On Adversarial Robustness of Demographic Fairness in Face Attribute Recognition

Huimin Zeng , Zhenrui Yue , Lanyu Shang , Yang Zhang , Dong Wang*
Unversity of Illinois at Urbana-Champaign
{huiminz3, zhenrui3, lshang3, yzhangnd, dwang24}@illinois.edu

Abstract

Demographic fairness has become a critical objective when developing modern visual models for identity-sensitive applications, such as face attribute recognition (FAR). While great efforts have been made to improve the fairness of the models, the investigation on the adversarial robustness of the fairness (e.g., whether the fairness of the models could still be maintained under potential malicious fairness attacks) is largely ignored. Therefore, this paper explores the *adversarial robustness of demographic fairness* in FAR applications from both attacking and defending perspectives. In particular, we firstly present a novel fairness attack, who aims at corrupting the demographic fairness of face attribute classifiers. Next, to mitigate the effect of the fairness attack, we design an efficient defense algorithm called robust-fair training. With this defense, face attribute classifiers learn how to combat the bias introduced by the fairness attack. As such, the face attribute classifiers are not only trained to be fair, but the fairness is also robust. Our extensive experimental results show the effectiveness of both our proposed attack and defense methods across various model architectures and FAR applications. We believe our work could be strong baselines for future work on robust-fair AI models.

1 Introduction

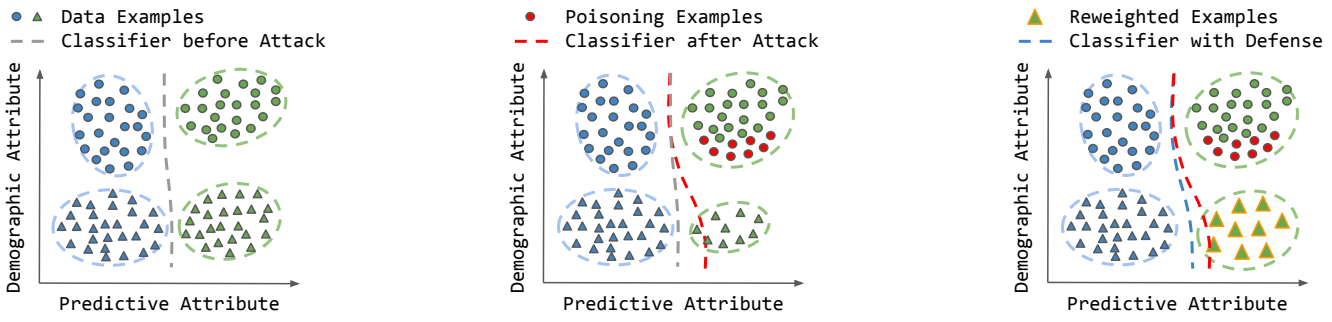
In recent years, deep neural networks (DNNs) have been thriving in face attribute recognition (FAR) applications (e.g., recidivism prediction and justice systems) [Ding *et al.*, 2018; Kärkkäinen and Joo, 2019; Tariq *et al.*, 2022; Tariq *et al.*, 2022]. Despite the impressive performance of DNNs on these applications, they usually suffer from performance bias against *vulnerable* demographic groups (e.g., under-represented races or gender) [Bellamy *et al.*, 2018; Buolamwini and Gebru, 2018]. For instance, [Bellamy *et al.*, 2018; Buolamwini and Gebru, 2018] found that AI facial services (e.g., IBM Watson Visual Recognition [1]) often perform much better for light-skin or male images than dark-skin or female images, resulting in race/gender inequality. As such, it is vital to train models that do not discriminate for FAR services and other identity-sensitive applications.

To address the bias issue, [Zhao *et al.*, 2017; Kou *et al.*, 2021; Zeng *et al.*, 2023; Zeng *et al.*, 2022] focus on creating balanced training distributions, and [Torralba and Efros, 2011; Geirhos *et al.*, 2020; Scimeca *et al.*, 2021] strive to suppress the shortcut learning phenomenon. However, merely pursuing a high degree of model fairness cannot automatically provide the *robustness* of the fairness: could the model always maintain its fairness under potential malicious fairness attacks? For instance, one can develop a fair model for the FAR-based crime prediction system. However, there might exist a malicious attacker, who can easily corrupt the system and make it biased against a specific demographic group (e.g., African-Americans). Under the fairness attack, the system is prone to produce wrong predictions for the attacked demographic group. Therefore, it is not sufficient to only develop fair models, but the robustness of the fairness should also be improved against potential adversarial threats. Compared to [Torralba and Efros, 2011; Geirhos *et al.*, 2020; Scimeca *et al.*, 2021; Zhao *et al.*, 2017; Kou *et al.*, 2021], this work studies the fairness from an adversary perspective.

With the goal of exploring robustness of fairness for FAR models, we firstly present a novel fairness attack, whose attack objective is the fairness of the model instead of its overall performance. Under our proposed attack, we show that the fairness of face attribute classifiers in different FAR applications could be easily corrupted. Then, the observed vulnerability of fairness motivates us to design a defense mechanism to enable the robustness of fairness against the potential attack. In particular, regarding our fairness attack, it is formulated as *test-data-free*, *cross-attribute*, and *clean-label* data poisoning at *group level*. *Test-data-free*: the test-data-free design guarantees the attacker could generate poisoned face data without accessing the test data. Compared to traditional poisoning attacks [Shafahi *et al.*, 2018; Huang *et al.*, 2020; Chen *et al.*, 2017], where selected test samples must be accessible to launch attack, our test-data-free poisoning is not limited by the accessibility of test data, and thus more dangerous. *Cross-attribute*: the cross-attribute feature also makes the attack more insidious: the poisons are generated w.r.t. de-

*Contact Author

[1] https://www.ibm.com/fi-en/cloud/watson-visual-recognition



(a) Train a fair classifier. (b) Launch fairness attack. (c) Defense: robust-fair training.

Figure 1: Overview: (a) Assume we have a trained fair face attribute classifier. (b) Now, an attacker launches the attack against the vulnerable group (green triangles). The attacker generates cross-attribute poisons (red circles) to skew the training distribution, such that the classifier becomes biased against the attacked group. (c) To defend against the fairness attack, we propose a robust-fair training, where a re-weighted training loss is optimized to achieve robust fairness of the classifier.

mographic attributes, which are orthogonal to the predictive attribute of the FAR task. As such, the attacked demographic attribute could be one of all possible demographic attributes, making the attack highly unpredictable. *Clean-label*: under the clean-label constraint, the labels of all poisons remain clean as they appear to human eyes. Along with other training samples, the poisoned dataset remains visually benign but is already biased against a targeted demographic group. The clean-label design makes the attack less detectable and could easily surpass human inception [Shafahi *et al.*, 2018]. *Group-level:* our attack focuses on an entire targeted demographic group instead of a few targeted test samples. The group-level design liberates the power of the attack, because it does not assume specific victim test samples to attack and generalizes to all test samples from the attacked demographic group by design. These four key features fundamentally distinguish our novel attack from existing poisoning attack (e.g., backdoor poisoning or test-data-required fairness poisoning). The intuition of this attack is illustrated in Figure 1. Next, motivated by the adversarial threat on model fairness, we further design an efficient defense mechanism called robust-fair training to enable the robustness of fairness against the fairness attack. Under our defense scheme, fairness adversarial examples will be generated, such that the bias within training data could be maximized. Under the biased training distribution, the face attribute classifiers are then trained to be fair using a fairness-aware loss. To this end, the face attribute classifiers learn to combat the bias introduced by the fairness adversarial examples. Eventually, the fairness of the classifier becomes robust against the fairness poisoning attack.

Consider a recidivism prediction system based on face attribute recognition. Assume the system is developed to be fair using existing methods, e.g., fair sampling. Now, a malicious fairness attacker attacks the system to be biased against a target demographic population (e.g. Asian or African-American). Since the attacker is *test-data-free*, it needs not to know which exact person to attack, but the attack effect is automatically generalized to all people from the attacked demographic group by wrongly predicting them as criminals. Moreover, the *clean-label* attacker does not flip the ground-truth labels of the poisoned data. As such, the attack could easily surpass the data checking process, as the poisoned data is visually benign and its label is clean. In addition, the cross-attribute design of the attack manipulates the demographic attribute of the poisoned data (e.g., races) instead of the predictive attribute (being criminal or not) to bias the training model. As a consequence, the attacked population would be misclassified as criminals by the biased system, leading to injustice in law enforcement.

To the best of our knowledge, our work is the first work investigating the adversarial robustness of demographic fairness in FAR applications under a novel cross-attribute, test-data-free, clean-label and group-level poisoning attack. We select FAR as our study case because FAR is highly identity-sensitive and the demographic information of FAR applications is usually implicitly integrated into face data by default. As such, launching fairness attack in FAR applications becomes more challenging compared to other applications, where vectorized data format (e.g., COMPAS [Angwin *et al.*, 2016] or Income Census [Kohavi, 1996]) is used. Under a vectorized data format, data attributes are stored in each entry of the data vector (e.g., one-hot representing the gender/race). When launching attack on vectorized data, one can easily flip the specific entry of the vector to modify the demographic information of that data point. In contrast, when the attacker tries to modify the demographic information of face images, it is non-trivial for the attacker to guarantee the efficiency of the manipulation while keeping the manipulation imperceptible.

After launching the proposed attack on different FAR tasks, we observe that the fairness property of different face attribute classifiers could be easily corrupted under our attack. Moreover, compared to existing fairness attacks (e.g., adversarial sampling, adversarial labeling, hard adversarial examples, and Fairness Poisoning (FP) [Chang *et al.*, 2020]), our attack is more dangerous and efficient because of the novel design of our attack. Our superior attack results indicate that our method could be used as a stronger baseline for future work. Besides, we also demonstrate that our defense mechanism shows impressive performance in terms of increasing the robustness of model fairness across various FAR applications. Finally, we highlight that while our attack scheme is

designed for general multi-class classification problems, we select binary attributes in our experiments for a fair comparison against baseline attack methods as in [Chang *et al.*, 2020].

We summarize the contributions of this work as follows:

- We introduce a new type of data poisoning attack against demographic fairness of FAR models. The key novel features (i.e., cross-attribute, test-data-free, clean-label and group-level) of our proposed fairness attack distinguish our work fundamentally from other existing poisoning attacks (e.g., backdoor attacks, single instance poisoning and test-data-required fairness poisoning).
- We present two concrete methods to launch the attack. With systematic evaluation on various FAR applications, we show that fair face attribute classifiers are severely biased by the attacker using both methods, and our attack is significantly more efficient than other baselines.
- We design an efficient defense mechanism called robust-fair training to improve the robustness of the fairness against the potential fairness attack. In addition to the successful attacks, our experimental results also demonstrate the efficacy of the proposed defense method.

2 Related Work

Fair Machine Learning and Fairness Attacks. Efforts have been made to address the bias issue in AI models [Li *et al.*, 2021; Wen *et al.*, 2022]. In these studies, it is usually assumed that the data source is clean, leading to desired model fairness. However, this assumption leaves a fatal loophole: the fairness derived using existing debiasing methods could be easily broken, when the data source is hacked [Chang *et al.*, 2020; Van *et al.*, 2021; Mehrabi *et al.*, 2020]. [Mehrabi *et al.*, 2020] leverages test data to craft poisons to corrupt model fairness. Similarly, online fairness attacks are proposed in [Chang *et al.*, 2020] and [Van *et al.*, 2021], where poisons are continuously generated and injected into the training process. However, in this work, we consider a more practical and dangerous setting, where the attack model has no access to test data nor the labeling function of poisons. Moreover, the online fairness attacks [Chang *et al.*, 2020; Van *et al.*, 2021] can only 'select' training samples as poisons based on a fairness metric. Such attack is not targeted, since the attacker could not specify a demographic group to attack. Moreover, we observed that online fairness attacks on large models are extremely slow due to selection process, and does not necessarily generate a visually fair poisoned dataset, which make the attack less effective but more detectable. In contrast, our proposed cross-attribute attack directly manipulates the demographic information of poisoned data as the attacker desires, which drastically increases the targeted attack effect. We require the manipulation of poisoned data to be imperceptible, which guarantees a fair-look of the poisoned dataset and makes the attack less detectable.

Data Poisoning. To launch traditional poisoning attacks, the prerequisite of generating poisoned samples is to use the targeted test instances [Chen *et al.*, 2017; Shafahi *et al.*, 2018; Huang *et al.*, 2020; Geiping *et al.*, 2020]. That is, the attacker has to access targeted test samples at first place, then apply specific poisoning strategies, such as placing perturbed fake images geometrically [Shafahi *et al.*, 2018; Huang *et al.*, 2020] close to the clean test images or plugging in backdoor trigger in test samples [Chen *et al.*, 2017]. Moreover, traditional targeted attacks focus on attacking targeted samples individually: after training the model on the poisoned data, the decision boundary near the targeted test sample is *locally* distorted (sample-level) to cause a wrong prediction [Chen *et al.*, 2017; Shafahi *et al.*, 2018; Huang *et al.*, 2020]. Meanwhile, clean-label and test-data-free unlearning poisoning was studied in [Huang *et al.*, 2021; Fowl *et al.*, 2021]. In contrast, our attack is **group-level** and **test-data-free** : we propose to inject test-data-irrelevant perturbed samples to the training dataset to deform the decision boundary statistically. As such, the attacker needs no access to test data and its attack effect automatically generalizes the entire targeted demographic group rather than just a selected test sample. The test-data-free and group-level design fundamentally distinguish our attack from existing targeted poisoning attacks (including backdoor attacks).

3 Problem Statement

Notations. In an FAR application, a face dataset is denoted by $\boldsymbol{D}_N = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\} \in \boldsymbol{\mathcal{X}} \times \mathcal{Y}$, where N is the total number of images, $\boldsymbol{\mathcal{X}}$ refers to the input image space and $\mathcal{Y}$ represents the face attribute space. Technically, a dataset could be divided into a training set $\boldsymbol{X}_{train}$ and a test set $\boldsymbol{X}_{test}$. Our goal is to train a face attribute classifier $f_{\boldsymbol{\theta}}$ who predicts a face attribute y for a given face image $\boldsymbol{x}$. However, face images are usually characterized by a set of sensitive demographic attributes such as gender or race. Without loss of generality, we consider C demographic attributes for all faces and denote them as an attribute list $\mathcal{A} = \{a^{(1)}, ..., a^{(C)}\}$. Given a face image $\boldsymbol{x}^{(i)}$, its demographic labels are defined as $\mathcal{A}(\boldsymbol{x}^{(i)}) = \{e^{(i,1)}, ..., e^{(i,C)}\}$. For instance, we consider one specific demographic attribute for each FAR application, where we denote the demographic label as $e^{(c)}$. Finally, to properly measure the bias of the models, we adopt the commonly used demographic parity ϕ_D and equalized odds ϕ_E [Hardt *et al.*, 2016].

Threat Model. We assume the worst-case threat model, where the attacker can access all information of the attacked face attribute classifier (e.g., architecture, parameters) and its training specification (e.g., learning rate, batch size). Moreover, we aim at designing a more dangerous attacker that generates poisons even without any knowledge about the test data $\boldsymbol{X}_{test}$ (no access to test data). Also, the attacker is cross-attribute: perturbing demographic information w.r.t. $a^{(c)} \in \mathcal{A}$ to corrupt face attribute predictions w.r.t. $y \in \mathcal{Y}$. Finally, our attack is clean-label: the labels of the poisoned data remain correct and clean as it appear to human eyes.

4 Fairness Attack and Defense

4.1 Crafting Cross-attribute Poisons

How can the attacker bias the classifier in terms of a demographic attribute (e.g., gender) when the model is originally designed to predict an orthogonal face attribute (e.g., attractiveness)? Consider an attacker aiming to bias the model

against smiling males, and increase gender inequality in the FAR application. In this case, we define the smiling male group as the attacked demographic group ($\boldsymbol{X}_{att}$). To bias the model, the attacker pulls the decision boundary towards the attacked demographic group by perturbing the images from the attacked demographic group away from their original data domain towards another data domain of a different demographic group (defined as the reference demographic group $\boldsymbol{X}_{ref}$). To generate cross-attribute poisons, we present two different methods.

Domain-level perturbation. The first method is domain-level: the attacker treats each demographic group as a separate data domain and will partially move the data domain of the attacked demographic group away from its original location. In other words, the attacker tries to perturb the domain of $\boldsymbol{X}_{att}$ towards $\boldsymbol{X}_{ref}$ by reducing the distance between these two domains. To measure the domain distance, we use the maximum mean discrepancy (MMD) distance. The MMD estimates the domain distance between two data domains using samples drawn from them [Gretton *et al.*, 2012]. Given two data domains $\mathcal{P}$ and $\mathcal{Q}$, MMD $\mathcal{D}_{\text{MMD}}$ is defined as: $\mathcal{D}_{\text{MMD}} = \sup_{k\in\mathcal{H}}\left(\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}}[k(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{x}\sim\mathcal{Q}}[k(\boldsymbol{x})]\right)$, where k is a function (kernel) in the reproducing kernel Hilbert space $\mathcal{H}$. In our implementation, the expectation is simplified by using the latent representations of the drawn samples from two different demographic groups as in [Yue *et al.*, 2021]. k is realized using Gaussian kernel [Yue *et al.*, 2021], i.e., $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\gamma})$. As such, the MMD between $\boldsymbol{X}_{att}$ and $\boldsymbol{x}_{ref}$ could be simplified as:

$$\begin{aligned}\mathcal{D}^{\text{MMD}} = &\frac{1}{|\boldsymbol{X}_{att}||\boldsymbol{X}_{att}|}\sum_{i=1}^{|\boldsymbol{X}_{att}|}\sum_{j=1}^{|\boldsymbol{X}_{att}|} k(f_e(\boldsymbol{x}_{att}^{(i)}), f_e(\boldsymbol{x}_{att}^{(j)}))\\ &+\frac{1}{|\boldsymbol{X}_{ref}||\boldsymbol{X}_{ref}|}\sum_{i=1}^{|\boldsymbol{X}_{ref}|}\sum_{j=1}^{|\boldsymbol{X}_{ref}|} k(f_e(\boldsymbol{x}_{ref}^{(i)}), f_e(\boldsymbol{x}_{ref}^{(j)}))\\ &-\frac{2}{|\boldsymbol{X}_{att}||\boldsymbol{X}_{ref}|}\sum_{i=1}^{|\boldsymbol{X}_{att}|}\sum_{j=1}^{|\boldsymbol{X}_{ref}|} k(f_e(\boldsymbol{x}_{att}^{(i)}), f_e(\boldsymbol{x}_{ref}^{(j)})),\end{aligned} \tag{1}$$

where f_e is the feature extractor of the classifier $f_{\boldsymbol{\theta}}$. $\boldsymbol{x}_{att}$ and $\boldsymbol{x}_{ref}$ are face images sampled from the attacked demographic group $\boldsymbol{X}_{att}$ and the reference demographic group $\boldsymbol{X}_{ref}$. To partially perturb the data domain of $\boldsymbol{X}_{att}$ towards the data domain of $\boldsymbol{X}_{ref}$, the attacker perturb $\boldsymbol{x}_{att} \sim \boldsymbol{X}_{att}$ with $\boldsymbol{\delta}$ by reducing the MMD distance. Since $\boldsymbol{\delta}$ is independent from $\boldsymbol{x}_{ref}$, $\mathcal{D}^{\text{MMD}}$ is thus reduced to:

$$\begin{aligned}&\mathcal{D}^{\text{MMD}}(\boldsymbol{\delta}) =\\ &\frac{1}{|\boldsymbol{X}_{att}||\boldsymbol{X}_{att}|}\sum_{i=1}^{|\boldsymbol{X}_{att}|}\sum_{j=1}^{|\boldsymbol{X}_{att}|} k(f_e(\boldsymbol{x}_{att}^{(i)}+\boldsymbol{\delta}^{(i)}), f_e(\boldsymbol{x}_{att}^{(j)}+\boldsymbol{\delta}^{(j)}))\\ &-\frac{2}{|\boldsymbol{X}_{att}||\boldsymbol{X}_{ref}|}\sum_{i=1}^{|\boldsymbol{X}_{att}|}\sum_{j=1}^{|\boldsymbol{X}_{ref}|} k(f_e(\boldsymbol{x}_{att}^{(i)}+\boldsymbol{\delta}^{(i)}), f_e(\boldsymbol{x}_{ref}^{(j)})).\end{aligned} \tag{2}$$

Instance-level perturbation. In addition to the domain-level perturbations, we present another instance-level perturbation by increasing the specificity of the attacker: the attacker solves the optimal perturbation individually for each face image from the attacked demographic group. To generate instance-level perturbations, an additional neural network is trained to extract the demographic information of face images. In this way, the attacker's manipulation on demographic information could be more efficient. For a better understanding of the attack model, we propose to split a neural network into two sub-networks, a feature-extracting encoder f_e (which is already used in Equation 1), and a feature classifier f_y or f_a (for face attribute y or any other demographic attribute $a^{(c)}$), as shown in Figure 2. To train the demographic attribute classifier f_a, we initialize a classification module f_a for the original face attribute classifier $f_{\boldsymbol{\theta}} = f_y(f_e)$. f_a is the classifier w.r.t. the demographic attribute that the attacker is interested in (i.e., $a^{(c)}$). Similar to train f_y, the parameter of f_{a_c} is optimized using the available training dataset $\boldsymbol{X}_{train}$:

$$\min_{f_a}\mathbb{E}_{(\boldsymbol{x},y)\sim\boldsymbol{X}_{train}}\left[l(f_a(f_e(\boldsymbol{x})), e^{(c)})\right], \tag{3}$$

where $e^{(c)}$ is the demographic label of the face images for demographic attribute $a^{(c)}$.

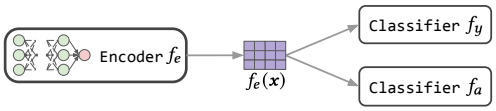


Figure 2: Split a face attribute classifier into an feature extractor f_e and a classifier f_y. To generate instance-level perturbations, we need an extra demographic attribute classifier f_a, which shares the same feature extractor with f_y.

After deriving f_a, the attacker could perturb the face images from the attacked demographic group towards another nearby demographic group. To measure the distance between individual images from different demographic groups, we use KL-divergence:

$$\mathcal{D}^{\text{KL}}(\boldsymbol{\delta}) = KL[f_a(f_e(\boldsymbol{x}_{att}+\boldsymbol{\delta})), f_a(f_e(\boldsymbol{x}_{ref}))], \tag{4}$$

where $\boldsymbol{x}_{att}$ is a face image sampled from the attacked demographic group $\boldsymbol{X}_{att}$, and $\boldsymbol{x}_{ref}$ is a reference face image from reference demographic group $\boldsymbol{X}_{ref}$.

Crafting Cross-attribute Poisons. With both perturbation strategies introduced above, we present the objective for generating cross-attribute poisons. Note that both Equation 2 and Equation 4 only considered the perturbation along the demographic attribute dimension. However, without controlling the perturbation along the face attribute dimension, it is likely that the resulted poisons could distort the decision boundary in an undesired direction and suppress the attack efficiency. For instance, the attacker could wrongly perturb a smiling male image to be a unsmiling female without controlling the perturbation along the face attribute dimension. In this case the poison does not directly contribute to the objective of attacking smiling male group. Therefore, we propose to jointly consider the perturbation constraints from both the face attribute dimension and the demographic attribute dimension:

- Domain-level:

$$\begin{aligned} \min_{\boldsymbol{\delta}} \quad & \mathcal{D}^{\mathrm{MMD}}(\boldsymbol{\delta}) + \alpha \cdot \mathbb{E}\Big[l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{att}+\boldsymbol{\delta}), y))\Big] \\ \text{s.t.} \quad & \|\boldsymbol{\delta}\| < \epsilon, \quad \mathbb{E} := \mathbb{E}_{\boldsymbol{x}_{att}\sim\boldsymbol{X}_{att}, \boldsymbol{x}_{ref}\sim\boldsymbol{X}_{ref}} \end{aligned} \tag{5}$$

- Instance-level:

$$\begin{aligned} \min_{\boldsymbol{\delta}} \quad & \mathcal{D}^{\mathrm{KL}}(\boldsymbol{\delta}) + \alpha \cdot l(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{att}+\boldsymbol{\delta}), y) \\ \text{s.t.} \quad & \|\boldsymbol{\delta}\| < \epsilon, \quad \boldsymbol{x}_{att} \sim \boldsymbol{X}_{att}, \quad \boldsymbol{x}_{ref} \sim \boldsymbol{X}_{ref} \end{aligned} \tag{6}$$

Note that a scaling factor α is introduced for both Equation 5 and Equation 6 to control the trade-off between the constraints from the demographic attribute and face attribute. In our experiments, we implement the iterative projected gradient descend [Madry *et al.*, 2017] to approximate the optimal solutions to both equations. We highlight that in both poisoning strategies, the attacker does not require the access to test data to generate poisons, which differentiates our framework from previous studies on data poisoning [Chen *et al.*, 2017; Shafahi *et al.*, 2018; Huang *et al.*, 2020]. Moreover, the proposed attacker is clean-label and does not manipulate the labels of the poisoned data. The poisoned data still has correct labels for both face attribute and the attacked demographic attribute as it appears to eyes. More importantly, if the original face dataset is fair, the poisoned dataset is still *visually* fair.

After solving Equation 5 and Equation 6 for the selected attacked demographic group $\boldsymbol{X}_{att}$ and the reference demographic group $\boldsymbol{X}_{ref}$, the attacker obtains a poisoned dataset $\boldsymbol{Z}$ and launches the attack.

4.2 Poisoning Attack on Fairness

For face attribute classifiers in real-world FAR applications, transfer learning is a commonly used methodology to update the parameters of the ML models [Shafahi *et al.*, 2018]. Therefore, inspired by [Shafahi *et al.*, 2018], we use transfer learning to simulate the attack process:

$$\min f_y \mathbb{E}_{(\boldsymbol{x},y)\sim\boldsymbol{X}'}\Big[l(f_y(f_e(\boldsymbol{x})), y)\Big], \tag{7}$$

where $\boldsymbol{X}' = \boldsymbol{X}_{\boldsymbol{train}} \cup \boldsymbol{Z}$ is the poisoned dataset. Since the attacker is clean-label and does not change the ground truth labels of the poisoned data, the poisoned dataset $\boldsymbol{X}'$ looks fair to human eyes if the original training set is fair. During the poisoning phase, however, $\boldsymbol{X}'$ is severely biased w.r.t. the demographic attribute $a^{(c)}$ from the classifier f_y's perspective.

4.3 Defending against the Fairness Attack

Given the threat above, how to defend against the fairness attack? In this section, we propose a defense named robust-fair training to enable the robustness of fairness for the face attribute classifiers.

To obtain *robust* fairness, the robust-fair training tries to train a fair model under a worst-case adversarial distribution. The adversarial distribution is crafted to be least-fair (or most unfair) w.r.t. a fairness metric (e.g., Φ_D or Φ_E). We formulate this process as a minimax game between a fairness adversary and a fairness booster. The fairness adversary tries to generate fairness adversarial examples by biasing the training samples with perturbations w.r.t. a fairness metric, whereas the fairness booster tries to learn a fair model via a fairness-aware loss even if the training data is deliberately biased. Formally, the minimax game is defined as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \frac{1}{C}\sum_{j=1}^{C}\left[\frac{1}{M_j}\sum_{i=1}^{M_j}\Big[l(f_{\boldsymbol{\theta}}(\boldsymbol{x}^{(j,i)}+\boldsymbol{\eta}^{(j,i)}), y^{(j,i)})\Big]\right] \\ s.t. \qquad \boldsymbol{\eta}^{(j,i)} = \arg\max_{\boldsymbol{\eta}} \hat{\Phi}_*(f_{\boldsymbol{\theta}}, \boldsymbol{x}^{(j,i)}), \\ \|\boldsymbol{\eta}^{(j,i)}\| \le \epsilon, \quad \boldsymbol{x}^{(j,i)} \in \boldsymbol{X}' \end{aligned} \tag{8}$$

where C denotes the number of demographic groups for the FAR application and M_j denotes the number of training samples of the j-th demographic group. $\boldsymbol{\eta}^{(j,i)}$ is the bias perturbation deliberately crafted by the fairness adversary to bias the training sample $\boldsymbol{x}^{(j,i)}$. $\hat{\Phi}_*$ is the differentiable version of any fairness notion (Φ_D or Φ_E) (in Section 3). Note that the 'training samples' $\boldsymbol{x}^{(j,i)}$ in Equation 8 is sampled from the poisoned dataset $\boldsymbol{X}'$. As such, the defense mechanism is attack-agnostic: the robustness of fairness would still be improved no matter whether there exists poisoned data or not. Moreover, the bias perturbation $\boldsymbol{\eta}$ is different from the perturbation $\boldsymbol{\delta}$ defined in Section 4.1 and 4.2. $\boldsymbol{\delta}$ is the perturbation generated during the poisoning phase to bias the model against the attacked demographic group. In contrast, $\boldsymbol{\eta}$ is generated during the defense phase, so that the classifier learns to combat the bias introduced by adversarial examples.

The maximization in Equation 8 specifies the goal of the fairness adversary is to generate the fairness adversarial examples with a specific budget ϵ [2]. The budget ϵ avoids infinity solutions to the maximization problem. To obtain $\boldsymbol{\eta}^{(j,i)}$, the maximization is solved using projected gradient descent (PGD). Regarding the computation of the gradients for the perturbations, we highlight that $\hat{\Phi}_*$ is the differentiable version of Φ_D or Φ_E. When computing Φ_D or Φ_E, the $\arg\max$ operation will be applied to the output logits to produce the predictions, which makes Φ_D or Φ_E non-differentiable. Therefore, in our implementation, we compute soft scores for these metrics by plugging in the normalized output logits instead of the predictions to enable the PGD.

The minimization in Equation 8 implies that the fairness booster increase model fairness by minimizing a fairness-aware training loss. Herein, the fairness-aware training is enabled by re-weighting each training sample inversely proportional to the data frequency (IDF) within each demographic group [Han *et al.*, 2021]. That is, the empirical risk is averaged within each demographic group, then the group-level risk is further averaged over the number of demographic groups. Moreover, note that Equation 8 is evaluated over the perturbed training samples $\boldsymbol{x}^{(j,i)} + \boldsymbol{\eta}^{(j,i)}$. Since the perturbation $\boldsymbol{\eta}^{(j,i)}$ is generated in a way such that the selected fairness metric would be maximized, optimizing the fairness-aware loss over the fairness-adversarial examples is essentially performing robust-fair training. With robust-fair training, the face attribute classifier is trained to be fair under a

[2] Note the same ϵ was used to generate poisons. As such, the power of the fairness attacker and defender is balanced for a fair comparison.

Face Application	Model	LightCNN-9			ResNet-18			VGG-19		
	Baselines	$\Phi_D \uparrow$	$\Phi_E \uparrow$	$\mathcal{A}_B \downarrow$	$\Phi_D \uparrow$	$\Phi_E \uparrow$	$\mathcal{A}_B \downarrow$	$\Phi_D \uparrow$	$\Phi_E \uparrow$	$\mathcal{A}_B \downarrow$
Attractivness (CelebA)	Balanced	0.0092	0.0392	0.7734	0.0164	0.0328	0.7610	0.0284	0.0558	0.7882
	Adv. Sampling	0.0513	0.1026	0.7744	0.0492	0.0984	0.7722	0.0212	0.0488	0.7910
	Adv. Labeling	0.0532	0.1064	0.7689	0.0464	0.0928	0.7692	0.0564	0.1128	0.7766
	Adv. Hard	0.0631	0.1262	0.7801	0.0384	0.0768	0.7748	0.0072	0.0336	0.7884
	Online FP	0.0446	0.0891	0.7742	0.0964	0.1928	0.7258	0.0236	0.0472	0.7778
	Domain (Ours)	**0.5080**	**1.0160**	**0.6804**	**0.5264**	**1.0528**	**0.6484**	**0.2928**	**0.5856**	**0.7312**
	Instance (Ours)	0.1948	0.3896	0.7314	0.2680	0.5360	0.7040	0.1096	0.2192	0.7744
Smiling (CelebA)	Balanced	0.0184	0.0544	0.9200	0.0196	0.0392	0.9198	0.0292	0.0060	0.9234
	Adv. Sampling	0.0253	0.0656	0.9202	0.0232	0.0464	0.9232	0.0292	0.0584	0.9222
	Adv. Labeling	0.0237	0.0666	0.9142	0.0260	0.0520	0.9246	0.0320	0.0640	0.9212
	Adv. Hard	0.0226	0.0654	0.9222	0.0216	0.0464	0.9264	0.0180	0.0392	0.9242
	Online FP	0.0242	0.0694	0.9176	0.0112	0.0368	0.9272	0.0252	0.0632	**0.9146**
	Domain (Ours)	**0.1548**	**0.3096**	**0.8942**	**0.1520**	**0.304**	**0.8948**	**0.0616**	**0.1232**	0.9148
	Instance (Ours)	0.0668	0.1336	0.9142	0.0620	0.1240	0.9150	0.0436	0.0872	0.9218
Age (CelebA)	Balanced	0.0288	0.0576	0.8032	0.0512	0.1024	0.8000	0.0364	0.0728	0.8186
	Adv. Sampling	0.0344	0.0688	0.8000	0.0660	0.1320	0.8030	0.0404	0.0808	0.8166
	Adv. Labeling	0.0564	0.1128	0.7958	0.0716	0.1432	0.7898	0.0408	0.0816	0.8064
	Adv. Hard	0.0524	0.1048	0.8114	0.0828	0.1656	0.7966	0.0376	0.0752	0.8160
	Online FP	0.0496	0.0992	0.8076	0.0332	0.0664	0.7982	0.0336	0.0672	0.8268
	Domain (Ours)	0.3568	0.7136	0.7440	0.4272	0.8533	0.6940	**0.2056**	**0.4112**	0.7976
	Instance (Ours)	**0.3736**	**0.7472**	**0.7436**	**0.4744**	**0.9488**	**0.6856**	0.1096	0.2192	**0.7744**
Age (FairFace)	Balanced	0.0648	0.1296	0.7424	0.0524	0.1048	0.7498	0.0180	0.0360	0.8186
	Adv. Sampling	0.0992	0.1984	0.7536	0.0956	0.1912	0.7606	0.0300	0.0600	0.8162
	Adv. Labeling	0.0548	0.1096	0.7494	0.0632	0.1264	0.7428	0.0256	0.0512	0.8164
	Adv. Hard	0.0812	0.1624	0.7374	0.0708	0.1416	0.7690	0.0268	0.0536	0.8190
	Online FP	0.0684	0.1368	0.7490	0.0720	0.1440	0.7588	0.0184	0.0368	0.8192
	Domain (Ours)	0.3320	0.6640	**0.7232**	0.2892	0.5784	**0.7250**	0.1728	0.3456	0.7764
	Instance (Ours)	**0.3736**	**0.7472**	0.7436	**0.3120**	**0.6240**	0.7332	**0.2292**	**0.4584**	**0.7702**

Table 1: Evaluation results of launching poisoning attacks against fairness, where the attacker wants to *increase* Φ_D and Φ_E, but *decrease* $\mathcal{A}_B$. The best results are highlighted in bold and the second best results are highlighted with underline.

worst-case adversarial scenario where the training data distribution is least-fair. To this end, the classifier learns to combat the bias introduced by the skewed training distribution and becomes robustly fair.

5 Experiments

5.1 Dataset and Experimental Setup

We use the large scale CelebA [Liu *et al.*, 2015] [3] and FairFace dataset [Karkkainen and Joo, 2021] for our experiments. Due to space limit, we chose *attractiveness*, *smiling* and *age* as our predictive face attributes. As mentioned in [Shen *et al.*, 2017], attractiveness prediction is highly challenging due to its subjectivity whereas smiling detection is simpler since smiling or not is easier to judge. Therefore, we pick these two FAR applications of different difficulty to understand the vulnerability of demographic fairness in FAR applications. To test the generality of our method across different datasets, we choose age detection, as only the age label is available in both Fairface and Celeba. Without loss of generality, we choose gender as the demographic attribute. To simulate the attack on *fair* classifiers and demonstrate the bias introduced by various fairness attacks, we sample two fair subsets to train and test the victim models. To maintain the size of the training set and guarantee the 'fair look' of the training set, when inserting the poisons into the training set, we also inject the same number of clean images from the other demographic groups. To measure model performance, we use fairness metrics Φ_D and Φ_E as well as the accuracy metric $\mathcal{A}_B$ (balanced accuracy). For reproducibility, all details (e.g., experimental setup, attacker specification) and code are uploaded within the supplementary materials of this submission.

[3] dataset link: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

5.2 Baselines

Baseline Architecture. Without the loss of generality and due to the space limit, we pick LightCNN [Wu *et al.*, 2018], ResNet [He *et al.*, 2016] and VGGNet [Simonyan and Zisserman, 2014] as the network architectures to perform all FAR applications. Moreover, for our attack, we build the demographic attribute classification module f_a (defined in Section 4) with exactly the same structure as the modified classification module for all baseline architectures. The face attribute classifiers as well as the demographic attribute classifiers for all three FAC applications are trained using our sampled fair training sets, respectively.

Baseline Poisoning Attacks. As discussed in Section 1 and Section 2, it is impractical to implement traditional data poisoning algorithms [Chen *et al.*, 2017; Shafahi *et al.*, 2018;

Face Application	Model	LightCNN			ResNet-18			VGG-9		
	Baselines	$\Phi_D \downarrow$	$\Phi_E \downarrow$	$\mathcal{A}_B \uparrow$	$\Phi_D \downarrow$	$\Phi_E \downarrow$	$\mathcal{A}_B \uparrow$	$\Phi_D \downarrow$	$\Phi_E \downarrow$	$\mathcal{A}_B \uparrow$
Attractivness (CelebA)	Before Attack (Balanced)	0.0092	0.0392	0.7734	0.0164	0.0328	0.7610	0.0284	0.0558	0.7882
	Attack (Domain-level)	0.4509	0.9008	0.6884	0.4396	0.8792	0.6730	0.2320	0.4640	0.7448
	Defense (Domain-level)	**0.1820**	**0.3640**	**0.7594**	**0.2052**	**0.4104**	**0.7546**	**0.0996**	**0.1992**	**0.7814**
	Attack (Instance-level)	0.2716	0.5432	0.7206	0.3364	0.6728	0.6798	0.2120	0.4240	0.7484
	Defense (Instance-level)	**0.1228**	**0.2456**	**0.7522**	**0.2612**	**0.5224**	**0.7250**	**0.0940**	**0.1880**	**0.7650**
Smiling (CelebA)	Before Attack (Balanced)	0.0184	0.0544	0.9200	0.0196	0.0392	0.9198	0.0292	0.0060	0.9234
	Attack (Domain-level)	0.1272	0.2544	0.9044	0.1284	0.2568	0.8890	0.0544	0.1088	0.9172
	Defense (Domain-level)	**0.0624**	**0.1248**	**0.9172**	**0.0628**	**0.1256**	**0.9154**	**0.0148**	**0.0584**	**0.9266**
	Attack (Instance-level)	0.0696	0.1392	0.9136	0.0920	0.1840	0.9088	0.0364	0.0728	0.9210
	Defense (Instance-level)	**0.0664**	**0.1328**	**0.9164**	**0.0856**	**0.1712**	**0.9128**	**0.0148**	**0.0568**	**0.9274**
Age (CelebA)	Before Attack (Balanced)	0.0288	0.0576	0.8032	0.0512	0.1024	0.8000	0.0364	0.0728	0.8186
	Attack (Domain-level)	0.3132	0.6264	0.7498	0.3280	0.6560	0.7344	0.1728	0.3456	0.8044
	Defense (Domain-level)	**0.1284**	**0.2568**	**0.7738**	**0.1608**	**0.3216**	**0.7884**	**0.0900**	**0.1800**	**0.8262**
	Attack (Instance-level)	0.3132	0.6264	0.7610	0.4060	0.8210	0.6994	0.3180	0.6360	0.7366
	Defense (Instance-level)	**0.1024**	**0.2048**	**0.7828**	**0.1472**	**0.2944**	**0.7900**	**0.1208**	**0.2416**	**0.8218**
Age (FairFace)	Before Attack (Balanced)	0.0648	0.1296	0.7424	0.0524	0.1048	0.7498	0.0180	0.0360	0.8186
	Attack (Domain-level)	0.3284	0.6568	0.7246	0.3172	0.6344	0.7218	0.1760	0.3520	0.7916
	Defense (Domain-level)	**0.1428**	**0.2856**	**0.6654**	**0.1408**	**0.2816**	**0.7660**	**0.0664**	**0.1328**	**0.8028**
	Attack (Instance-level)	0.3504	0.7008	0.7152	0.3300	0.6600	0.7170	0.2372	0.4744	0.7594
	Defense (Instance-level)	**0.1580**	**0.3160**	**0.6847**	**0.1544**	**0.3088**	**0.7563**	**0.0796**	**0.1592**	**0.8109**

Table 2: Evaluation results of defending against the fairness poisoning attacks, where the defender aims to *decrease* Φ_D and Φ_E, and *increase* $\mathcal{A}_B$. The bold results suggest that the defender improves the robustness of fairness under each attack case.

Huang *et al.*, 2020] and the test-data-required fairness attack [Mehrabi *et al.*, 2020] for comparison in our setting. Firstly, our attack is at group-level while the standard poisoning is at instance-level. Second, our attack does not use test data information that is used by existing fairness poisoning work [Mehrabi *et al.*, 2020]. To build baseline comparisons, we adopt the baseline fairness poisoning attacks (Adv. sampling, Adv. labeling, Adv. Hard) and the fairness poisoning algorithm (**F**airness **P**oisoning) in [Chang *et al.*, 2020].

5.3 Evaluation Results

Attack Evaluation. The results on attacking fairness of face attribute classifiers are shown in Table 1. In this set of experiments, 5% of the training data is poisoned. We selected 5% because the attack efficacy of all baseline methods would be too marginal when fewer poisons are used. We observe that both our domain-level attack and instance-level attack outperform all baseline methods in terms of corrupting the *fairness* of the face attribute classifiers. For instance, in attractiveness prediction, the equalized odds of LightCNN-9 is raised from 0.0392 to 0.3272 (domain-level) and 0.3896 (instance-level). Similar trends are observed for other FAR applications and datasets. In addition, we also perform a robustness study to investigate two important aspects of the proposed fairness attack, namely the number of poisons and the perturbation radius ϵ (the results of robustness study are in Appendix).

Defense Evaluation. The results on defending the proposed attack are reported in Table 2. To show the efficacy of the proposed robust-fair training method, we doubled the number of poisons to attack the fairness of the face attribute classifiers. It is observed that the defense could effectively reduce the bias introduced by the proposed fairness attack. For instance, in attractiveness prediction, the equalized odds of the LightCNN classifier was 0.5432 after the attack (instance-level). With robust-fair training (our defense), the attacked classifier becomes more fair with an equalized odds of 0.2456. The improved fairness as well as the model accuracy could also be observed for other network architectures and FAR applications. However, we acknowledge that the model's fairness after defense is still worse than the unattacked model, which indicates that future work is needed to further increase the robustness of the fairness in FAR applications.

6 Conclusion

In this work, we present a novel poisoning attack against demographic fairness of face attribute classifiers. To the best of our knowledge, our work is the first to explore test-data-free, cross-attribute and clean-label poisoning attacks against fairness of face attribute classifiers at group level. Experiments on various FAR applications show that our method could easily bias the fairly trained models. Motivated by the observed vulnerability of fairness, we further propose an efficient defense mechanism to increase the robustness of the fairness. Through this work, we stress the significance of studying *robustness of fairness* for AI models, as the fairness obtained with traditional fair algorithms could be vulnerable under fairness attacks. Therefore, decent efforts should be made to improve the *robustness of the fairness* of ML/AI models.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. IIS-2202481, CHE-2105005, IIS-2008228, CNS-1845639, CNS-1831669. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Ethics Statement

Understanding the robustness of fairness is extremely important in identity-sensitive applications. In addition to fair ML and AI algorithms, we argue that it is equally important to develop robust fairness in identify-sensitive applications, as fair models could be vulnerable under the fairness attack and fair ML algorithms do not automatically enable the robustness of the fairness. Ignoring the robustness of demographic fairness in ML could harm the social equality. Therefore, this work investigates the robustness of fairness from both attacking and defense perspective, such that robust fair models could be developed for identity-sensitive applications.

Our work serves as a prototype and baseline for future studies on robust fair machine learning. In this work, we designed a prototype poisoning attack against fairness of visual models. Motivated by the vulnerability of the attack, we further present a novel defense mechanism to increase the robustness of the fairness. We stress that the core intention of this work is to inspire future work on the robustness of demographic fairness in DNNs. We believe both our attack and defense could be used as baselines for future work on robust fair ML and AI systems.

References

[Angwin *et al.*, 2016] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. How we analyzed the compas recidivism algorithm, 2016.

[Bellamy *et al.*, 2018] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

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

[Chang *et al.*, 2020] Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv preprint arXiv:2006.08669*, 2020.

[Chen *et al.*, 2017] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[Ding *et al.*, 2018] Hui Ding, Hao Zhou, Shaohua Zhou, and Rama Chellappa. A deep cascade network for unaligned face attribute classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Fowl *et al.*, 2021] Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, and Tom Goldstein. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.

[Geiping *et al.*, 2020] Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Han *et al.*, 2021] Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through training reweighting. *arXiv preprint arXiv:2109.08253*, 2021.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2020] W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. *arXiv preprint arXiv:2004.00225*, 2020.

[Huang *et al.*, 2021] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. *arXiv preprint arXiv:2101.04898*, 2021.

[Kärkkäinen and Joo, 2019] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[Karkkainen and Joo, 2021] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[Kohavi, 1996] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[Kou *et al.*, 2021] Ziyi Kou, Lanyu Shang, Huimin Zeng, Yang Zhang, and Dong Wang. Exgfair: A crowdsourcing data exchange approach to fair human face datasets augmentation. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1285–1290. IEEE, 2021.

[Li *et al.*, 2021] Xiaoxiao Li, Ziteng Cui, Yifan Wu, Lin Gu, and Tatsuya Harada. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*, 2021.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 12 2015.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Mehrabi *et al.*, 2020] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. *arXiv preprint arXiv:2012.08723*, 2020.

[Scimeca *et al.*, 2021] Luca Scimeca, Seong Joon Oh, Sanghyuk Chun, Michael Poli, and Sangdoo Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. *arXiv preprint arXiv:2110.03095*, 2021.

[Shafahi *et al.*, 2018] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *arXiv preprint arXiv:1804.00792*, 2018.

[Shen *et al.*, 2017] Sijie Shen, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Fooling neural networks in face attractiveness evaluation: Adversarial examples with high attractiveness score but low subjective score. In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, pages 66–69. IEEE, 2017.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Tariq *et al.*, 2022] Shahroz Tariq, Sowon Jeon, and Simon S Woo. Am i a real or fake celebrity? evaluating face recognition and verification apis under deepfake impersonation attack. In *Proceedings of the ACM Web Conference 2022*, pages 512–523, 2022.

[Torralba and Efros, 2011] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[Van *et al.*, 2021] Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. Poisoning attacks on fair machine learning. *arXiv preprint arXiv:2110.08932*, 2021.

[Wen *et al.*, 2022] Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H Chi. Distributionally-robust recommendations for improving worst-case user experience. In *Proceedings of the ACM Web Conference 2022*, pages 3606–3610, 2022.

[Wu *et al.*, 2018] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[Yue *et al.*, 2021] Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. Contrastive domain adaptation for question answering using limited text corpora. *arXiv preprint arXiv:2108.13854*, 2021.

[Zeng *et al.*, 2022] Huimin Zeng, Zhenrui Yue, Lanyu Shang, Yang Zhang, and Dong Wang. Boosting demographic fairness of face attribute classifiers via latent adversarial representations. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1588–1593. IEEE, 2022.

[Zeng *et al.*, 2023] Huimin Zeng, Zhenrui Yue, Ziyi Kou, Yang Zhang, Lanyu Shang, and Dong Wang. Fairness-aware training of face attribute classifiers via adversarial robustness. *Knowledge-Based Systems*, 264:110356, 2023.

[Zhao *et al.*, 2017] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.