Utility-Fairness Trade-Offs and How to Find Them

Sepehr Dehdashtian Bashir Sadeghi Vishnu Naresh Boddeti
Michigan State University
{sepehr, sadeghib, vishnu}@msu.edu

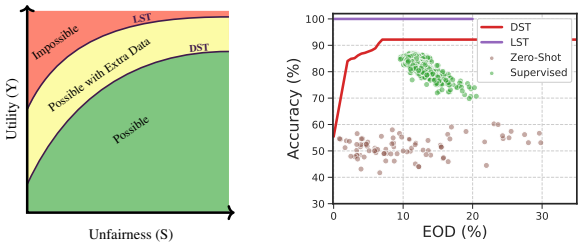
Abstract

When building classification systems with demographic fairness considerations, there are two objectives to satisfy: 1) maximizing utility for the specific task and 2) ensuring fairness w.r.t. a known demographic attribute. These objectives often compete, so optimizing both can lead to a trade-off between utility and fairness. While existing works acknowledge the trade-offs and study their limits, two questions remain unanswered: 1) What are the optimal trade-offs between utility and fairness? and 2) How can we numerically quantify these trade-offs from data for a desired prediction task and demographic attribute of interest? This paper addresses these questions. We introduce two utility-fairness trade-offs: the Data-Space and Label-Space Trade-off. The trade-offs reveal three regions within the utility-fairness plane, delineating what is fully and partially possible and impossible. We propose U-FaTE, a method to numerically quantify the trade-offs for a given prediction task and group fairness definition from data samples. Based on the trade-offs, we introduce a new scheme for evaluating representations. An extensive evaluation of fair representation learning methods and representations from over 1000 pre-trained models revealed that most current approaches are far from the estimated and achievable fairness-utility trade-offs across multiple datasets and prediction tasks.

1. Introduction

As learning-based systems are increasingly being deployed in high-stakes applications, there is a dire need to ensure that they do not propagate or amplify any discriminative tendencies inherent in the training datasets. An ideal solution would impart fairness to prediction models while retaining the performance of the same model when learned without fairness considerations.

Realizing this goal necessitates optimizing two objectives: maximizing utility in predicting a label Y for a target task (e.g., face identity) while minimizing the unfairness w.r.t. a demographic attribute S (e.g., age or gender). However, when the statistical dependence between Y and S is not negligible, learning with fairness considerations will necessarily degrade the performance of the target predictor, i.e., *a trade-off will exist between utility and fairness*.



(a) (b)

Figure 1. **The utility-fairness trade-offs.** (a) Classification systems can be evaluated by their utility (e.g., accuracy) w.r.t. a target label Y and their unfairness w.r.t. a demographic label S. We introduce two trade-offs, *Data Space Trade-Off* (DST) and *Label Space Trade-Off* (LST). (b) We empirically estimate DST and LST on CelebA and evaluate the utility (high cheekbones) and fairness (gender & age) of over 100 zero-shot and 900 supervised models.

The existence of a utility-fairness trade-off has been well established, theoretically [11, 22, 29, 36, 37] and empirically [29], in multiple prior works. However, the focus of this body of work has been limited in multiple respects. First, prior work [29, 36] focused on just one type of trade-off, ignoring other possible trade-offs between utility and fairness. Second, prior work [36, 37] focused on establishing bounds or identifying the end-points of the trade-off of interest rather than attempting its precise characterization. Third, the majority of the prior work [11, 29, 36, 37] has investigated the utility-fairness trade-offs for one definition of group fairness, namely, demographic parity (DP). There are multiple fairness definitions [2], including those more practically relevant than DP, such as Equalized Opportunity (EO), for which the trade-offs have not been studied.

Despite these attempts, several questions related to the utility-fairness trade-offs remain outstanding.

1. *What are the optimal utility-fairness trade-offs?*
2. *For a given prediction task and a demographic attribute, we wish to be fair w.r.t., how can we empirically estimate the trade-offs from data?*

Addressing these questions by identifying and quantifying the trade-offs is the primary goal of this paper. The trade-offs are a function of the data triplet (X, Y, S), where X is the data (e.g., images), Y is the target label, and S is the sensitive demographic label. Figure 1 illustrates the plausible trade-offs, their empirical estimation on CelebA [19], and their utility in empirically evaluating representations from pre-trained models.

Identifying Trade-Offs (§3). We identify two trade-offs: the *Label-Space Trade-Off* (LST) and *Data-Space Trade-Off* (DST). They can be defined for *any* group fairness definitions that can be expressed via *independence* and *separation* relations [2]. The LST corresponds to the trade-off obtained by an *oracle fair classifier* that depends only on the distributions of Y and S. Similarly, DST is the trade-off obtained by an *optimally learned fair classifier* and depends on (X, Y, S). By definition, LST necessarily dominates DST since it does not depend on the data X.

The trade-offs divide the utility-fairness plane into three regions shown in Fig. 1a. A *possible* region that can be attained by algorithms learned on the given data, a *possible with extra data* region that can be attained by learning on data beyond the given data, and an *impossible* region that cannot be attained by any algorithmic scheme due to the inherent dependence between the distributions of Y and S.

Quantifying Trade-Offs (§4). Characterizing the exact trade-offs from data for a given task, demographic attribute, and fairness definition affords multiple benefits. It will allow researchers and practitioners to identify the achievable solution space for the task, evaluate how far a given predictor is from the optimal solution, and identify performance gaps and trends among existing solutions. To this end, we propose U-FaTE (Utility-Fairness Trade-Off Estimator), a method for quantifying the trade-offs from data triplets numerically. U-FaTE is an end-to-end model that adopts a statistical dependence measure as a proxy for utility and fairness and optimizes their weighted linear combination. U-FaTE can be flexibly adapted to estimate both the DST and LST from a finite labeled dataset.

Usefulness of Trade-Offs (§5). The trade-offs illuminate the fundamental limits of learning algorithms in mitigating unfairness and present a new avenue to evaluate a given image representation in terms of its distance from the estimated trade-offs. We adopt this scheme to evaluate the representations of over 900 supervised and 100 zero-shot publicly available pre-trained models, derive insights, and identify trends and models that are close and far from the empirical trade-off estimates (§6).

Notation: We denote scalars using lowercase letters, e.g., d and λ. We denote deterministic vectors by boldface lowercase letters, e.g., $\boldsymbol{x}$, $\boldsymbol{y}$. Both scalar-valued and multidimensional random variables (RV)s are denoted by regular upper case letters, e.g., X, Y. We denote deterministic matrices by boldface upper case letters, e.g., $\boldsymbol{K}$, $\boldsymbol{\Theta}$. Finite or infinite sets are denoted by calligraphic letters, e.g., $\mathcal{A}$, $\mathcal{H}$.

2. Related Works

A vast majority of prior work on designing fair classifiers focused primarily on uncovering disparities in practical tasks [5, 31] and learning a fair predictor [20, 32, 35] for a given fairness measure. An extended discussion of this body of work can be found in the supplementary material.

Utility-Fairness Trade-Offs: Many attempts on learning fair models [10, 20, 31, 32] ignored the existence of trade-offs. They sought to maximize accuracy on target tasks while minimizing unfairness, thus perhaps seeking an infeasible solution. Most studies on utility-fairness trade-offs are theoretical and under restricted settings in terms of the type of labels, notion of fairness, and bounds or extreme limits of trade-offs. For example, Zhao *et al.* [38] obtained a lower bound on DST when both Y and S are binary labels. McNamara *et al.* [21] provided both upper and lower bounds for binary labels. Only a couple of attempts [27, 29] have been made to numerically estimate utility-fairness trade-offs for *independence* related-based measures like demographic parity, both of them on features from pre-trained models, rather than raw data. Sadeghi *et al.* [27] obtained a simplified version of DST, but for linear models. Later on, in the context of invariant representation learning, this was extended to estimate a near-optimal DST-like trade-off called $\mathcal{T}_{\text{Opt}}$ in [29]. But as we demonstrate in §6.2, the estimate of $\mathcal{T}_{\text{Opt}}$ called K-$\mathcal{T}_{\text{Opt}}$ does not span the entire trade-off.

In contrast to this body of work, we identify two types of trade-offs, DST and LST, and propose a method to numerically quantify them from data. Our trade-offs and their empirical estimates apply to a wide range of prediction tasks for two different categories of fairness notions without any restrictions on the type of labels.

Learning Fair Classifiers: Over the last decade, many methods have been developed for learning fair classifiers. These approaches follow the template of adopting a fairness constraint as a regularizer in addition to the objective for the target task. The approaches differ in the choice of measure as a proxy for quantifying the level of unfairness between the target label Y and the prediction $\hat{Y}$, and the associated optimization technique. From an optimization perspective, they can be classified into three major categories–i.e., iterative adversarial methods (ARL[34], SARL[27], and MaxEnt-ARL[26]), non-iterative adversarial methods (FairHSIC [23], OptNet-ARL [28]), and closed-form solver methods (SARL [27], K-$\mathcal{T}_{\text{Opt}}$ [29], LEACE [3], FairerCLIP [7]). Among these, ARL, SARL, MaxEnt-ARL, and OptNet-ARL measure mean dependence [1, 12], FairHSIC, FairerCLIP and K-$\mathcal{T}_{\text{Opt}}$

measure full statistical dependence, i.e., all modes of dependence, and SARL and LEACE measures linear dependence.

U-FaTE draws inspiration from K-$\mathcal{T}_{\text{Opt}}$ [29]. By using a closed-form solver and a universal dependence measure that captures all non-linear dependencies, K-$\mathcal{T}_{\text{Opt}}$ achieves a better utility-fairness trade-off and is more stable than the other fair learning methods discussed above. However, K-$\mathcal{T}_{\text{Opt}}$ is limited in multiple respects and cannot be directly employed for estimating the trade-offs. 1) It operates on features and does not generalize to learning directly from high-dimensional raw data representations such as pixels for images. 2) K-$\mathcal{T}_{\text{Opt}}$ optimizes an unconditional dependence measure, which limits its applicability to *independence* relation-based fairness definitions such as demographic parity. 3) As we demonstrate in §6.2, for demographic parity, K-$\mathcal{T}_{\text{Opt}}$'s trade-off is the closest to our DST estimate, but it does not span the entire utility-fairness trade-off front. Therefore, we adopt the positive aspects of K-$\mathcal{T}_{\text{Opt}}$, namely universal dependence measure and closed-form solver, into U-FaTE and overcome its drawbacks.

3. The Utility-Fairness Trade-Offs

Fairness Notions: Group fairness notions are typically categorized into three classes [2], namely *independence*, *separation*, and *sufficiency*, each corresponding to different societal desiderata. We focus on the *idependence* and *separation* relations, which can be expressed as independence ($\hat{Y} \perp\!\!\!\perp S$) and conditional independence ($\hat{Y} \perp\!\!\!\perp S|Y = y$) relations, respectively.

We consider frequently used fairness criteria including Demographic Parity (DP) [17] which is an example of an *independence* relation, and Equalized Opportunity (EO) [14], and Equality of Odds (EOO) [14] which are both examples of *separation* relations. The corresponding unfairness metrics are Demographic Parity Violation $DPV := |P(\hat{Y} = 1|S = 0) - P(\hat{Y} = 1|S = 1)|$, Equalized Opportunity Difference $EOD := |P(\hat{Y} = 1|Y = 1, S = 0) - P(\hat{Y} = 1|Y = 1, S = 1)|$, and Equality of Odds Difference $EOOD := \frac{1}{2}\sum_{y\in\{0,1\}} |P(\hat{Y} = 1|Y = y, S = 0) - P(\hat{Y} = 1|Y = y, S = 1)|$, respectively.

We now introduce the *Data-Space Trade-Off* and the *Label-Space Trade-Off*. In both, we employ a dependence measure $\text{Dep}(\cdot,\cdot|\cdot)$ to enforce the *independence* and *separation* based fairness constraints. The function $\text{Dep}(\cdot,\cdot|\cdot) \geq 0$ is a parametric or non-parametric measure of statistical dependence. $\text{Dep}(P, Q|R = r) = 0$ implies that conditioned on $R = r$, the random variables (RVs) P and Q are independent. $\text{Dep}(P, Q|R = r) > 0$ means that conditioned on $R = r$, P, and Q are dependent, with larger values indicating larger degrees of dependence. When r is the empty set $\emptyset$, we assume that $\text{Dep}(P, Q|R = \emptyset)$ simply reduces to the unconditional dependence $\text{Dep}(P, Q)$.

Definition 1. *Data Space Trade-Off* (DST)

$$f_\lambda^{DST} := \arg\inf_{f\in\mathcal{H}_X} \Big\{(1-\lambda) \inf_{g_Y\in\mathcal{H}_Y} \mathbb{E}_{X,Y}\left[\mathcal{L}_Y\left(g_Y\left(\boldsymbol{f}(X)\right), Y\right)\right] + \lambda \,\text{Dep}\big(\boldsymbol{f}(X), S|Y = y\big)\Big\}, \quad 0 \leq \lambda < 1$$

Here f is the encoder that maps data X to a representation Z, and g is a classifier that predicts $\hat{Y}$ from Z. $\mathcal{H}_X$ and $\mathcal{H}_Y$ are the hypothesis classes of functions for f and g respectively. $\mathcal{L}_Y(\cdot,\cdot)$ is the loss function corresponding to the utility, and λ controls the trade-off between utility and fairness, i.e., $\lambda = 0$ corresponds to ignoring the fairness constraint and only optimizing the utility, while, $\lambda \to 1$ corresponds to the total fairness. The outcome f_λ^{DST} corresponds to the encoder for a given value of λ. This definition corresponds to the **DST** curve in Fig. 1a, where the utility-fairness plane below the DST corresponds to the region achievable by algorithms designed for this prediction task that learn from the data triplet $(X, Y, S) \sim p(X, Y, S)$.

Definition 2. *Label Space Trade-Off* (LST)

$$Z_\lambda^{LST} := \arg\inf_{Z\in L^2} \Big\{(1-\lambda) \inf_{g_Y\in\mathcal{H}_Y} \mathbb{E}_Y\Big[\mathcal{L}_Y\big(g_Y(Z), Y\big)\Big] + \lambda \,\text{Dep}\big(Z, S|Y = y\big)\Big\}, \quad 0 \leq \lambda < 1$$

Here L^2 is the space of all square-integrable RVs $\big($i.e. $\mathbb{E}_Z\left[\|Z\|^2\right] < \infty\big)$ in the probability space generated by the joint RV (Y, S). LST corresponds to the trade-off from an *ideal* representation space Z_λ^{LST} that is not constrained to be learned from the input data X. It is the trade-off *inherent* to the task itself and is the best that *any* algorithm can hope to achieve for this task. Therefore, it necessarily dominates (or is equal to) DST in Definition 1. This definition corresponds to the LST curve in Fig. 1a, where the utility-fairness plane above the LST corresponds to the region that *any* algorithm cannot achieve.

We stress that the above trade-offs are intrinsic to the underlying data, specifically the underlying distributions that generated that data. So, *the trade-offs are a property of the data, not of any particular learning algorithm.*

The LST and DST Divide: As illustrated by the yellow region in Fig. 1a, there is a potential gap between LST and DST. This gap at $\lambda = 0$ stems from the irreducible error from the prediction $\mathbb{E}[Y|X]$ of a Bayes Classifier, or when Y is fully recoverable from X. And, when $\lambda > 0$, the gap between LST and DST widens in two scenarios: 1) $Y \not\perp\!\!\!\perp S$: The model starts discarding S from the representation Z, which will lead to Y being even less recoverable from X compared to $\lambda = 0$. and 2) $Y \perp\!\!\!\perp S$: If X entangles Y and S in such a way that Y is not recoverable from X when S is discarded, it will lead to Y being even less recoverable from X compared to $\lambda = 0$.

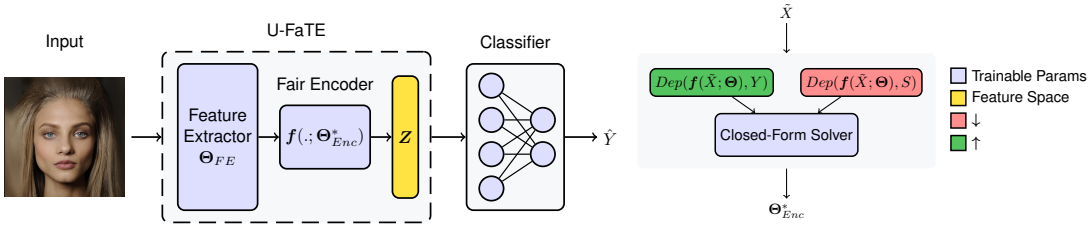


Figure 2. **Overview of U-FaTE:** (Left) It comprises two components, a feature extractor and a fair encoder, that are trained end-to-end. Once U-FaTE is trained, the MLP classifier is trained to predict Y from which fairness metrics can be computed. (Right) The fair encoder parameters are optimized through a closed-form solver operating on the features from the feature extractor. See text for more details.

4. Numerically Quantifying the Trade-Offs

Now, we turn to the second goal of this paper, numerically quantifying the trade-offs from data. Fig. 2 shows a high-level overview of U-FaTE to learn a fair representation for a given trade-off parameter λ. U-FaTE comprises a feature extractor and a fair encoder. It receives raw data as input and uses a feature extractor to provide features for the fair encoder. The encoder uses the extracted features and employs a closed-form solver to find the optimum function that maps these features to a new feature space that minimizes the dependency on the sensitive attribute while maximizing the dependency on the target attribute. Following this, to predict the target Y, a classifier is trained with the standard cross-entropy loss for classification problems. This process is repeated for multiple values of λ with $0 \leq \lambda < 1$ to obtain the full trade-off curves.

4.1. Problem Setup

We start from Definition 1 and model the function $\boldsymbol{f}$ as a composition $\boldsymbol{f}_{FE} \circ \boldsymbol{f}_{Enc}$ of the feature extractor and a fair encoder i.e., $\boldsymbol{f}(X;\boldsymbol{\Theta}) = f_{Enc}(f_{FE}(X;\boldsymbol{\Theta}_{FE});\boldsymbol{\Theta}_{Enc})$. We parameterize $\boldsymbol{f}$ with $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_{FE};\boldsymbol{\Theta}_{Enc}]$ where $\boldsymbol{\Theta}_{FE}$ are the parameters of $\boldsymbol{f}_{FE}$ and $\boldsymbol{\Theta}_{Enc}$ are the parameters of $\boldsymbol{f}_{Enc}$. The objective function in Definition 1 is now

$$\min_{\boldsymbol{\Theta}} \Big\{ (1-\lambda) \inf_{\boldsymbol{\Theta}_Y} \mathbb{E}_{X,Y} \left[L_Y \left(g_Y \left(\boldsymbol{f}(X;\boldsymbol{\Theta});\boldsymbol{\Theta}_Y \right), Y \right) \right] + \lambda \operatorname{Dep} \left(\boldsymbol{f}\left(X_c;\boldsymbol{\Theta}\right), S_c \right) \Big\}, \quad 0 \leq \lambda < 1. \quad (1)$$

where $\operatorname{Dep}\left(\boldsymbol{f}\left(X_c;\boldsymbol{\Theta}\right), S_c\right)$ is equivalent to the term $\operatorname{Dep}(\boldsymbol{f}(X), S|Y = y)$ in Definition 1 when Y is not a continuous label. In this case, $X_c \sim P(X|Y = y)$ and $S_c \sim P(S|Y = y)$ are the random variables that represent the data and sensitive attribute conditioned on $Y = y$, respectively. The fair representation is $Z = \boldsymbol{f}(X;\boldsymbol{\Theta})$.

4.2. Optimization via Dependence Measures

The formulation in (1) can be directly optimized for an appropriate choice of dependence measure. Different choices of Dep lead to different fair representation learning methods. For instance, measuring Dep through an adversary leads to the class of adversarial representation learning (ARL) methods [26–28, 34]. Similarly, employing the Hilbert Schmidt Independence Criterion (HSIC) [13] as Dep leads to FairHSIC [23]. However, due to challenges in optimization [26–28] and as we demonstrate in §6.2, these approaches are either very unstable, fail to span the trade-off or lead to sub-optimal trade-offs.

Recently, Sadeghi *et al.* [29] demonstrated that adopting an HSIC-like dependence measure for the fairness objective and the target loss leads to a closed-form solution that is both efficient and effective at finding a near-optimal trade-off. Therefore, we incorporate the HSIC-like dependence measure and the closed-form solver into U-FaTE. Thus (1) can be expressed as,

$$\sup_{\boldsymbol{f} \in \mathcal{A}_r} \Big\{ (1-\lambda) \operatorname{Dep} \left(\boldsymbol{f}(X;\boldsymbol{\Theta}), Y \right) - \lambda \operatorname{Dep} \left(\boldsymbol{f}(X_c;\boldsymbol{\Theta}), S_c \right) \Big\}, \quad (2)$$

where $\mathcal{A}_r$ is a function space that encourages the representations to be uncorrelated. It does not affect the optimality of the learned encoder [29] and improves the compactness of representation [4]. Note that while the first term involves all data X, the second involves the conditional data X_c.

Choice of Dependence Measure: We adapt the dependence measure from [29] since it lends itself to a closed-form solution while capturing linear and non-linear dependencies under mild assumptions. While the dependence measure in [29] has been defined for absolute independence, our formulation in (2) also requires conditional independence to be compatible with *separation* based fairness definitions. Therefore, when Y is not a continuous label, we define the conditional dependence measure as,

$$\operatorname{Dep}(\boldsymbol{f}(X), S|Y = y) := \sum_{j=1}^{r} \sum_{\beta_S \in \mathcal{U}_S} \mathbb{E}\left[\left(f_j(X_c) - \mathbb{E} f_j(X_c) \right) \left(\beta_S(S_c) - \mathbb{E}\beta_S(S_c) \right) \right] \quad (3)$$

where $\mathcal{U}_S$ is a countable orthonormal basis set for the separable universal RKHS $\mathcal{H}_S$ and $X_c \sim P(X|Y = y)$ and

$S_c \sim P(S|Y=y)$ are data and sensitive attributes, respectively. *Empirically* it can be estimated as,

$$\text{Dep}(\boldsymbol{f}(X), S|Y=y) := \frac{1}{n^2} \left\| \boldsymbol{\Theta} \boldsymbol{K}_{X_c} \boldsymbol{H} \boldsymbol{L}_{S_c} \right\|_F^2, \quad (4)$$

where n is the number of data samples, $\boldsymbol{K}_{X_c} \in \mathbb{R}^{n \times n}$ is the Gram matrix corresponding to $\mathcal{H}_X$, $\boldsymbol{\Theta}$ is the encoder parameter in $\boldsymbol{f}(X) = \boldsymbol{\Theta}[k_{X_1}, k_{X_2}, \cdots, k_{X_n}]^T$, $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the centering matrix, and $\boldsymbol{L}_{S_c}$ is a full column-rank matrix such that $\boldsymbol{L}_{S_c}\boldsymbol{L}_{S_c}^T = \boldsymbol{K}_{S_c}$ (Cholesky factorization).

4.3. A Solution to the Optimization Problem

Closed-Form Solver via Functions in RKHSs: Directly solving for all the parameters $\boldsymbol{\Theta}$ through (2) and (4) leads to abysmal performance in practice since the kernel $\boldsymbol{K}_X$ has to be computed over the raw data space. Therefore, we instead define the fair encoder on the co-domain of the feature extractor $\boldsymbol{f}(\cdot; \boldsymbol{\Theta}_{FE})$. So, in this case, (2) reduces to,

$$\sup_{\boldsymbol{f}_{Enc} \in \mathcal{A}_r} \Big\{ \quad (1-\lambda)\,\text{Dep}\left(\boldsymbol{f}_{Enc}(\tilde{X}; \boldsymbol{\Theta}_{Enc}), Y\right) - \lambda\,\text{Dep}\left(\boldsymbol{f}_{Enc}(\tilde{X}_c; \boldsymbol{\Theta}_{Enc}), S_c\right) \Big\}, \quad (5)$$

where $\tilde{X} = f(X; \boldsymbol{\Theta_{FE}})$, and the first and second terms are $\frac{1}{n^2}\left\|\boldsymbol{\Theta}_{Enc}\boldsymbol{K}_{\tilde{X}}\boldsymbol{H}\boldsymbol{L}_Y\right\|_F^2$ and $\frac{1}{n^2}\left\|\boldsymbol{\Theta}_{Enc}\boldsymbol{K}_{\tilde{X}_c}\boldsymbol{H}\boldsymbol{L}_{S_c}\right\|_F^2$, respectively. The parameters $\boldsymbol{\Theta}_{Enc}$ can now be solved exactly via a closed-form solution:

Theorem 1. *A global optimizer of (5) is*

$$\boldsymbol{f}_{\mathcal{H}_{\tilde{X}}}^{opt}(\tilde{X}; \boldsymbol{\Theta}_{Enc}) = \boldsymbol{\Theta}_{Enc}^{opt}\left[k_{\tilde{X}}(\tilde{\boldsymbol{x}}_1, \tilde{X}), \cdots, k_{\tilde{X}}(\tilde{\boldsymbol{x}}_n, \tilde{X})\right]^T$$

where $\boldsymbol{\Theta}_{Enc}^{opt} = \boldsymbol{U}^T\boldsymbol{L}_{\tilde{X}}^{\dagger} \in \mathbb{R}^{r \times n}$ *and the columns of* $\boldsymbol{U}$ *are eigenvectors corresponding to the* r *largest eigenvalues of the following generalized eigenvalue problem.*

$$\left((1-\lambda)\boldsymbol{L}_{\tilde{X}}^T\boldsymbol{H}\boldsymbol{K}_Y\boldsymbol{H}\boldsymbol{L}_{\tilde{X}} - \lambda\boldsymbol{L}_{\tilde{X}_c}^T\boldsymbol{H}\boldsymbol{K}_{S_c}\boldsymbol{H}\boldsymbol{L}_{\tilde{X}_c}\right)\boldsymbol{u} = \lambda\left(\frac{1}{n}\boldsymbol{L}_{\tilde{X}}^T\boldsymbol{H}\boldsymbol{L}_{\tilde{X}} + \gamma\boldsymbol{I}\right)\boldsymbol{u}. \quad (6)$$

Here $\boldsymbol{L}_{\tilde{X}}\boldsymbol{L}_{\tilde{X}}^T = \boldsymbol{K}_{\tilde{X}}$, $\tilde{X}_c \sim p(\tilde{X}|Y=y)$ *and* $S_c \sim p(S|Y=y)$.

Proof. The objective in (5) reduces to a generalized eigenvalue problem [18] by expressing it as a trace optimization problem. See supplementary for detailed proof. □

While this is a general solution to (5), the solution for each group fairness case is detailed in the supplementary.

Alternating Optimization: Now we present our full algorithm to optimize (2). We adopt standard minibatch to learn the feature encoder's parameters $\boldsymbol{\Theta}_{FE}$ and the closed-form solver for the fair encoder parameters $\boldsymbol{\Theta}_{Enc}$. We optimize them alternatively where in each iteration, we update $\boldsymbol{\Theta}_{Enc}$ while freezing $\boldsymbol{\Theta}_{FE}$ and vice-versa. Specifically, to optimize the fair encoder's parameters $\boldsymbol{\Theta}_{Enc}$, we extract features from the data using the frozen feature extractor and use the closed-form solution in (5) to update $\boldsymbol{\Theta}_{Enc}$. Then, we update the feature extractor's parameters $\boldsymbol{\Theta}_{FE}$ through minibatch SGD in (2) while freezing the encoder parameters. We repeat this process for every minibatch iteration. More details and an illustration of this alternating algorithm can be found in the supplementary material.

4.4. Numerically Estimating the LST

The Label Space Trade-off (LST) arises when the representation Z is not restricted to be a function of X. Following the discussion in the previous subsection, this trade-off can be formulated as,

$$\sup_{Z \in L_r^2} \Big\{(1-\lambda)\,\text{Dep}\big(Z, Y\big) - \lambda\,\text{Dep}\big(Z, S|Y=y\big)\Big\}, \quad (7)$$

where L_r^2 is the space of all RVs of dimension r with finite variance, i.e., $\mathbb{E}_Z\left[\left\|Z - \mathbb{E}[Z]\right\|^2 < \infty\right]$. From (7), observe that the optimal Z is a function of $\boldsymbol{p}_{Y,S}$ only. Therefore, instead of directly optimizing Z over L_r^2, equivalently, we optimize for $\boldsymbol{\Theta}_{FE}$ and $\boldsymbol{\Theta}_{Enc}$ as

$$\max_{\boldsymbol{\Theta}_{FE}, \boldsymbol{\Theta}_{Enc}} \Big\{(1-\lambda)\,\text{Dep}\left(\boldsymbol{f}\left(Y, S; \boldsymbol{\Theta}_{FE}, \boldsymbol{\Theta}_{Enc}\right), Y\right) + \lambda\,\text{Dep}\left(\boldsymbol{f}\left(Y, S; \boldsymbol{\Theta}_{FE}, \boldsymbol{\Theta}_{Enc}\right), S|Y=y\right)\Big\}. \quad (8)$$

Here $\boldsymbol{f}$ is a function of the labels Y and S, i.e., the model takes as input Y and S and seeks to remove the information corresponding to S, including that present in Y. In practice, to improve the stability of the optimization and facilitate learning, in addition to Y and S, we also use X.

5. Scheme for Evaluating Representations

The primary utility of the trade-offs is in illuminating the fundamental limits of learning algorithms in mitigating unfairness and evaluating the effectiveness of learned representations w.r.t. utility and fairness. This includes representations that provide a single solution in the utility-fairness plane and multiple solutions that span the trade-off between utility and fairness. Fairness evaluations in prior literature focused primarily on relative comparisons of the models to each other. Such a comparison, however, precludes an understanding of how far the solution is from the inherent limits of the task. Elucidating and numerically quantifying the inherent trade-offs facilitates such an understanding and drives further algorithmic development.

A standard way to evaluate bi-objective solutions is to plot them on a 2-D plane, identify non-dominated solutions,

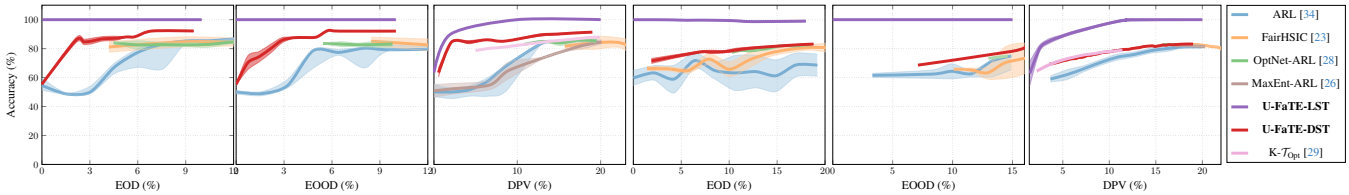


Figure 3. **Evaluating Fair Representation Learning Methods:** Accuracy versus fairness trade-offs on CelebA (a)-(c) and FolkTable (d)-(f). (a) and (d) show the trade-off for Equalized Opportunity as the fairness constraint. (b) and (e) show the trade-off for Equality of Odds as the fairness constraint, and (c) and (f) show the trade-off for Demographic Parity as the fairness constraint. The solid lines represent the mean accuracy at a given fairness value, and the shaded region shows the uncertainty of the trade-off. Both DST and LST estimates from U-FaTE are stable. Among the FRL methods, K-$\mathcal{T}_{\text{Opt}}$ is closest to the DST, while ARL has the most variance.

or compare their dominance w.r.t. each other. More details can be found in the supplementary material.

6. Experimental Evaluation

We designed experiments to answer the following:

1. How far are existing supervised fair representation learning methods from the two trade-offs? (§6.2)
2. How far are zero-shot representations from the two trade-offs? What is the effect of network architecture and pre-training dataset? (§6.3)
3. How far are pre-trained image representations trained in a supervised fashion from the two trade-offs? (§6.4)

6.1. Experimental Setup

Datasets: We estimate the trade-offs through U-FaTE on an assortment of datasets. 1) **CelebA [19]** consists of more than 200K face images of celebrities in the wild annotated with 40 binary attributes. 2) **FairFace [16]** consists of face images from 7 different race groups labeled with race, sex, and age groups. 3) **FolkTables [8]** is a tabular dataset of individuals from fifty states derived from the US Census.

For experiments on CelebA, the target attribute is *high cheekbones*, and the sensitive attribute is a combination of *sex* and *age* for a total of four classes (young woman, young man, old woman, and old man). For experiments on the FairFace dataset, *sex* (binary) is the target attribute with *race* (7 groups: East Asian, White, Black, Indian, Latino, South Asian, and Middle Eastern) being the sensitive attribute. For the FolkTables dataset, we consider data from Washington State and choose *employment status* (binary) and *age* (discrete value between 1 and 96) of the individuals as the target and sensitive attributes, respectively.

Metrics: In all experiments, the utility is measured via classification accuracy. We consider three group fairness metrics (EOD, EOOD, and DPV). The zero-shot CLIP models are evaluated via cosine similarity. Moreover, on the FairFace dataset, we compared the models to the ideal solutions estimated by LST and DST using a weighted normalized Euclidean distance (see supplementary for details). We refer to the distance as Dist_{LST} and Dist_{DST}, respectively.

Implementation Details: We use ResNet-18 as the feature extractor for U-FaTE and the FRL methods. The final classifier is a two-layer MLP. We evaluate the representations from pre-trained vision models by learning a logistic classifier. To obtain the trade-offs, we run U-FaTE and other FRL methods for multiple values of λ between zero and one, where zero corresponds to no fairness constraint, and one corresponds to only fairness.

Optimizing U-FaTE: The trade-offs are defined through the dependence terms in (3) which involves an expectation over the joint distribution $p(X, Y, S)$. However, due to practical considerations, we only have access to a finite set of samples to estimate the trade-offs. Therefore, we estimate the trade-offs using all the samples available in each dataset without splitting it into train, validation, and test sets. This choice ensures that the estimates account for any possible generalization gap between the train and test distributions and identify the best achievable utility-fairness trade-offs.

6.2. Evaluating FRL Methods

FRL Baselines: We consider a wide range of FRL methods based on adversarial learning (ARL [34] and MaxEnt-ARL [26]), dependence measures (FairHSIC [23], OptNet-ARL [28]), and closed-form solvers (K-$\mathcal{T}_{\text{Opt}}$ [29]).

Results: We estimate the LST and DST through U-FaTE and the trade-offs from the other baselines across various settings and datasets. Figs. 3a to 3c show the trade-offs on the CelebA dataset for EOD, EOOD, and DPV, respectively. Similarly, Figs. 3d to 3f show the trade-offs on the Folktable dataset for EOD, EOOD, and DPV, respectively. In the plots, the solid lines represent the mean, and the light shadows represent the variance of the accuracy for a given fairness value. On FairFace, observe that trade-offs do not exist since, on this task, it is possible to mitigate unfairness without sacrificing accuracy. Hence, we present the results of FRL methods and the estimated LST and DST in Tab. 1.

	Method	Accuracy (↑)	Unfairness (↓)	Dist_{DST} (↓)	Dist_{LST} (↓)
EOD	ARL [34]	93.39	1.34	0.448	0.559
	FairHSIC [23]	91.02	1.33	0.445	0.557
	OptNet-ARL [28]	92.94	1.70	0.598	0.709
	U-FaTE-DST	96.17	0.263	-	0.133
	U-FaTE-LST	100.0	0.0	-	-
EOOD	ARL [34]	91.60	3.04	0.447	0.71
	FairHSIC [23]	93.43	2.13	0.236	0.498
	OptNet-ARL [28]	93.39	2.34	0.284	0.546
	U-FaTE-DST	97.93	1.126	-	0.262
	U-FaTE-LST	100.0	0.0	-	-
DPV	ARL [34]	92.49	6.09	0.350	0.351
	FairHSIC [23]	91.41	5.91	0.329	0.332
	OptNet-ARL [28]	93.33	5.80	0.316	0.317
	U-FaTE-DST	94.39	3.082	-	0.04
	U-FaTE-LST	100.0	3.10	-	-

Table 1. Evaluation of FRL methods on FairFace based on the distance to DST and LST estimated by U-FaTE. Color corresponds to the DST and LST trade-offs.

Observations: Although K-$\mathcal{T}_{\text{Opt}}$, OptNet-ARL and FairHSIC can achieve near-optimal accuracy in most cases, they are unable to span the whole range of fairness values. ARL is the most unstable but can span the whole range of fairness values. K-$\mathcal{T}_{\text{Opt}}$ is the most stable method due to its closed-form solver, but is unable to span the whole range of fairness values (Figs. 3c and 3f).

The gap between LST and DST demonstrates the information gap in X for predicting Y. From Fig. 3, we observe that the gap is $\sim$20% of accuracy in low fairness regions and $\sim$40% in high fairness regions for EOD and EOOD. For DPV, the trend reverses with a gap of $\sim$20% for low fairness and gradually decreases with increasing unfairness.

Observe that the LST in Figs. 3a and 3b and Figs. 3d and 3e for EOD and EOOD is almost flat at 100% accuracy. This observation, however, is unsurprising since EOD and EOOD both condition on the label Y, and thus an ideal classifier with 100% accuracy (i.e., $\hat{Y} = Y$) will have zero EOD and EOOD. And, in LST, the Oracle classifier is 100% accurate since it has access to Y and S. So, the LST has sufficient information to minimize EOD and EOOD without sacrificing utility. The same, however, does not hold for DPV since it does not consider the target labels in its definition. Based on the above discussion, we deduce that EOD and EOOD are more pragmatic fairness metrics than DPV since they do not force the model to sacrifice predictive accuracy to ensure fairness. Thus, both offer a more balanced and practical approach to measuring fairness. Our empirical results provides independent confirmation of the same observations in [6, 14].

The comparison of FRL methods in Tab. 1 based on Dist_{LST} suggests that when models are optimized for EOD, ARL and FairHSIC find solutions closer to LST and DST than OptNet-ARL. When models are optimized for EOOD, FairHSIC finds the closest point to the LST and DST. OptNet-ARL performs slightly better than the other FRL methods when optimized for reducing DPV.

6.3. Evaluating Zero-Shot CLIP Models

Zero-Shot Models: To study the fairness of current zero-shot models, we consider more than 100 pre-trained models from OpenCLIP [15] and evaluate them on CelebA and FairFace for the same target and sensitive labels as before.

Results: Figs. 4a and 4b show results on CelebA and FairFace, respectively, for three group fairness definitions. Each point represents the result of one zero-shot CLIP model, with the color denoting the model's pre-training dataset. Plots also include DST and LST for comparison.

Observations: From Fig. 4a, we observe that zero-shot models perform poorly, in terms of accuracy, on the CelebA task. We hypothesize that it is so since the target task (predicting *high cheekbones*) is uncommon, even in the large datasets the models have been trained on. From a fairness perspective, we observe that models trained on CommonPool [9] (red dots) are more likely to be fair, while models trained on DataComp [9] have marginally better accuracy over the other models. Finally, the CLIP models are very far from the DST across the board.

On FairFace (Fig. 4b), since the target task (sex) is abundantly represented in all large-scale pre-training datasets, we observe that the CLIP models exhibit high levels of accuracy. Two CLIP models pre-trained on OpenAI WIT [24] are the closest to DST and LST w.r.t. EOD. Similar to our observations on CelebA, models pre-trained on the CommonPool dataset are more likely to be fair but at the cost of accuracy. In contrast, models pre-trained on DataComp are more unfair but have greater accuracy. Finally, there is a clear positive correlation between accuracy and unfairness across all fairness metrics on FairFace (Fig. 4b).

6.4. Evaluating Supervised Representations

Supervised Baselines: To study the fairness of image representations from models pre-trained in a supervised fashion, we consider more than 900 models from Pytorch Image Models [33] and evaluate them on CelebA and FairFace for the same target and sensitive labels as before.

Results: Figs. 4c and 4d show results on CelebA and FairFace datasets, respectively. Each point represents the result of one supervised model, with color denoting the model's pre-training dataset. Plots also include DST and LST for comparison, but some plots were magnified for better resolution. So, the LST may only be partially visible.

Observations: From the CelebA results in Fig. 4c, we observe a high positive correlation between the accuracy and EOD and EOOD. Thus, models with better accuracy are also more fair. And, in contrast to the zero-shot models, even though *high cheekbones* is a rare label, the representations have sufficient information to accurately detect it with a logistic regression classifier. Specifically, models pre-

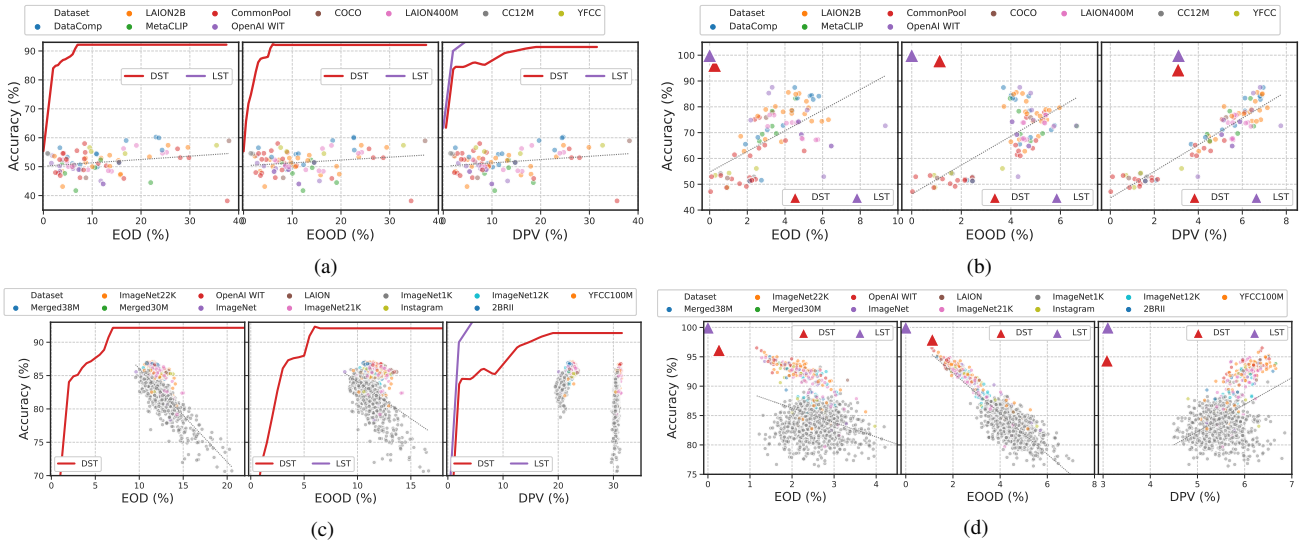


Figure 4. **Evalauting Pre-Trained Zero-Shot and Supervised Models:** Accuracy-fairness evaluation of more than 100 pre-trained zero-shot models on CelebA (a) and FairFace (b), and over 900 pre-trained image representations on CelebA (c), and FairFace (d).

trained on ImageNet22K [25] (orange dots), and OpenAI WIT [24] have the best accuracy and reach the DST trade-off in high unfairness regions. However, with more than 11% EOD and EOOD and more than 20% DPV, they have significant levels of bias between the two sexes.

Results on FairFace in Fig. 4d also reiterate that models trained on OpenAI WIT and ImageNet22K are more fair and more accurate than other datasets. We also observe that the models are generally more fair on FairFace than on CelebA. A similar positive correlation exists between accuracy and fairness for EOD and EOOD. We also make an interesting observation from Fig. 4d. Some models surpass the DST and enter the *"Possible with Extra Data"* region of the utility-fairness plane illustrated in Fig. 1a. Recall that DST estimates the upper bound of the *possible* region for models trained with the same data as DST. The DST can be surpassed to enter the *"Possible with Extra Data"* region if the model is trained on additional data beyond what is used for estimating DST. From Fig. 4d (middle), we observe that a few models trained on OpenAI, LAION [30], ImageNet22K, ImageNet21K, and ImageNet12K can surpass the DST, plausibly since these datasets contain sufficient—both quality and quantity—samples from the distribution of the target attribute (*sex*).

7. Concluding Remarks

As image classification systems are widely deployed in high-stakes applications, ensuring that their predictions do not exhibit demographic bias is paramount for gaining user trust. While it is desirable to mitigate bias without sacrificing accuracy, this is not always possible. This paper studied such inherent trade-offs between utility and fairness. First, we identified two types of trade-offs called *Data-Space* and *Label-Space* trade-offs corresponding to those achievable with and without data restrictions. But unlike prior theoretical studies on utility-fairness trade-offs, next, we focused on developing algorithmic tools for quantifying the trade-offs from data and proposed U-FaTE. As an illustration of its practical utility, we estimated the trade-offs on several image classification tasks, facilitating a large-scale evaluation of over 100 zero-shot and 900 supervised pre-trained models. The results revealed that, out of the box, pre-trained models are far from the best achievable limits of accuracy and fairness. Furthermore, we identified that, in some cases, larger datasets can improve accuracy and fairness and surpass the solutions represented by the DST.

U-FaTE was designed as a composition of a neural network with the last layer optimized to global optimality through a closed-form solver for a given feature representation. This design allowed it to estimate the two trade-offs reliably. However, U-FaTE does not provide convergence or optimality guarantees. Therefore, the estimated DST and LST are likely to be suboptimal. Nonetheless, they can serve as a valuable tool for understanding the nature of the problem (e.g., Does there exist a trade-off? or What are the *possible*, *impossible* and *possible with extra data* regions in the utility-fairness plane?) and how far a given fair learning algorithm is from the achievable limits. Furthermore, depending on which region of the utility-fairness plane a solution is and how far it is from the DST and LST reveals whether to focus on better optimization or better data.

Acknowledgements: This work was supported by the National Science Foundation (award #2147116).

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M Pohl. Representation learning with statistical independence to mitigate bias. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 2

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023. 1, 2, 3

[3] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*, 2023. 2

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 4

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 2018. 2

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. big data 5, 2 (2017), 153–163. *arXiv preprint arXiv:1610.07524*, 2017. 7

[7] Sepehr Dehdashtian, Lan Wang, and Vishnu Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. In *International Conference on Learning Representations*, 2024. 2

[8] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 2021. 6

[9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 7

[10] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[11] Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to fairness in statistical learning. *arXiv preprint arXiv:2005.11720*, 2020. 1

[12] Vincent Grari, Oualid El Hajouji, Sylvain Lamprier, and Marcin Detyniecki. Learning unbiased representations via Rényi minimization. *arXiv preprint arXiv:2009.03183*, 2020. 2

[13] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(12):2075–2129, 2005. 4

[14] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016. 3, 7

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 7

[16] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 6

[17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 2017. 3

[18] Effrosini Kokiopoulou, Jie Chen, and Yousef Saad. Trace optimization and eigenproblems in dimension reduction methods. *Numerical Linear Algebra with Applications*, 18(3): 565–602, 2011. 5

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015. 2, 6

[20] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018. 2

[21] Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Costs and benefits of fair representation learning. *AAAI/ACM Conference on AI, Ethics, and Society*, 2019. 2

[22] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. *Conference on Fairness, Accountability and Transparency*, 2018. 1

[23] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4, 6, 7

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 7, 8

[25] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 8

[26] Proteek Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4, 6

[27] Bashir Sadeghi, Runyi Yu, and Vishnu Boddeti. On the global optima of kernelized adversarial representation learning. *IEEE International Conference on Computer Vision*, 2019. 2

[28] Bashir Sadeghi, Lan Wang, and Vishnu Naresh Boddeti. Adversarial representation learning with closed-form solvers. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021. 2, 4, 6, 7

[29] Bashir Sadeghi, Sepehr Dehdashtian, and Vishnu Boddeti. On characterizing the trade-off in invariant representation

learning. *Transactions on Machine Learning Research*, 2022. Featured Certification. 1, 2, 3, 4, 6

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 8

[31] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2

[32] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[33] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 7

[34] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. *Advances in Neural Information Processing Systems*, 2017. 2, 4, 6, 7

[35] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. *International Conference on Machine Learning*, 2013. 2

[36] Han Zhao. Costs and benefits of wasserstein fair regression. *arXiv preprint arXiv:2106.08812*, 2021. 1

[37] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *arXiv preprint arXiv:1906.08386*, 2019. 1

[38] Han Zhao, Jianfeng Chi, Yuan Tian, and Geoffrey J Gordon. Trade-offs and guarantees of adversarial representation learning for information obfuscation. *arXiv preprint arXiv:1906.07902*, 2019. 2