FairSAM: Fair Classification on Corrupted Data Through Sharpness-Aware Minimization

Anonymous authors
Paper under double-blind review

Abstract

Image classification models trained on clean data often suffer from significant performance degradation when exposed to corrupted testing or deployment data, such as images with impulse noise, Gaussian noise, or environmental noise. This degradation not only impacts overall performance but also disproportionately affects various demographic subgroups, raising critical algorithmic bias concerns. Although robust learning algorithms such as Sharpness-Aware Minimization improve overall model robustness and generalization, they do not address biased performance degradation across demographic subgroups. Existing fairness-aware machine learning methods aim to reduce performance disparities but struggle to maintain robust and equitable accuracy across demographic subgroups when faced with data corruption. This reveals an inherent tension between robustness and fairness when dealing with corrupted data. To address these challenges, we introduce a newly-designed metric to assess performance degradation across subgroups under data corruption. We propose **FairSAM**, a framework that integrates Fairness-oriented strategies into SAM to deliver equalized performance across demographic groups under corrupted conditions. Our experiments on multiple real-world datasets and various predictive tasks show that FairSAM reconciles robustness and fairness. The framework yields a structured solution for fair and robust image classification in the presence of data corruption.

1 Introduction

Deep neural networks have shown remarkable success in various AI applications, including image classification, image segmentation, and object detection. However, corrupted data poses a set of challenges in the applications of deep neural networks on common AI applications. In this paper, we focus on one critical scenario in machine learning systems where image classification models face corrupted data from diverse sources including user-uploaded images, mobile device captures, and network transmission artifacts. We investigate unequal model degradation when well-trained models (on clean data) are applied to corrupted images caused by impulse noise, Gaussian noise, weather effects (e.g., snow, fog), and motion blur during capture or transfer. Recent work (Nanda et al., 2021) shows that this degradation disproportionately affects demographic subgroups, i.e., accuracy drop significantly varies across subgroups and raises algorithmic bias concerns. Thereby, the application of well-trained models (on clean data) to corrupted images raise critical algorithmic bias concerns.

Although various fairness-aware methods, such as fairness constraints (Calders & Verwer, 2010; Kamiran et al., 2010; Corbett-Davies et al., 2017; Zafar et al., 2017) or reweighing strategies (Calders et al., 2009), have been proposed to address bias in machine learning models, they often fail to mitigate unequal accuracy degradation across subgroups and do not address the broader robustness challenge.

Recent research frames the classification of corrupted data as a robustness and generalization challenge, as its learning objective is to maintain performance when exposed to noise perturbations that differ from the training distribution. Sharpness-Aware Minimization (SAM) (Foret et al., 2021) has emerged as an effective approach to tackle robustness issues by promoting “flat" minima in the loss landscape, where the loss changes gradually with parameter variations. This property enhances generalization and resilience to data corruption

and improves overall model robustness. However, while SAM improves robustness at a broad level, it does not inherently address fairness across demographic subgroups. The gains in performance achieved through SAM tend to be unevenly distributed, leaving certain disadvantaged subgroups more susceptible to accuracy degradation. This disparity highlights a critical limitation of SAM in scenarios where both robustness and fairness are equally essential.

To address these inherent challenges, we first formulate the fair classification problem in the context of corrupted data. **Specifically, we focus on a setting where a model is trained on a clean and noise-free dataset and is tested on a corrupted dataset containing various types of noise.** This setting is inspired by the real-world scenario where training data are carefully curated while the trained model might be tested on unexpected conditions, including corrupted data in the wild. We then evaluate model performance and assess performance degradation across demographic subgroups, such as age (young/non-young) and gender (female/male), to understand both robustness and fairness under corrupted conditions. We introduce a novel metric to quantify fairness in performance degradation under corruption, **which differs subtly from existing one-shot fairness notions that mandate equal robustness across population partitions to imperceptible input perturbations.** The metric, ***Corrupted Degradation Disparity***, captures the difference in accuracy degradation (i.e., the drop in accuracy between clean and corrupted data) between specific subgroups, such as young and non-young individuals. Based on this metric, we propose **FairSAM**, the first framework to incorporate fairness-oriented strategies into SAM. Specifically, we develop an instance-reweighed SAM and approximate the per-sample perturbation in a per-batch perturbation learning algorithm to promote fairness and robustness simultaneously. FairSAM distributes robustness improvements equitably across demographic subgroups. It is designed to address fairness concerns and maintains high overall accuracy under corrupted conditions.

We conduct experiments on multiple real-world datasets, including the imbalanced CelebA dataset and the balanced FairFace dataset, across various prediction tasks with different combinations of target and sensitive attributes to thoroughly evaluate FairSAM's effectiveness in terms of both robustness and fairness. Our results demonstrate that FairSAM addresses the tension between robustness and fairness. It attains significantly better values on our proposed fairness metric *Corrupted Degradation Disparity* (lower is better).

Our contributions are as follows: 1) We identify and formalize the robustness bias challenge in image classification under data corruption, introducing a novel fairness metric *Corrupted Degradation Disparity* to evaluate the fairness of performance degradation across demographic subgroups. 2) We introduce **FairSAM**, a novel framework that integrates SAM with fairness-enhancing strategies to achieve both robustness and fairness in corrupted image classification tasks. 3) We validate the effectiveness of FairSAM on multiple datasets and multiple configuration conditions, demonstrating its consistent superior performance in both accuracy and fairness compared with various baselines.

2 Related Works

2.1 Fairness in ML

Algorithmic fairness has emerged as a critical topic in machine learning, with increasing awareness of biases that disproportionately affect marginalized groups based on demographic factors like gender, race, or age. In the context of image classification, these biases can manifest as unequal performance across subgroups, especially under challenging conditions like image corruption. Despite its importance, fairness in the presence of image corruption remains an underexplored area. This gap is significant, as machine learning models deployed in real-world environments frequently encounter corrupted data, which can amplify existing inequalities by disproportionately impacting certain demographic groups.

Existing research on fair machine learning focuses primarily on two objectives: (1) defining and identifying bias in machine learning models and (2) developing algorithms to effectively mitigate bias. Various fairness definitions have been proposed, with *statistical parity* being one of the most widely recognized. Statistical parity ensures that the likelihood of favorable outcomes remains similar across protected and non-protected groups. This can be quantified through metrics like *risk difference*, *risk ratio*, *relative change*, and *odds ratio* (Žliobaite, 2017). These metrics provide a structured way to quantify fairness and to assess and compare bias

in model predictions. Bias mitigation strategies fall into three main categories: pre-processing, in-processing, and post-processing methods. *Pre-processing* techniques modify the training data to remove potential biases before model training. Examples include *Massaging* (Kamiran & Calders, 2009), *reweighing* (Calders et al., 2009), and *Preferential Sampling* (Kamiran & Calders, 2012), which adjust data distributions to promote fairness. In contrast, *in-processing* methods (Calders & Verwer, 2010; Corbett-Davies et al., 2017; Kamiran et al., 2010; Zafar et al., 2017) introduce fairness constraints or regularization terms directly into the model's objective function, ensuring that the learning algorithm prioritizes fairness alongside accuracy. Lastly, *post-processing* techniques (Awasthi et al., 2020; Hardt et al., 2016; Kamiran et al., 2012) adjust model predictions after training to correct for any biases detected in model outputs.

Despite significant progress, most of these methods have not specifically addressed fairness issues arising from image corruption, where different groups may experience varying degrees of accuracy loss. This gap in the literature motivates our work as we seek to address both robustness and fairness under image corruption conditions.

2.2 Fairness and Robustness

While the foundational work in algorithmic fairness primarily focused on achieving fairness across different demographic subgroups, real-world deployments have revealed the fragility of these guarantees. A model deemed "fair" in the laboratory can exhibit significant bias when faced with the natural variations of a production environment or the deliberate manipulations of an adversary. This has led to a critical expansion of the fairness research to include robustness to various forms of bias, including distribution shift and adversarial attacks.

Fairness and Distribution Shift: Distribution shift poses significant challenges to maintaining fairness across different demographic groups. Sagawa et al. (2020) demonstrated that models can exhibit disparate performance across subgroups when the test distribution differs from the training distribution, particularly affecting minority groups. Koh et al. (2021) introduced the WILDS benchmark to systematically evaluate model performance under various types of distribution shift. This work highlights how demographic subgroups can be disproportionately affected. Liu et al. (2021) showed that standard domain adaptation techniques can inadvertently amplify bias, while Zhao et al. (2024) proposed methods to maintain fairness guarantees under covariate shift.

Fairness and Adversarial Attacks: The vulnerability of fair models to adversarial perturbations has received considerable attention. Xu et al. (2021) demonstrated that adversarially trained models often exhibit increased bias against minority groups. This reveals a fundamental tension between adversarial robustness and fairness. Sun et al. (2022) proposed adversarial training methods that explicitly account for fairness constraints, while Tran et al. (2022) showed that certain demographic groups are more susceptible to adversarial attacks than others. Mehrabi et al. (2021); Van et al. (2022); Solans et al. (2020) further explored how adversarial examples can be crafted to specifically target fairness, leading to disproportionate performance degradation across subgroups. Nanda et al. (2021) showed that different demographic subgroups exhibit different levels of robustness, and that this disparity can produce unfairness.

Most current work emphasizes adversarial robustness rather than robustness to natural corruption, and few methods explicitly tackle the intersection of fairness and robustness under image corruption. This work investigates that intersection and proposes a direction to mitigate such bias.

2.3 SAM and its Variants

Sharpness-Aware Minimization (Foret et al., 2021) was introduced to improve neural network generalization by identifying flatter minima in the loss landscape. SAM minimizes the maximum loss within a neighborhood around the current parameter setting rather than minimizing the loss at a single point. This approach yields solutions that are more resilient to small parameter perturbations, enhancing the model's generalization and robustness.

ImbSAM (Zhou et al., 2023) extends SAM's applicability to settings with extremely imbalanced data distributions, addressing the trade-off between sharpness and data imbalance. By incorporating strategies to

handle imbalance, ImbSAM enhances generalization for certain long-tail classes and improves robustness in challenging training scenarios. Adaptive Sharpness-Aware Pruning (AdaSAP) (Bair et al., 2023) advances SAM's concept by focusing on pruning models to enhance both compactness and robustness. AdaSAP employs adaptive weight perturbations to regularize pruned models, improving their resilience against corrupted data while maintaining efficient model size. However, none of these SAM variants specifically target fairness across demographic subgroups, particularly under image corruption.

Our work advances SAM in a novel direction, focusing on fairness and robustness to image corruption. We incorporate fairness mechanisms to ensure robust performance without sacrificing equity across sensitive demographic groups. This work addresses critical limitations in both SAM and its variants, bridging the gap between robustness and fairness in corrupted image classification.

The foregoing lines of work address fairness under distribution shift or under adversarial perturbation, but not under *natural* corruption (e.g., sensor noise and weather conditions) that differs from both full distribution shift and adversarial attacks. Fairness under such natural corruption remains underexplored. This work fills this gap by introducing a metric and a method for equitable robustness when test data are corrupted in this way.

3 Preliminary

3.1 Fair Classification

We first formulate the fair classification problem under corrupted conditions. Consider a clean training dataset, denoted as $\mathcal{D}_\mathcal{T} = \left\{(\mathbf{x}_i, \mathbf{y}_i, s_i)\right\}_{i=1}^N$, and a corrupted testing dataset, represented by $\mathcal{D}_\mathcal{C} = \left\{(\mathbf{x}_i^c, \mathbf{y}_i, s_i)\right\}_{i=1}^M$, where $\mathbf{x}_i \in \mathcal{X}$ denotes an input image, $\mathbf{y}_i \in \mathcal{Y}$ denotes the ground truth target, and $s_i \in \mathcal{S} = \{s^+, s^-\}$ represents a sensitive attribute, with s^+ and s^- denoting the advantaged and disadvantaged groups, respectively. Here $\mathbf{x}_i^c$ denotes a corrupted input (e.g., an image with noise). The classification hypothesis space is formalized as $f(\mathbf{w}) : \mathcal{X} \to \mathcal{Y}$, parameterized by $\mathbf{w}$. Traditionally, fair classification aims to ensure that the model has equal outcome (e.g., demographic parity) or performance (e.g., equalized odds) among different demographic subgroups.

3.2 Sharpness-Aware Minimization

Sharpness-Aware Minimization (SAM) is an optimization technique that enhances the generalization capability of neural networks by mitigating overfitting. Unlike traditional methods that solely minimize the loss at the current parameter values, SAM minimizes the maximum loss within a neighborhood around the current parameters. This strategy encourages the model to find "flatter" minima in the loss landscape, where surrounding regions exhibit uniformly low loss, improving generalization and robustness.

Consider a family of models parameterized by $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$, where $\mathcal{L}$ is the loss function, and $\mathcal{D}_\mathcal{T}$ represents the training dataset. SAM aims to minimize an upper bound on the PAC-Bayesian generalization error. For a given $\rho > 0$ and norm $\|\cdot\|_p$ (typically $p = 2$), this bound is expressed as follows:

$$\mathcal{L}(\mathbf{w}) \leq \max_{\|\epsilon\|_p \leq \rho} \left[\mathcal{L}_{\mathcal{D}_\mathcal{T}}(\mathbf{w} + \epsilon) - \mathcal{L}_{\mathcal{D}_\mathcal{T}}(\mathbf{w})\right] + \mathcal{L}_{\mathcal{D}_\mathcal{T}}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{1}$$

Therefore, the problem is a minimax problem:

$$\min_{\mathbf{w}} \max_{\|\epsilon\|_p \leq \rho} \mathcal{L}_{\mathcal{D}_\mathcal{T}}(\mathbf{w} + \epsilon) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{2}$$

To solve the above minimax problem, SAM performs an iterative update at each iteration t, where the update steps are as follows:
Firstly, compute the perturbation ϵ_t as:

$$\epsilon_t = \frac{\rho \cdot \operatorname{sign}\left(\nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}}\left(\mathbf{w}_{t-1}\right)\right) \left|\nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}}\left(\mathbf{w}_{t-1}\right)\right|^{q-1}}{\left(\left\|\nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}}\left(\mathbf{w}_{t-1}\right)\right\|_q^q\right)^{1/p}} \tag{3}$$

where $1/p + 1/q = 1$, $\rho > 0$ is a hyperparameter controlling the neighborhood size, and p, q denote norm parameters. Typically, p and q are set to 2.
Secondly, update the model parameter $\mathbf{w}_t$ as:

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \left(\nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}} \left(\mathbf{w}_{t-1} + \epsilon_t \right) + \lambda \mathbf{w}_{t-1} \right) \tag{4}$$

where $\lambda > 0$ is the parameter for weight decay, and $\eta_t > 0$ is the learning rate.
With $p = q = 2$, introducing an intermediate variable $\mathbf{u}_t$, we have:

$$\mathbf{u}_t = \mathbf{w}_{t-1} + \frac{\rho \nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}} \left(\mathbf{w}_{t-1} \right)}{\| \nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}} \left(\mathbf{w}_{t-1} \right) \|}, \tag{5}$$

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \left(\nabla \mathcal{L}_{\mathcal{D}_\mathcal{T}} \left(\mathbf{u}_t \right) + \lambda \mathbf{w}_{t-1} \right). \tag{6}$$

By minimizing loss over a neighborhood rather than a single point, SAM finds parameter configurations that are less sensitive to small perturbations and that generalize better under corruption. This property of SAM motivates our development of FairSAM, which extends SAM to also address fairness concerns across demographic subgroups under corrupted conditions.

4 Proposed Research

4.1 New Fairness Notions for Corrupted Data

Traditional fair machine learning focuses primarily on performance disparity. However, performance degradation is unevenly distributed across subgroups when data is corrupted. Our objective is to address this robustness-based fairness by training a model $f(\mathbf{w})$ that maintains consistent performance degradation across distinct subgroups, as measured by a specified metric. We formally define a new fairness metric for corrupted data as follows:

Definition 1 (Corrupted Degradation Disparity)**.** Firstly, given a model f trained on clean training data $\mathcal{D}_\mathcal{T}$, we define *Corrupted Degradation* for a specific demographic group s as:

$$\Delta p^s = |\mathbb{M}({\mathcal{D}_\mathcal{T}}^{S=s}, f) - \mathbb{M}({\mathcal{D}_\mathcal{C}}^{S=s}, f)|,$$

where $\mathbb{M}({\mathcal{D}_\mathcal{T}}^{S=s}, f)$ and $\mathbb{M}({\mathcal{D}_\mathcal{C}}^{S=s}, f)$ represent the performance metrics on clean training and corrupted testing data, respectively, for subgroup s. We then define *Corrupted Degradation Disparity* as:

$$\Delta p = |\Delta p^{s^+} - \Delta p^{s^-}|.$$

□

This definition quantifies how differently subgroups are impacted by data corruption. By measuring this disparity, we can identify subgroups that are disproportionately affected. Specifically, a smaller Δp value indicates that the model's robustness is more equally distributed across subgroups.

4.2 Fair SAM: Instance-Reweighed SAM

Sharpness-Aware Minimization (SAM) has demonstrated effectiveness in enhancing model robustness by encouraging "flat" minima in the loss landscape, where loss exhibits gradual changes with respect to parameter variations. This characteristic enhances generalization and increases resilience to data corruption. However, SAM does not adequately address fairness concerns, as the resulting robustness improvements are not uniformly distributed across demographic subgroups. In particular, the accuracy gains achieved by SAM tend to be unevenly allocated, with certain disadvantaged subgroups experiencing disproportionately higher levels of accuracy degradation.

To address this limitation, we introduce a reweighing mechanism that adjusts sample importance across subgroups so that SAM prioritizes both robustness and fairness. By allocating greater attention to samples

from disadvantaged subgroups, this reweighing scheme balances gradient contributions from each group, mitigates robustness disparities, and ensures more equitable performance across demographic subgroups.

Following this intuition, we derive the gradient update for the SAM training procedure. Given a clean training dataset $\mathcal{D}_{\mathcal{T}} = \{(\mathbf{x}_i, \mathbf{y}_i, s_i)\}_{i=1}^{N}$, the vanilla training objective without the reweighing mechanism is formulated as:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i}^{n} \max_{\|\epsilon_i\|_2 \leq \rho} \ell_i(\mathbf{w} + \epsilon_i) \tag{7}$$

where ℓ_i is the loss for sample i. In this unweighted form, the classifier tends to focus more on the majority-favored group. Perturbation magnitudes ϵ then differ across subgroups and yield different generalization capabilities across groups. We propose assigning initial weights to samples within each group as $\gamma_i = \frac{c}{n'}$, where n' represents the number of samples in the group to which the i-th sample belongs, and c is a scaling constant. By assigning weights to samples of different groups, the classifier is encouraged to focus more on samples that are either misclassified or likely to be misclassified. Moreover, this approach ensures that the weighted representation of samples remains balanced across different groups. This can be achieved by constraining the sum of weights within groups and is formulated as:

$$\max_{\gamma} \sum_{s} \sum_{i \in g_s} \gamma_i \max_{\|\epsilon_i\|_2 \leq \rho} \ell_i(\mathbf{w} + \epsilon_i) \ s.t. \sum_{i \in g_s} \gamma_i = c, p_i \geq 0 \tag{8}$$

where g_s collects the indices of samples belonging to the demographic group s. The optimization problem equation 8 can be partitioned by groups as follows for each demographic group:

$$\max_{\gamma} \sum_{i}^{n'} \gamma_i \underbrace{\max_{\|\epsilon_i\|_2 \leq \rho} \ell_i(\mathbf{w} + \epsilon_i)}_{\text{per-sample SAM } \ell_s} \ s.t. \gamma^T 1 = c, \gamma_i \geq 0. \tag{9}$$

Here we perform reweighing within each specific group to ensure the weighted samples are balanced between groups.

4.3 Perturbation for Per-Batch Calculation

SAM calculates the per-batch perturbation ϵ instead of per-instance perturbation ϵ_i, practically.

$$\min_{\mathbf{w}} \max_{\|\epsilon\|_p \leq \rho} \frac{1}{n} \sum_{i}^{n} \ell_i(\mathbf{w} + \epsilon). \tag{10}$$

To facilitate the calculation, we seek a per-batch perturbation ϵ' that is aligned with per-instance perturbation ϵ_i. We start by considering the second-order expansion of a general empirical risk ℓ_i for instance i:

$$\ell_i(\mathbf{w} + \epsilon) \approx \ell_i(\mathbf{w}) + \nabla \ell_i(\mathbf{w})^T \epsilon + \frac{1}{2} \epsilon^T H_i(\mathbf{w}) \epsilon. \tag{11}$$

For a tractable theoretical analysis, we assume the Hessian to be a low-rank, positive definite matrix $H_i(w) = a_i \nabla \ell_{i,p}(w) \nabla \ell_{i,p}(w)^T, (a_i > 0)$. Then the gradient of per-instance SAM can be calculated by using one-time backpropagation on a reweighed batch:

$$\begin{aligned} \nabla \ell_b(w) &= \nabla \big(\sum_{i}^{N} g_i \ell_i(w) \big) \\ \epsilon^* &= \rho \frac{\nabla \ell_b(w)}{\|\nabla \ell_b(w)\|_2} \end{aligned} \tag{12}$$

where $g_i = a_i \|\nabla \ell_i\|_2$ provides the group-specific weight for each instance, ensuring both fairness and robustness in the model. Then, we use the estimated ϵ^* to update the fairness-aware weight p_i.

5 Experiments

5.1 Datasets and Experiment Settings

We evaluate our proposed method, FairSAM, on several widely-used datasets, including CelebA (Liu et al., 2015), FairFace (Karkkainen & Joo, 2021), and LFW (Huang et al., 2007), to examine its effectiveness in achieving both robustness and fairness under corrupted data conditions. Specifically, CelebA, an imbalanced dataset with a substantial disparity in sample sizes between the advantaged and disadvantaged groups, presents a significant challenge for ensuring fairness across these subgroups. Specifically, we select "Big Nose" and "Blond Hair" as target attributes and "Gender" and "Age" as sensitive attributes. In contrast, FairFace, a balanced dataset with a roughly equal distribution of samples across demographic groups, provides a controlled setting for evaluating FairSAM's performance in maintaining fairness under balanced conditions. We choose "Age" as the sensitive attribute and "Gender" as the target attribute. These datasets are representative of corrupted-data scenarios and support comprehensive evaluation of the proposed method across diverse settings. To investigate the fairness-aware generalization capabilities of FairSAM, we conduct out-of-distribution robustness experiments using the CelebA and LFW datasets, where models are trained on one dataset and tested on the other, further highlighting the method's ability to address fairness in challenging scenarios. The models are all implemented using PyTorch and evaluated on an Ubuntu 20.04 LTS server with an Intel(R) Core(TM) i9-10900X CPU, 128GB memory, and an Nvidia GeForce RTX 3070 GPU. We set the hyperparameters to $c = 1$ and $\epsilon = 0.05$. All source code is included in a code appendix.

Methods	Test Data	Acc s^+	Δp^{s^+}	Acc s^-	Δp^{s^-}	Accuracy ↑	ΔAcc ↓	Δp ↓
Vanilla	clean	0.8572	0.0115	0.7171	0.0862	0.8232	0.2148	0.0747
	corrupted	0.8457		0.6309		**0.7901**		
FairReg	clean	0.6530	0.0215	0.6492	0.0588	0.6517	**0.0411**	<u>0.0373</u>
	corrupted	0.6315		0.5904		0.6217		
Reweighed	clean	0.8527	0.0436	0.7156	0.0836	0.7983	0.1771	0.0400
	corrupted	0.8091		0.6320		0.7662		
SAM	clean	0.8590	0.0090	0.7043	0.0666	0.8215	0.2123	0.0576
	corrupted	0.8500		0.6377		0.7984		
GroupSAM	clean	0.8571	0.0074	0.7046	0.0534	0.8199	0.1985	0.0460
	corrupted	0.8497		0.6512		0.7809		
FairSAM (Ours)	clean	0.8574	0.0399	0.7480	0.0499	0.8310	<u>0.1194</u>	**0.0100**
	corrupted	0.8175		0.6981		<u>0.7885</u>		

Table 1: **Performance and fairness trade-off comparison on the CelebA Dataset.** The target attribute is "Big Nose" and the sensitive attribute is "Age". The corruption is set to level-3 snow noise. FairSAM achieves the best performance in terms of Δp and demonstrates superior fairness in accuracy degradation across subgroups. It ranks second in ΔAcc while incurring the smallest accuracy drop compared to the vanilla accuracy and effectively balances robustness and fairness.

5.2 Baseline Methods

All experiments are conducted using the ResNet-18 model architecture. We adapt ImbSAM (Zhou et al., 2023) to enhance fairness-aware robustness by selectively applying SAM to the disadvantaged subgroup during training, introducing an approach we term Group-specific SAM (**GroupSAM**). Specifically, we first reformulate Eq. 2 as follows:

$$\min_{\mathbf{w}} \max_{\epsilon \leq \rho} \left[\mathcal{L}_{s^+}(\mathbf{w}+\epsilon) + \mathcal{L}_{s^-}(\mathbf{w}+\epsilon)\right] + \frac{\lambda}{2}\|\mathbf{w}\|^2, \tag{13}$$

where $\mathcal{L}_{s^+}$ and $\mathcal{L}_{s^-}$ represent the losses for the advantaged and disadvantaged groups, respectively. To enhance fairness, we then modify this by applying SAM only to the disadvantaged group, resulting in:

$$\min_{\mathbf{w}} \max_{\epsilon \leq \rho} \mathcal{L}_{s^-}(\mathbf{w}+\epsilon^-) + \mathcal{L}_{s^+}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{14}$$

where ϵ^- is a perturbation specific to the disadvantaged group. The perturbation ϵ^- in Eq. 14 differs from ϵ in Eq. 13 because SAM is applied only to the disadvantaged group, so the perturbation is group-specific. To make explicit our sharpness-aware term, the above optimization target can be rewritten as follows:

$$\min_{\mathbf{w}} \overbrace{\max_{\epsilon \leq \rho} \left[\mathcal{L}_{s^-}(\mathbf{w}+\epsilon) - \mathcal{L}_{s^-}(\mathbf{w})\right] + \mathcal{L}_{s^-}(\mathbf{w})}^{\text{Disadvantaged group-specific SAM term}} + \mathcal{L}_{s^+}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2. \tag{15}$$

Additionally, we train several comparison models, including vanilla ResNet-18 (**Vanilla**), ResNet-18 with fairness regularizers (**FairReg**), Reweighed ResNet-18 (**Reweighed**), Group-Specific SAM (**GroupSAM**), SAM, and FairSAM. We evaluate model performance on both clean and corrupted versions of the test data, focusing on accuracy across demographic subgroups (Young and Non-Young) under noise-free and noise-corrupted conditions. Following the noise settings from ImageNet-C (Hendrycks & Dietterich, 2019), we add various noise at 5 levels of severity (1 to 5) and with varying noise types, such as snow, Gaussian, and blur.

Methods	Test Data	Acc s^+	Δp^{s^+}	Acc s^-	Δp^{s^-}	Accuracy ↑	ΔAcc ↓	Δp ↓
Vanilla	clean	0.9769	0.0006	0.9318	0.1083	0.9493	0.1528	0.1077
	corrupted	0.9763		0.8235		0.8827		
FairReg	clean	0.9442	0.0281	0.9532	0.1094	0.9387	0.1465	0.0813
	corrupted	0.9723		0.8258		0.8824		
Reweighed	clean	0.9627	0.0074	0.9351	0.1056	0.9457	0.1406	0.1130
	corrupted	0.9701		0.8295		0.8839		
SAM	clean	0.9797	0.0168	0.9363	0.0575	0.9531	0.0841	0.0407
	corrupted	0.9629		0.8788		0.9114		
GroupSAM	clean	0.9780	0.0094	0.9406	0.0468	0.9551	<u>0.0748</u>	<u>0.0374</u>
	corrupted	0.9686		0.8938		<u>0.9228</u>		
FairSAM (Ours)	clean	0.9734	0.0202	0.9412	0.0275	0.9570	**0.0395**	**0.0073**
	corrupted	0.9532		0.9137		**0.9291**		

Table 2: **Performance and fairness trade-off comparison on the CelebA Dataset.** The target attribute is "Blond Hair" and the sensitive attribute is "Gender". The corruption is set to level-3 Gaussian noise. FairSAM demonstrates superior performance compared to all baseline methods, achieving higher accuracy while outperforming others on fairness metrics Δp and ΔAcc, showcasing its effectiveness in balancing robustness and fairness.

Methods	Test Data	Acc s^+	Δp^{s^+}	Acc s^-	Δp^{s^-}	Accuracy ↑	Δp ↓
Vanilla	clean	0.7396	0.1704	0.8296	0.2008	0.7800	0.0304
	corrupted	0.5692		0.6288		0.5961	
SAM	clean	0.7727	0.1518	0.8781	0.1732	0.8201	0.0214
	corrupted	0.6209		0.7049		<u>0.6885</u>	
GroupSAM	clean	0.7550	0.1253	0.8333	0.1441	0.7904	<u>0.0188</u>
	corrupted	0.6297		0.6892		0.6563	
FairSAM	clean	0.7837	0.1467	0.9075	0.1539	0.8394	**0.0072**
	corrupted	0.6370		0.7536		**0.6893**	

Table 3: **Performance and fairness trade-off comparison on the FairFace Dataset.** The target attribute is "Gender" and the sensitive attribute is "Age". The corruption is set to level-5 Gaussian noise. FairSAM outperforms all baseline methods and achieves the highest accuracy and the best fairness as measured by Δp. This demonstrates its effectiveness under severe corruption.

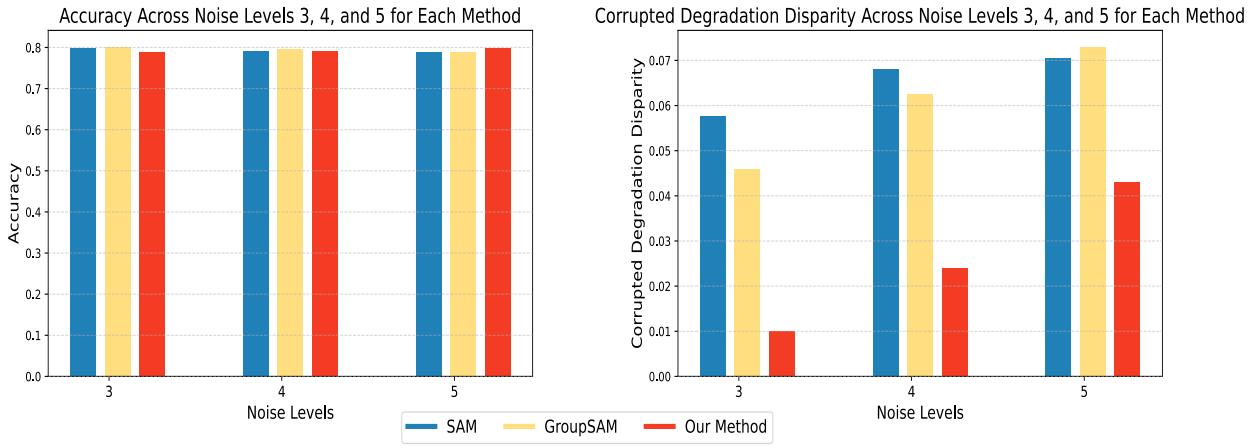

Figure 1: Comparison of SAM-based methods among varying noise levels. FairSAM achieves comparable or even better accuracy while maintaining the lowest bias Δp.

Method	Train → Test	Acc s^+	Acc s^-	Δp ↓
SAM	CA → CA	0.8590	0.7043	0.0210
	CA→LFW	0.5946	0.4189	
GroupSAM	CA→CA	0.8570	0.7042	0.1865
	CA→LFW	0.5363	0.5700	
FairSAM	CA→CA	0.8575	0.7480	**0.0188**
	CA→LFW	0.5341	0.4058	

Table 4: **Results of training on the CelebA (CA for abbr.) dataset and testing on the LFW dataset**. Target attribute is “Big Nose" and the sensitive attribute is “Age".

5.3 Trade-off between Fairness and Performance

We investigate robustness disparity across subgroups on both balanced and unbalanced datasets, examining various sensitive and target attributes. We train models using the clean training data and evaluate those models on noisy data. Tab. 1, Tab. 2, and Tab. 3 show the comparison of methods regarding accuracy, corrupted degradation disparity, and fairness promotion. Notably, values in bold indicate the **best** performance among all methods, while values that are underlined represent the second-best performance. We run all experiments *three times* and report the average accuracy and omit the standard deviation since the training is relatively stable (i.e., usually less than 0.1% standard deviation).

Corrupted Degradation Disparity. The result for degradation disparity is reported in the Δp column in Tab. 1 and Tab. 2. Our proposed method, FairSAM, consistently achieves the best balance of fairness and accuracy among all comparable baselines, indicating its effectiveness in mitigating the robustness bias gap. Specifically, FairSAM consistently outperforms fairness-aware methods, such as FairReg and Reweighed, which improve subgroup fairness but suffer from notable accuracy degradation. In contrast, while SAM and GroupSAM show strong generalization and robustness against corrupted data, they fall short in ensuring fairness across demographic subgroups, as evidenced by their higher disparities in performance degradation.

To further validate our methods, we conduct experiments on FairFace, a balanced dataset designed for fairness assessment. As shown in Tab. 3, the results on FairFace are consistent with the trends observed on CelebA. FairSAM achieves the lowest level of corrupted bias among all baseline methods, demonstrating its effectiveness in promoting fairness across demographic groups. Additionally, FairSAM attains the highest accuracy, benefiting from its balanced approach that considers samples from both advantaged and disadvantaged groups. This result underscores FairSAM's ability to maintain robust performance while achieving fairness, even in datasets with balanced demographic distributions.

Method	Train → Test	Acc s^+	Acc s^-	$\Delta p \downarrow$
SAM	LFW→LFW	0.7714	0.7687	0.1456
	LFW→CA	0.6863	0.5380	
GroupSAM	LFW→LFW	0.7784	0.7963	0.1499
	LFW→CA	0.6590	0.5270	
FairSAM	LFW→LFW	0.7893	0.7984	**0.1066**
	LFW→CA	0.6668	0.5511	

Table 5: **Results of training on the LFW dataset and testing on the CelebA (CA for abbr.) dataset.** Target attribute is "Big Nose" and the sensitive attribute is "Age".

Performance Disparity. In the column of ΔAcc, we focus on performance disparity among subgroups in corrupted images. Tab. 1 and Tab. 2 demonstrate that FairSAM effectively enhances fairness across diverse target and sensitive attributes, maintaining consistent performance. Although FairReg achieves a notable fairness level, this improvement comes at a substantial cost to the performance of the advantaged group, resulting in reduced overall accuracy. FairSAM balances fairness and accuracy and yields equitable outcomes without compromising the performance of any subgroup.

5.4 Ablation Study

To thoroughly evaluate the robustness and fairness of the proposed FairSAM framework, we conduct an ablation study across multiple noise levels. This study is designed to determine whether FairSAM consistently achieves optimal performance in both accuracy and fairness metrics, regardless of image corruption severity, and to benchmark its performance against baseline SAM-based variants. Specifically, we introduce incremental levels of snow noise, ranging from mild (level 1) to severe (level 5), to the test datasets. For each noise level, we measure accuracy and *Corrupted Degradation Disparity* across all methods to assess the balance between robustness and fairness. Our results, as illustrated in Fig. 1, show that the proposed method FairSAM consistently outperforms all baseline methods in terms of fairness across all noise levels. FairSAM maintains the lowest bias at every noise level while maintaining comparable accuracy, indicating a better trade-off between fairness and accuracy across subgroups under varied corruption conditions and levels.

5.5 Out-of-distribution Generalization

To further assess the fairness-aware generalization capabilities of our method, we conduct an out-of-distribution experiment. This evaluation involves measuring the performance degradation difference between in-distribution and out-of-distribution test data for each demographic subgroup, using a method similar to the *Corrupted Degradation Disparity* metric. This approach allows us to estimate the model's robustness and its ability to maintain fairness across diverse data sets. As shown in Tab. 4 and Tab. 5, the proposed FairSAM consistently demonstrates the lowest bias among the methods evaluated. In contrast, GroupSAM, by disregarding the loss landscape flatness for the advantaged group, risks shifting this group into a disadvantaged position, potentially creating new imbalances.

6 Discussion

Our work reveals several important insights about the intersection of robustness and fairness in machine learning systems. First, traditional robustness methods improve overall performance under corruption but can inadvertently worsen fairness disparities across demographic subgroups. This gap is a blind spot in robustness research as aggregated metrics can mask inequities. Second, fairness-aware optimization during training improves equitable outcomes without large sacrifices in overall accuracy. FairSAM's balanced perturbation strategy shows that incorporating fairness into the optimization is feasible and useful for robust systems. Third, we observe that the benefits of fairness-aware robustness extend beyond imbalanced datasets

to balanced ones, indicating that demographic disparities in model performance under corruption are not solely attributable to class imbalance but reflect deeper algorithmic biases that require targeted intervention.

The implications of this work extend far beyond technical contributions to machine learning. As AI systems become increasingly deployed in high-stakes applications such as healthcare, criminal justice, and employment screening, ensuring that these systems maintain fairness under real-world conditions becomes paramount. This work addresses a gap in AI safety research that robustness and fairness are interconnected and must be addressed jointly. Robust models often exhibit disparate performance degradation across demographic groups. If unaddressed, this disparity can perpetuate or amplify societal inequalities.

Limitations and future work. FairSAM yields promising results, but several limitations remain. Our current evaluation focuses primarily on image classification tasks with binary sensitive attributes. Future work should explore the framework's effectiveness across multi-class sensitive attributes and other domains (Kong, 2022; Foulds et al., 2020; Pastor & Bonchi, 2024). Additionally, our fairness metrics, while comprehensive, represent one perspective on measuring equitable performance degradation. The development of alternative fairness criteria (Hardt et al., 2016; McNamara, 2019; Awasthi et al., 2020) and their integration into robust optimization frameworks remains an important area for future research.

7 Conclusion

In this work, we address the dual challenges of fairness and robustness in corrupted image classification. We introduce novel metrics for assessing unequal performance degradation in corrupted environments. We further develop FairSAM, a new framework that effectively couples robustness and fairness to ensure equitable performance across demographic subgroups. Our experimental results across multiple datasets and corruption conditions demonstrate that FairSAM consistently outperforms baseline methods in balancing the trade-off between fairness and performance. By maintaining both robust performance and fairness across subgroups, FairSAM represents a significant step forward in creating machine learning models that are resilient and fair in real-world applications. Future work will explore additional corruption scenarios and extend FairSAM as a framework for fair and robust image classification.

References

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, Proceedings of Machine Learning Research, pp. 1770–1780. PMLR, 2020. URL `http://proceedings.mlr.press/v108/awasthi20a.html`.

Anna Bair, Hongxu Yin, Maying Shen, Pavlo Molchanov, and Jose Alvarez. Adaptive sharpness-aware pruning for robust sparse networks. *arXiv preprint arXiv:2306.14306*, 2023.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, pp. 277–292, 2010.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pp. 13–18. IEEE, 2009.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pp. 1918–1921. IEEE, 2020. doi: 10.1109/ICDE48307.2020.00203.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3315–3323, 2016. URL `https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html`.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, February 2009. ISBN 978-1-4244-3313-1. doi: 10.1109/IC4.2009.4909197. URL `http://ieeexplore.ieee.org/document/4909197/`.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, (1):1–33, 2012.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pp. 869–874. IEEE, 2010.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, pp. 924–929. IEEE, 2012.

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*, pp. 1548–1558, 2021.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-Wild distribution shifts. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 5637–5664. PMLR, 2021. URL `http://proceedings.mlr.press/v139/koh21a.html`.

Youjin Kong. Are "Intersectionally Fair" AI algorithms really fair to women of color? A philosophical analysis. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pp. 485–494. ACM, 2022. doi: 10.1145/3531146.3533114.

Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 6781–6792. PMLR, 2021. URL `http://proceedings.mlr.press/v139/liu21f.html`.

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)*, December 2015.

Daniel McNamara. Equalized odds implies partially equalized outcomes under realistic assumptions. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor (eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 313–320. ACM, 2019. doi: 10.1145/3306618.3314290.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium*

on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 8930–8938. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17080.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel (eds.), *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp. 466–477. ACM, 2021. doi: 10.1145/3442188.3445910.

Eliana Pastor and Francesco Bonchi. Intersectional fair ranking via subgroup divergence. *Data Mining and Knowledge Discovery*, May 2024. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-024-01029-8. URL https://link.springer.com/10.1007/s10618-024-01029-8.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization, April 2020. URL http://arxiv.org/abs/1911.08731.

David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera (eds.), *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part I*, Lecture Notes in Computer Science, pp. 162–177. Springer, 2020. doi: 10.1007/978-3-030-67658-2_10.

Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, XiangLong Liu, and Aishan Liu. Improving Robust Fairness via Balance Adversarial Training, September 2022. URL http://arxiv.org/abs/2209.07534.

Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. Fairness Increases Adversarial Vulnerability, November 2022. URL http://arxiv.org/abs/2211.11835.

Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. Poisoning attacks on fair machine learning. In Arnab Bhattacharya, Janice Lee, Mong Li, Divyakant Agrawal, P. Krishna Reddy, Mukesh K. Mohania, Anirban Mondal, Vikram Goyal, and Rage Uday Kiran (eds.), *Database Systems for Advanced Applications - 27th International Conference, DASFAA 2022, Virtual Event, April 11-14, 2022, Proceedings, Part I*, Lecture Notes in Computer Science, pp. 370–386. Springer, 2022. doi: 10.1007/978-3-031-00123-9_30.

Indre Žliobaite. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017. doi: 10.1007/s10618-017-0506-1. URL https://doi.org/10.1007/s10618-017-0506-1.

Han Xu, Xiaorui Liu, Yaxin Li, Anil K. Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pp. 11492–11501. PMLR, 2021. URL http://proceedings.mlr.press/v139/xu21b.html.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017.

Chen Zhao, Kai Jiang, Xintao Wu, Haoliang Wang, Latifur Khan, Christan Grant, and Feng Chen. Algorithmic fairness generalization under covariate and dependence shifts simultaneously. In Ricardo Baeza-Yates and Francesco Bonchi (eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pp. 4419–4430. ACM, 2024. doi: 10.1145/3637528.3671909.

Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11345–11355, 2023.