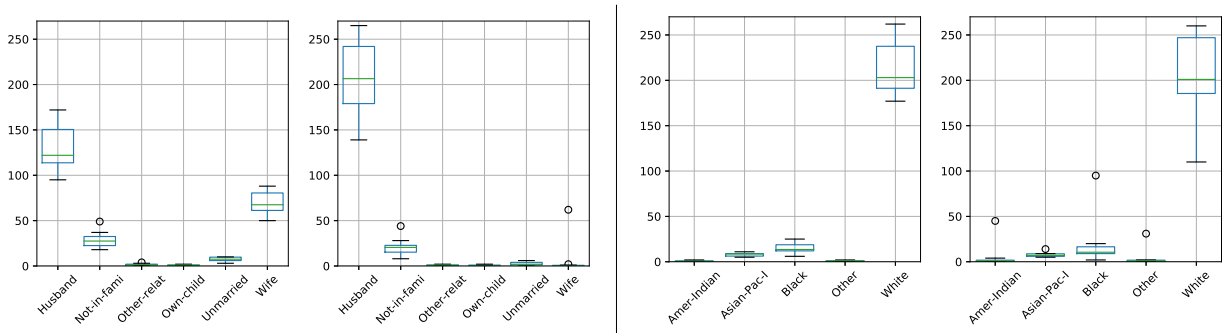
[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[4] https://www.research.ibm.com/artificial-intelligence/trusted-ai/diversity-in-faces/

[5] https://archive.ics.uci.edu/ml/datasets/adult

	original $\mathbf{x}$		fair interpretable $\tilde{\mathbf{x}}$		latent embedding $\mathbf{z}$	
	Accuracy ↑	Eq. Opp ↓	Accuracy ↑	Eq. Opp ↓	Accuracy ↑	Eq. Opp ↓
1: `LR`	85.1 ± 0.2	$\mathbf{9.2 \pm 2.3}$	84.2 ± 0.3	$\mathbf{5.6 \pm 2.5}$	81.8 ± 2.1	$\mathbf{5.9 \pm 4.6}$
2: `SVM`	85.1 ± 0.2	$\mathbf{8.2 \pm 2.3}$	84.2 ± 0.3	$\mathbf{4.9 \pm 2.8}$	81.9 ± 2.0	$\mathbf{6.7 \pm 4.7}$
3: `Fair Reduction LR` [1]	85.1 ± 0.2	$\mathbf{14.9 \pm 1.3}$	84.1 ± 0.3	$\mathbf{6.5 \pm 3.2}$	81.8 ± 2.1	$\mathbf{5.6 \pm 4.8}$
4: `Fair Reduction SVM` [1]	85.1 ± 0.2	$\mathbf{8.2 \pm 2.3}$	84.2 ± 0.3	$\mathbf{4.9 \pm 2.8}$	81.9 ± 2.0	$\mathbf{6.7 \pm 4.7}$
5: `Kamiran & Calders LR` [26]	84.4 ± 0.2	$\mathbf{14.9 \pm 1.3}$	84.1 ± 0.3	$\mathbf{1.7 \pm 1.3}$	81.8 ± 2.1	$\mathbf{4.9 \pm 3.3}$
6: `Kamiran & Calders SVM` [26]	85.1 ± 0.2	$\mathbf{8.2 \pm 2.3}$	84.2 ± 0.3	$\mathbf{4.9 \pm 2.8}$	81.9 ± 2.0	$\mathbf{6.7 \pm 4.7}$
7: `Zafar et al.`* [47]	85.0 ± 0.3	$\mathbf{1.8 \pm 0.9}$	—	—	—	—

Table 1. Results of training multiple classifiers (rows 1–7) on 3 different representations, $\mathbf{x}$, $\tilde{\mathbf{x}}$, and $\mathbf{z}$. $\mathbf{x}$ is the original input representation, $\tilde{\mathbf{x}}$ is the interpretable, fair representation introduced in this paper, and $\mathbf{z}$ is the latent embedding representation of Beutel et al. [2]. We **boldface** Eq. Opp. since this is the fairness criterion (the lower the better). *The solver of `Zafar et al.` fails to converge in 4 out of 10 repeats. Our learned representation $\tilde{\mathbf{x}}$ achieves comparable fairness level to the latent representation $\mathbf{z}$, while maintaining the constraint of being in the same space as the original input.



('Relationship Status') ('Race')

Figure 1. **Left** Boxplots showing the distribution of the categorical feature 'Relationship Status' **Right** Boxplots showing the distribution of the categorical feature 'Race'. **Left of each**: original representation $\mathbf{x} \in \mathcal{X}$. **Right of each**: fair representation $\tilde{\mathbf{x}} \in \mathcal{X}$.

such, we only demonstrate (fair) classifier 7 on the unmodified $\mathbf{x}$.

Benchmarking We train our model for $50,000$ iterations using a network with 1 hidden layer of 40 nodes for both the encoder and decoder, with the encoded representation being 40 nodes. The predictor acts on the decoded output of this network. We set the trade-off parameters of the reconstruction loss (λ_1) and decomposition loss (λ_2) to 10^{-4} and 100 respectively. We then use this model to translate 10 different training and test sets into $\tilde{\mathbf{x}}$. Using a modified version of the framework provided by Friedler et al. [13] we evaluate methods 1–6 using $\mathbf{x}$ and $\tilde{\mathbf{x}}$ representations. To ensure consistency, we train the model of Beutel et al. [2] with the same architecture and number of iterations as our model.

Table 1 shows the results of these experiments. Our interpretable representation, $\tilde{\mathbf{x}}$ achieves similar fairness level to Beutel's state-of-the-art approach (**Q1**). Consistently, our representation $\tilde{\mathbf{x}}$ promoted the *fairness* criterion (Eq. Opp. close to 0), with only a small penalty in accuracy.

Interpretability We promote equality of opportunity for the positive class (actual salary $>$ \$50K). In Figure 1 we show the effect of learning a fair representation, showing changes in the 'Relationship Status' and 'Race' features of samples that were incorrectly classified by an SVM as earning $<$ \$50K in $\mathbf{x}$, but were correctly classified in $\tilde{\mathbf{x}}$. The visualization can be used for understanding how representation methods adjust the data for fairness. For example in Figure 1 (left) we can see that our method deals with the notorious problem of a husband or wife relationship status being a direct proxy for gender (**Q2**). Our method recognises this across all repeats in an unsupervised manner and reduces the wife category which is associated with a negative prediction. Other categories that have less correlation with the protected characteristic, such as race, largely remain unmodified (Figure 1 (right)).

3.2. The CelebA dataset

Our intention here is to investigate whether (**Q3**) performing an image-to-image translation brings us closer to achieving interpretability in how fairness is being satisfied, and whether (**Q4**) using semantic attribute representations of images reinforces similar interpretability conclusions as using image features directly.

Image-to-image translation Our autoencoder network

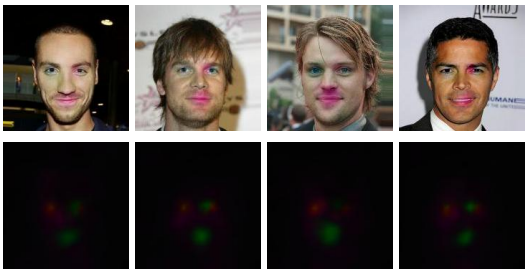


Figure 2. Examples of the translated and residual images on CelebA from the protected group of males (minority group) that have been classified correctly (as attractive) after transformation. These results are obtained with the transformer network for image-to-image translation. Best viewed in color.

	domain $\mathcal{X}$	Acc. ↑	Eq. Opp. ↓	TPR female	TPR male
orig. $\mathbf{x}$	*images*	80.6	**33.8**	90.8	57.0
orig. $\mathbf{x}$	*attributes*	79.1	**39.9**	90.8	50.9
fair $\tilde{\mathbf{x}}$	*images*	79.4	**23.8**	85.2	61.4
fair $\tilde{\mathbf{x}}$	*attributes*	75.9	**12.4**	87.2	74.8
fair $\tilde{\mathbf{x}}$	*fake images*	78.5	**23.0**	87.5	64.5

Table 2. Results on CelebA dataset using a variety of input domains. Prediction performance is measured by accuracy, and we use equality of opportunity, TPRs difference, as the fairness criterion. Here, domain of fake images (last row) denotes images synthesized by the StarGAN[7] model from the original images and their fair attribute representations. We **boldface** Eq. Opp. since this is the fairness criterion.

is based on the architecture of the transformer network for neural style transfer [25] with three convolutional layers, five residual layers and three deconvolutional/upsampling layers in combination with instance weight normalization [44]. The transformer network produces the residual image using a non-linear tanh activation, which is then subtracted from the input image to form the translated fair image $\tilde{\mathbf{x}}$. Similarly to neural style transfer [17, 16, 25], for computing the loss terms, we use the activations in the deeper layers of the 19-layered VGG19 network [41] as feature representations of both input and translated images. Specifically, we use activations in the conv3_1, conv4_1 and conv5_1 layers for computing the decomposition loss, the conv3_1 layer activations for the reconstruction loss, and the activations in the last convolutional layer pool_5 for the prediction loss and when evaluating the performance. Given a 176x176 color input image, we compute the activations at each layer mentioned earlier after ReLU, then we flatten and l_2 normalize them to form features for the loss terms. In the HSIC estimates of the decomposition loss, we use a Gaussian RBF kernel $k(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2)$ width $\gamma = 1.0$ for image features, and $\gamma = 0.5$ for protected characteristics (as one over squared distance in the binary space). To compute the decomposition loss, we add the contributions across the three feature layers. We set the trade-off parameters λ_1 and λ_2 of the reconstruction loss and the decomposition loss, respectively, to 1.0, and the TV regularization strength to 10^{-3}. Training was carried out for 50 epochs with a batch size of 80 images. We use minibatch SGD and apply the Adam solver [27] with learning rate 10^{-3}; our TensorFlow implementation is publicly available[6].

Benchmarking and interpretability We enforce equality of opportunity as the fairness criterion, and we consider attractiveness as the positive label. Attractiveness is what could give someone a job opportunity or an advantaged outcome as defined in [22]. To test the hypothesis that we have learned a fairer image representation, we compare the performance and fairness of a standard SVM classifier trained using original images and the translated fair images. We use activation in the pool_5 layer of the VGG19 network as features for training and evaluating the classifier[7].

We report the quantitative results of this experiment in Table 2 (first and third rows) and the qualitative evaluations of image-to-image translations in Figure 2. From the Table 2 it is clear that the classifier trained on fair/translated images $\tilde{\mathbf{x}}$ has improved over the classifier trained on the original images $\mathbf{x}$ in terms of equality of opportunity (reduction from 33.8 to 23.8) while maintaining the prediction accuracy (79.4 comparing to 80.6). Looking at the TPR values across protected features (females and males), we can see that the male TPR value has increased, but it has an opposite effect for females. In the CelebA dataset, the proportion of attractive to unattractive males is around 30% to 70%, and it is opposite for females; male group is therefore the minority group in this problem. Our method achieves better equality of opportunity measure than the baseline by increasing the minority group TPR value while decreasing the majority group TPR value. To understand the balancing mechanism of TPR values (**Q3**), we visualize a subset of test male images that have been classified correctly as attractive after transformation (those examples were misclassified in the original domain) in Figure 2.

We observe a consistent localized area in face, specifically *lips* and *eyes* regions. The CelebA dataset has a large diversity in visual appearance of females and males (hair style, hair color) and their ethnic groups, so more localized facial areas have to be discovered to equalize TPR values across groups. Lips are very often coloured in female (the majority group) celebrity faces, hence our method, to in-

[6] https://github.com/predictive-analytics-lab/Data-Domain-Fairness

[7] We deliberately evaluate the performance (accuracy and fairness) using an auxiliary classifier instead of using the predictor of the transformer network. Since the emphasis of this work is on representation learning, we should not prescribe what classifier the user chooses on top of learned representation.

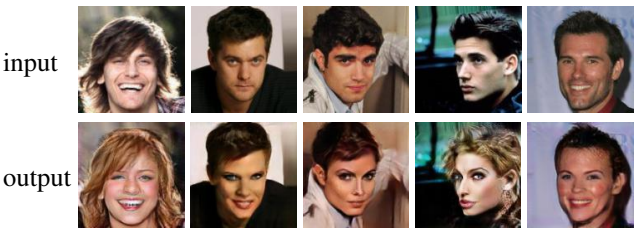


Figure 3. Results of our approach (image-to-image translation via attributes). Given N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^N$, our method transforms them into a new fair dataset $\{(\tilde{\mathbf{x}}^n, y^n)\}_{n=1}^N$ where $(\tilde{\mathbf{x}}^n, y^n)$ is the fair version of $(\mathbf{x}^n, y^n)$. The synthesized images are produced by the StarGAN model [7] conditioned on the original images and their fair attribute representation.

Figure 4. Results of Fainess GAN [40] (Fig.2) of non-attractive (left) and attractive (right) males after pre-processing. Given N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^N$, Fainess GAN transforms them into a new fair dataset $\{(\tilde{\mathbf{x}}^n, \tilde{y}^n)\}_{n=1}^{N'}$ where $N' \neq N$ and $(\tilde{\mathbf{x}}^n, \tilde{y}^n)$ has no correspondence to $(\mathbf{x}^n, y^n)$.

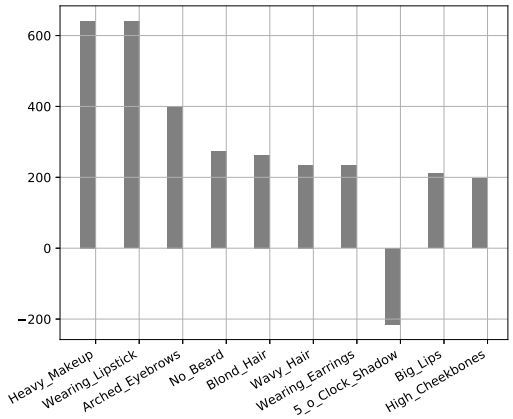


Figure 5. Top 10 semantic attribute features that have been changed in 647 males; those males were *incorrectly* predicted as not attractive, but are now correctly predicted as attractive. 641 and 639 males out of 647 are now with "Heavy_Makeup" and "Wearing_Lipstick" attributes, respectively, and 215 out of 647 males are now *without* a "5_o_Clock_Shadow" attribute.

crease the minority group TPR value, colorizes the lip regions of the minority group (males). Interestingly, female faces without prominent lipstick often got this transformation as well, prompting the decrease in the majority group TPR value. Regarding eye regions, several studies (e.g. [4] and references therein) have shown their importance in gender identification. Also, a heavy makeup that is often applied to female celebrity eyes can also support our visualization in Figure 2.

The image-to-image translation using transformer network learns to produce coarse-grained changes, i.e. masking/colorizing face regions. This is expected as we learn a highly unconstrained mapping from source to target domain, in which the target data is unavailable. To enable fine-grained changes and semantic transformation of the images, we now explore semantic attributes; attributes are well-established interpretable mid-level representations for images. We show how an attribute-to-attribute translation provides an alternative way in analysing and performing an image-to-image translation.

Attribute-to-attribute translation Images in the CelebA dataset come with 40 dimensional binary attribute annotations. We use all but two attributes (*gender* and *attractiveness*) as semantic attribute representation of images. We then perform attribute-to-attribute translation with the transformer network and consider the same attractive versus not attractive task and gender protected characteristic as with the image data. We report the results of this experiment in Table 2 (second and forth rows correspond to the domain of attributes). First, we observe that the predictive performance of the classifier trained on attribute representation is only slightly lower than the performance of the classifier trained on the image data (79.1 versus 80.6), which enables sensible comparison of the results in these two settings. Second, we observe better gain in equality of opportunity when using the transformed attribute representation comparing to transformed images (12.4 is the best Eq. Opp. result in this experiment). This comes at the cost of a drop in accuracy performance. The TPR rates for both groups are higher when using translated attribute representation than when using translated image representation (third row versus fourth row). The largest improvement of the TPR is observed in the group of males (from 50.9 in the original attribute to 74.8 in the translated attribute space). Further analysis of changes in attribute representation reveals that equality of opportunity is achieved by putting *lipstick* and *heavy-makeup* to the male group (Figure 5). These top 2 features have been mostly changed in the group of *males*. Very few changes happened in the group of females. This is encouraging as we have just arrived at the same conclusion (Figures 2 and 5), be it using images or using semantic attributes (**Q4**).

Image-to-image translation via attributes Given the remarkable progress that has been made in the field towards image synthesis with the conditional GAN models, we attempt to synthesize images with respect to the attribute description. Specifically, we use the StarGAN

model [7], the state-of-the-art model for image synthesis with multi-attribute transformation, to synthesize images with our learned fair attribute representation. For this, we pre-train the StarGAN model to perform image transformations with 38 binary attributes (excluding gender and attractive attributes) using training data. We then translate all images in CelebA with respect to their fair attribute representation. We evaluate the performance of this approach and report the results in Table 2 (last row). We also include the qualitative evaluations of image-to-image translations via attributes in Figure 3. These visualizations essentially generalize counterfactual explanations in the sense of [45] to the image domain. We have just shown the "closest synthesized world", i.e. the smallest change to the world that can be made to obtain a desirable outcome. Overall, the classifier trained using this fair representation shows similar Eq. Opp. performance and comparable accuracy to the classifier trained on representation learned with the transformer network. However, the TPR rates for both protected groups are higher (last row versus third row), especially in the group of males, when using this representation.

Pre-processing approaches The aim of the pre-processing approaches such as [40, 6] is to transform the given dataset of N i.i.d. samples $\{(\mathbf{x}^n, y^n)\}_{n=1}^N$ into a *new* fair dataset $\{(\tilde{\mathbf{x}}^n, \tilde{y}^n)\}_{n=1}^{N'}$. It is important to note that N' is not necessarily equal to N, and therefore $(\tilde{\mathbf{x}}^n, \bar{y}^n)$ has no correspondence to $(\mathbf{x}^n, y^n)$. [6] has proposed this approach for tabular (discrete) data, while [40] has explored image data. Here, we offer a unified framework for tabular (continuous and discrete) and image data that transforms the given dataset $\{(\mathbf{x}^n, y^n)\}_{n=1}^N$ into a new fair dataset $\{(\tilde{\mathbf{x}}^n, y^n)\}_{n=1}^N$ where $(\tilde{\mathbf{x}}^n, y^n)$ is the fair version of $(\mathbf{x}^n, y^n)$. *What is the advantage of creating a fair representation per sample (our method) rather than on the whole dataset at once [40, 6]?* The first can be used to provide an *individual*-level explanation of fair systems, while the latter can only be used to provide a *system*-level explanation. For comparison, we include here a snapshot of results presented in [40] using the CelebA dataset in Figure 4. The figure shows eigenfaces/eigensketches with *the mean image* of the new fair dataset $\{(\tilde{\mathbf{x}}^n)\}_{n=1}^{N'}$ (in the center) of the 3×3 grid. No per sample visualisation $(\tilde{\mathbf{x}}^n)$ was provided. Left/right/top/bottom images in Fig. 4 show variations along the first/second principal components. In contrast, Figure 3 shows a per sample visualisation $(\tilde{\mathbf{x}}^n)$ using our proposed method.

3.3. The Diversity in Faces dataset

We extract and align face crops from the images and use 128x128 facial images as the inputs. Our preliminary experiment has similar setup to the image-to-image translation on the CelebA dataset except that the prediction task has seven age groups to be classified. As the fairness criterion we enforce equality of opportunity considering the middle age group (31-45) to be desirable (as the positive label when conditioning). As before, to test the hypothesis that we have learned a fairer image representation, we compare the performance and fairness of the SVM classifier trained using original images and the translated fair images (with features as activations in the pool_5 layer of the VGG19 network). We achieve 52.85 as the overall classification accuracy over seven age groups when using original image features and an increased 60.26 accuracy when using translated images. The equality of opportunity improved from 27.21 using original image representation to 9.85 using fair image representation. Similarly to the CelebA dataset, the image-to-image translation using transformer network learns to produce coarse-grained changes, i.e. masking/colorizing nose regions (as opposed to lips and eyes regions on CelebA). These preliminary results are encouraging and further analysis will be addressed as a future extension.

4. Discussion and Conclusion

> *It is not clear if fairness and interpretability are conflicting requirements.* Reviewer #1

They are not, however interpretability in how fairness is enforced has so far been overlooked despite being an integral ingredient for encouraging productive public debates regarding fair machine learning systems. Interpretability in machine learning models can help to ascertain qualitatively *whether* fairness is met [46, 10]. This paper takes a step further and advocates interpretability to ascertain qualitatively *how* fairness is met, once we have agreed to enforce fairness (e.g. equality of opportunity) in machine learning models. We specifically focus on enforcing fairness in representation learning. Unlike other fair representation learning methods that learn *latent* embeddings, our method learns a representation that is in the same space as the original input data, therefore retaining the semantics of the input domain. Our method picks up consistently in 10 *out of* 10 *repeated experiments* whether a person is a husband or wife as a direct proxy for *gender*, and subsequently reduces the wife category which is associated with a negative prediction. In our experiments with people's faces, eyes and lips are considered to be the direct proxy for gender attractiveness, and nose regions for being in a certain age group. As a potential future direction, we plan to further analyze the interpretability in fairness using causal reasoning [31].

Acknowledgments

NQ is supported by the UK EPSRC project EP/P03442X/1 and the Russian Academic Excellence Project' 5-100'. VS is supported by the Imperial College Research Fellowship. We gratefully acknowledge NVIDIA for GPUs donation and Amazon for AWS Cloud Credits.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018. 4, 5

[2] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017. 1, 2, 4, 5

[3] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. Fair pipelines. *CoRR*, abs/1707.00391, 2017. 4

[4] Elizabeth Brown and David I Perrett. What gives a face its gender? *Perception*, 22(7):829–840, 1993. 1, 7

[5] V. Bruce, A. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. Sex discrimination: how do we tell the difference between male and female faces? *Perception*, 22, 1993. 1

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3992–4001, 2017. 8

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6, 7, 8

[8] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5, 2017. 2

[9] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. 4

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *CoRR*, abs/1702.08608, 2017. 8

[11] Cynthia Dwork and Christina Ilvento. Fairness under composition. In *Innovations in Theoretical Computer Science (ITCS)*, 2019. 4

[12] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations (ICLR)*, 2016. 2

[13] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. *CoRR*, abs/1802.04422, 2018. 5

[14] Siyao Fu, Haibo He, and Zeng-Guang Hou. Learning race from face: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 2483–2509, 2014. 1

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 17:2096–2030, 2016. 2

[16] Jacob R Gardner, Paul Upchurch, Matt J Kusner, Yixuan Li, Kilian Q Weinberger, Kavita Bala, and John E Hopcroft. Deep manifold traversal: Changing labels with convolutional features. In *European Conference on Computer Vision (ECCV)*, 2016. 6

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. 4, 6

[18] Global Future Council on Human Rights 2016-18. How to prevent discriminatory outcomes in machine learning. Technical report, World Economic Forum, 2018. 1, 2

[19] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 13:723–773, 2012. 3

[20] Arthur Gretton, Olivier Bousquet, Alexander J. Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory ALT*, 2005. 2, 3

[21] N. Grgic-Hlaca, E. Redmiles, K. P. Gummadi, and A. Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *The Web Conference (WWW)*, 2018. 2

[22] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS) 29*, 2016. 1, 6

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 4, 6

[26] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2012. 4, 5

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 6

[28] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *CoRR*, abs/1609.05807, 2016. 2

[29] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015. 1

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 4

[31] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[32] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. In *International Conference on Learning Representations (ICLR)*, 2016. 2, 3

[33] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*, 2018. 1, 2, 4

[34] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 4

[35] Michele Merler, Nalini K. Ratha, Rogério Schmidt Feris, and John R. Smith. Diversity in faces. *CoRR*, abs/1901.10436, 2019. 4

[36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 1

[37] Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *International Conference on Machine Learning (ICML)*, 2009. 3

[38] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 3

[39] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2008. 4

[40] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. Fairness GAN. *arXiv*, arXiv:1805.09910, 2018. 2, 7, 8

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[42] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research (JMLR)*, 13:1393–1434, 2012. 3

[43] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *International Conference on Machine Learning (ICML)*, 2016. 4

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 4, 6

[45] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2), 2018. 8

[46] Working Group. Machine learning: the power and promise of computers that learn by example. Technical report, The Royal Society, 2017. 8

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web (WWW)*, pages 1171–1180, 2017. 1, 4, 5

[48] Richard S. Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, 2013. 2, 4

[49] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018. 1

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017. 2