Preserving AUC Fairness in Learning with Noisy Protected Groups

Mingyang Wu [*1] **Li Lin** [*1] **Wenbin Zhang** [2] **Xin Wang** [3] **Zhenhuan Yang** [4] **Shu Hu** [1]

Abstract

The Area Under the ROC Curve (AUC) is a key metric for classification, especially under class imbalance, with growing research focus on optimizing AUC over accuracy in applications like medical image analysis and deepfake detection. This leads to fairness in AUC optimization becoming crucial as biases can impact protected groups. While various fairness mitigation techniques exist, fairness considerations in AUC optimization remain in their early stages, with most research focusing on improving AUC fairness under the assumption of clean protected groups. However, these studies often *overlook* the impact of noisy protected groups, leading to fairness violations in practice. To address this, we propose the *first* robust AUC fairness approach under noisy protected groups with fairness theoretical guarantees using distributionally robust optimization. Extensive experiments on tabular and image datasets show that our method outperforms state-of-the-art approaches in preserving AUC fairness. The code is in https://github.com/Purdue-M2/AUC_Fairness_with_Noisy_Groups.

1. Introduction

The *Area Under the ROC Curve* (AUC) (Hanley & McNeil, 1982) is one of the most widely used performance metrics in classification tasks, particularly when addressing challenges such as class imbalance or uncertain relative costs of false positives and false negatives. It provides a measure of a classifier's ability to distinguish between classes across all possible decision thresholds, making it especially relevant in domains like information retrieval (Cortes & Mohri, 2003), medical image analysis (Yuan et al., 2021), and deepfake detection (Pu et al., 2022). In particular, in deepfake detection, misclassifying fake content as real can lead to the widespread dissemination of misinformation, potentially undermining public trust, manipulating discourse, or enabling fraud. AUC is thus preferred over fixed-threshold metrics, as it captures model performance comprehensively across varying operational settings.

This has motivated numerous studies (Yang, 2021; Kumagai et al.; Guo et al., 2022; Zhang et al., 2023) focusing on training AI models to maximize AUC rather than relying on traditional loss functions (*e.g.*, cross-entropy loss), as this approach directly optimizes a metric better aligned with the desired application outcomes. By doing so, models achieve improved performance in scenarios where distinguishing between classes with high sensitivity and specificity is essential.

The optimization of AUC in machine learning models necessitates a focus on fairness, particularly in light of growing concerns that algorithmic decisions often exacerbate inequities faced by vulnerable groups defined by sensitive attributes such as gender and race, also known as *protected groups*. Recent studies (Caton & Haas, 2024; Kenfack et al., 2024) have highlighted how machine learning models, if left unchecked, can perpetuate or worsen biases in allocation decisions, leading to unfavorable outcomes for these groups. To address this, a range of bias mitigation techniques (Hu & Chen, 2024; Lin et al., 2024; Tian et al., 2025; Kollias et al., 2024; Ju et al., 2024) has been developed, including methods that focus on statistical fairness metrics derived from confusion matrices. However, fairness considerations in AUC optimization (Yang et al., 2023; Yao et al., 2023) remain an early stage of exploration despite its significance in many applications and most of them focus on the *group-level* AUC fairness.

The group-level fairness for AUC leads to three categories of metrics, each addressing different aspects of disparate impacts. First, *intra-group* AUC (Beutel et al., 2019; Yao et al., 2023) focuses constraining both positive and negative examples to the same group. Second, *inter-group* AUC (Beutel et al., 2019; Kallus & Zhou, 2019; Yao et al., 2023) considers ranking fairness between groups, evaluating the

[*]Equal contribution [1]Department of Computer and Information Technology, Purdue University, West Lafayette, USA [2]Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, USA [3]Department of Epidemiology and Biostatistics, University at Albany, State University of New York, New York, USA [4]Amazon, New York, USA. Correspondence to: Shu Hu <hu968@purdue.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

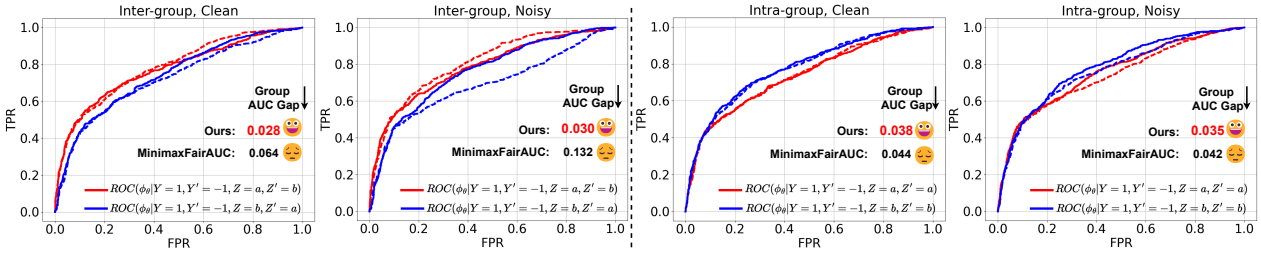

Figure 1. Illustrative inter-/intra-group AUC dicrepancy examples of existing MinimaxFairAUC method (Yang et al., 2023) (dashed curves) and our method (solid curves) on `Default` *(Yeh & Lien, 2009) dataset with noisy levels 0 and 0.3, respectively. Notations are defined in Section 4. In general, our method is better than MinimaxFairAUC in preserving AUC fairness, demonstrating robustness to noisy groups.*

metric by comparing positive examples from one group to negative examples from another. Lastly, a few works (Yang et al., 2023; Yao et al., 2023) have sought to considering both intra-group and inter-group AUC fairness during the learning process.

Nevertheless, existing AUC fairness studies often **overlook** the reliability of protected group information. This raises a crucial question: *Can AUC fairness notions be accurately measured or effectively enforced when the protected group data is noisy or unreliable?* Noisy protected group labels are prevalent in many scenarios. For instance, survey participants may intentionally obfuscate their responses due to concerns about privacy, fear of disclosure, or potential discrimination, leading to response biases (Krumpal, 2013). Similarly, in deepfake datasets, demographic annotations are often inferred using deep learning models (Lin et al., 2025). However, the accuracy of these annotations is inherently limited, as the true demographic information cannot be verified or tracked when the faces are AI-generated. Our practical evaluation (see Fig. 1) reveals that training with traditional AUC fairness under noisy protected group labels can result in significant group AUC gap in model deployment. *This highlights the critical need for designing a robust approach to ensure reliable AUC fairness.*

In this work, we propose the **first** robust AUC fairness approach for learning under noisy protected groups, providing theoretical fairness guarantees. We begin by conducting experiments to illustrate the adverse effects of noisy protected group labels on existing AUC fairness methods. Next, we introduce a novel and general AUC fairness metric that accounts for both intra-group and inter-group AUC. Based on this metric, we formulate a new learning objective for AUC fairness using a distributionally robust optimization (DRO) framework (Duchi & Namkoong, 2021), which bounds the Total Variation (TV) distance between clean and noisy group distributions. We also provide a theoretical analysis demonstrating that our approach ensures fairness even under noisy protected group labels. To estimate the TV distance bound, we reformulate it in terms of noisy label ratios and propose an empirical estimation method leveraging pre-trained multi-modal foundation models. Finally, we design an efficient stochastic gradient descent-ascent (SGDA) algorithm to optimize the proposed learning objective, enhancing both AUC fairness and the model's generalization capabilities in deployment scenarios. Our key contributions are:

1. We present the first experimental analysis of the impact of noisy protected group labels on the existing AUC fairness learning method.
2. We introduce a novel AUC fairness metric and propose the first approach to preserve AUC fairness under noisy protected groups with theoretical guarantees.
3. Extensive experiments on tabular and image datasets show that our method surpasses state-of-the-art approaches across applications like socioeconomic analysis and deepfake detection.

2. Related Work

AUC-based Fairness. A pioneering effort by Dixon et al. (2018) introduced the Pinned AUC metric for text classification tasks, which involves resampling the data so that each of the two groups constitutes half of the dataset, followed by calculating the AUC difference on the resampled data. Building on this foundation, Beutel et al. (2019) proposed intra-group and inter-group AUC metrics for recommender systems, assessing whether clicked items are ranked above unclicked items both within and across protected groups. Their method also incorporated a regularization term to reduce ranking unfairness. To address disparities across groups, Kallus & Zhou (2019) developed the cross-AUC (xAUC) metric, which identifies systematic biases where positive instances from one group may be ranked below negative instances from another.

In the context of general pairwise ranking, Narasimhan et al. (2020) proposed maximizing AUC under fairness constraints, further advancing the exploration of cross-group AUC fairness. Other works, such as Vogel et al. (2021),

focused on fairness defined directly in terms of the ROC curve, employing regularization to balance overall AUC performance with group-level fairness requirements. More recently, Yang et al. (2023) introduced a minimax fairness framework that simultaneously addresses intra-group and inter-group AUC disparities using a Rawlsian approach, supported by an efficient optimization algorithm with proven convergence guarantees. Similarly, Yao et al. (2023) proposed a scalable and efficient stochastic optimization framework for AUC-based fairness constraints, demonstrating its ability to balance classification performance and fairness in both online and offline learning scenarios. *However, all of the aforementioned works assume clean and accurately labeled protected groups, disregarding the impact of noisy group labels that are common in real-world datasets. Addressing this limitation is the central focus of our paper.*

Fairness with Noisy Protected Groups. Group fairness methods typically assume accurate knowledge of protected group labels, but in practice, these labels are often noisy or unreliable. Enforcing fairness constraints based on such noisy labels fails to guarantee fairness with respect to the clean labels (Gupta et al., 2018). To address this issue, Lahoti et al. (2020) proposed an adversarial reweighting approach that leverages correlations between non-protected features, task labels, and potentially unobserved group membership, demonstrating improved fairness under label uncertainty in tabular datasets. However, extending this approach to image data poses significant challenges.

Under the more conservative assumption of no protected group information, Hashimoto et al. (2018) applied distributionally robust optimization (DRO) to enforce what Lahoti et al. (2020) termed Rawlsian Max-Min fairness. Although DRO-based methods can achieve reasonable fairness results without explicit protected group labels, they are often less effective than approaches that incorporate such information. Building on Hashimoto et al. (2018), Wang et al. (2020) introduced a maximum total variation distance bound in the DRO procedure, offering the first fairness framework with guarantees under noisy protected-group labels. Further advancements include the work of Celis et al. (2021), who proposed an optimization framework with provable guarantees on both accuracy and fairness for classifiers trained with noisy protected attributes. Similarly, Mehrotra & Vishnoi (2022) introduced a novel method for improving fair rankings in the presence of noisy group labels. Additionally, Ghazimatin et al. (2022) and Ghosh et al. (2023) conducted empirical evaluations of fairness approaches under noisy sensitive information. *However, the aforementioned approaches are either not generalizable or are unsuitable for direct application to pairwise ranking optimization problems. These limitations leave AUC fairness largely unexplored in these contexts.*

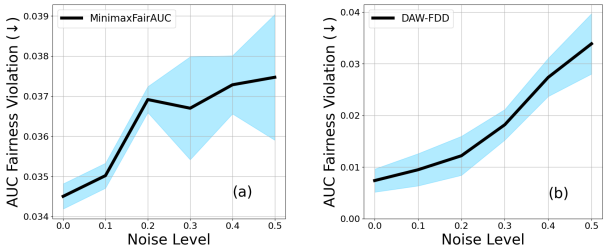

Figure 2. Impact of noisy protected group labels on AUC fairness violation (lower values indicate better AUC fairness) in two scenarios: (a) Socioeconomic Analysis and (b) Deepfake Detection. Mean value is shown in black line. The standard deviation is shown in blue background, where three random runs for each noise level.

3. Motivation

To demonstrate the impact of noisy protected group levels on AUC fairness, we conduct experiments on the tabular Adult dataset (for socioeconomic analysis) (Asuncion et al., 2007) and the image-based FF++ dataset (for deepfake detection) (Rossler et al., 2019; Lin et al., 2024). Noise is introduced by flipping a portion (ranging from 0 to 50%) of the protected group labels (*e.g.*, gender) in the training sets. For the Adult dataset, we evaluate the performance of the latest AUC fairness method, MinimaxFairAUC (Yang et al., 2023), while for the FF++ dataset, we assess the state-of-the-art fairness approach, DAW-FDD (Ju et al., 2024). All experiments follow the original settings specified in these methods. We use AUC fairness violation as the evaluation metric, which quantifies the maximum gap between group-level (intra-group or inter-group) AUC and the overall AUC. Then, we present the mean and standard deviation scores on the test sets over three random runs for each noise setting.

As shown in Fig. 2, AUC fairness violation increases significantly as the accuracy of protected group annotations declines across all three scenarios. Specifically, Fig. 2(a) demonstrates that conventional AUC fairness enhancement method is highly sensitive to noisy protected group labels in the training data. Similarly, while DAW-FDD is designed to improve general fairness, Fig. 2(b) reveals it struggles in maintaining AUC fairness under noisy group conditions. These findings underscore the critical need for robust AUC fairness approaches capable of tolerating noisy groups.

4. Methodology

In this section, we introduce a novel robust AUC fairness approach to address the challenges discussed in the previous section. Let $X \in \mathcal{X} \subseteq \mathbb{R}^d$ represent the random variable for input features, $Y \in \mathcal{Y} = \{\pm 1\}$ denote the binary label, and $Z \in \mathcal{Z} = \{1, \ldots, m\}$ represent the random protected group, where m is the total number of protected groups. We define a scoring function $\phi_\theta : \mathcal{X} \to \mathbb{R}$, parameterized by

$\theta \in \Theta$, which maps the input features to a real-valued score.

4.1. AUC Fairness

Overall AUC: The overall AUC quantifies the probability that a model correctly ranks a positive example (*e.g.*, X) higher than a negative example (*e.g.*, X'). It is formally defined as:

$$AUC(\theta) = \mathbb{E}[\mathbb{I}_{[\phi_\theta(X)>\phi_\theta(X')]} \mid Y = 1, Y' = -1], \quad (1)$$

where $\mathbb{I}_{[a]}$ is the indicator function, which equals 1 if a is true and 0 otherwise. To optimize the AUC score, it is common practice to minimize the complementary AUC risk (Hanley & McNeil, 1982), given by:

$$\begin{aligned} L_{AUC}(\theta) &= 1 - AUC(\theta) \\ &= \mathbb{E}[\mathbb{I}_{[\phi_\theta(X)\leq\phi_\theta(X')]} \mid Y = 1, Y' = -1]. \end{aligned} \quad (2)$$

Group-level AUC. Following (Yang et al., 2023), the group-level AUC is defined as:

$$\begin{aligned} &AUC_{z,z'}(\theta) \\ &= \mathbb{E}[\mathbb{I}_{[\phi_\theta(X)>\phi_\theta(X')]} | Y = 1, Y' = -1, Z = z, Z' = z']. \end{aligned} \quad (3)$$

This metric captures the AUC score for comparisons between positive examples from group z and negative examples from group z', reflecting a pairwise dependence with respect to the groups $Z, Z' \in \mathcal{Z}$. When $z = z'$, we refer to it as *intra-group* AUC, which measures ranking performance within a single group. Conversely, when $z \neq z'$, it is termed *inter-group* AUC, assessing the ranking performance across different groups.

AUC Fairness Metric. Then, we propose the target AUC fairness metric as follows:

$$\begin{aligned} h(\theta) =& \mathbb{I}_{[\phi_\theta(X)>\phi_\theta(X')]}\mathbb{I}_{[Y=1]}\mathbb{I}_{[Y'=-1]} \\ &- \mathbb{E}[\mathbb{I}_{[\phi_\theta(X)>\phi_\theta(X')]}\mathbb{I}_{[Y=1]}\mathbb{I}_{[Y'=-1]}]. \end{aligned} \quad (4)$$

Building on this, we define the group-level AUC fairness functions as:

$$\begin{aligned} g_{z,z'}(\theta) &= \mathbb{E}[h(\theta)|Z = z, Z' = z'] \\ &= AUC_{z,z'}(\theta) - AUC(\theta), \ \ \forall z, z' \in \mathcal{Z}. \end{aligned} \quad (5)$$

The above formulation quantifies the gap ($|g_{z,z'}(\theta)|$) between any group-level AUC score and the overall AUC score. To achieve AUC fairness, it suffices to enforce the constraint $g_{z,z'}(\theta) \leq 0 \ \forall z, z' \in \mathcal{Z}$, ensuring that all group-level AUCs are close to the overall AUC. This, in turn, reduces disparities among group-level AUCs, fostering a more equitable model performance across groups. This formulation also establishes meaningful connections with several existing fairness measures. For instance, when $z = z'$, it generalizes to intra-group pairwise fairness, as studied in Beutel et al. (2019); Yao et al. (2023). Conversely, when $z \neq z'$, it aligns with inter-group pairwise fairness, as explored in Beutel et al. (2019); Kallus & Zhou (2019); Yao et al. (2023). Furthermore, our AUC fairness function is closely related to the Rawlsian principle of justice (Rawls, 2001), particularly the Rawlsian AUC fairness framework proposed by Yang et al. (2023). Specifically, if we focus solely on maximizing the smallest group-level AUC among all groups, without considering the second term (*i.e.*, the overall AUC) in Eq. (5).

However, solely emphasizing and enforcing the fairness constraint can lead to a trivial solution where $AUC_{z,z'}(\theta) = AUC(\theta) = 0.5$, which reflects no discriminatory power in the model. To prevent this, it is necessary to simultaneously maximize the overall AUC score while ensuring AUC fairness. It is worth noting that AUC is not only a fairness-related metric but also a key performance measure for evaluating trained models. Consequently, directly optimizing AUC can enhance model performance, as demonstrated in various domains, including medical image analysis (Yuan et al., 2021) and deepfake detection (Pu et al., 2022). Instead of maximizing the overall AUC, we minimize its corresponding AUC risk directly. This leads us to formulate the following constrained AUC fairness problem:

$$\min_\theta L_{AUC}(\theta), \text{ s.t. } g_{z,z'}(\theta) \leq 0, \forall z, z' \in \mathcal{Z}. \quad (6)$$

4.2. Robust AUC Fairness

While minimizing Eq. (6) can achieve AUC fairness under clean protected groups, it does not guarantee fairness when the protected groups are noisy. To address this, we propose a robust AUC fairness learning objective by leveraging a distributionally robust optimization (DRO) approach (Duchi & Namkoong, 2021), inspired by Wang et al. (2020).

Formulation. Specifically, let $\widehat{Z} \in \mathcal{Z}$ be the random variable representing the noisy protected group associated with X. The learning objective can then be reformulated as:

$$\min_\theta L_{AUC}(\theta), \text{ s.t. } \widehat{g}_{z,z'}(\theta) \leq 0, \forall z, z' \in \mathcal{Z}, \quad (7)$$

where $\widehat{g}_{z,z'}(\theta) = \mathbb{E}[h(\theta)|\widehat{Z} = z, \widehat{Z}' = z']$. Next, we analyze how far a model trained with noisy protected group labels using Eq. (7) deviates from satisfying the fairness constraints defined for clean protected groups. Let p represent the distribution of pairwise data $((X, Y), (X', Y')) \sim p$, where $Y = 1$ and $Y' = -1$. Define $p_{z,z'}$ as the distribution of pairwise data conditioned on the clean groups $Z = z$ and $Z' = z'$, such that $((X, Y), (X', Y')) \mid (Z = z, Z' = z') \sim p_{z,z'}$. Similarly, let $\widehat{p}_{z,z'}$ represent the distribution of $((X, Y), (X', Y'))$ conditioned on the noisy groups $\widehat{Z} = z$ and $\widehat{Z}' = z'$, such that $((X, Y), (X', Y')) \mid (\widehat{Z} = z, \widehat{Z}' = z') \sim \widehat{p}_{z,z'}$. To quantify the difference between $p_{z,z'}$ and

$\widehat{p}_{z,z'}$, we use the Total Variation (TV) distance (Duchi & Namkoong, 2021), denoted as $TV(p_{z,z'}, \widehat{p}_{z,z'})$. Based on this measure, we establish the following theoretical results (the proof is provided in Appendix A.1).

Theorem 4.1. *Suppose a model is trained using Eq. (7) with noisy groups and satisfies $\widehat{g}_{z,z'}(\theta) \leq 0 \, \forall z, z' \in \mathcal{Z}$. Let $\gamma_{z,z'}$ be an upper bound on the TV distance, such that $\gamma_{z,z'} \geq TV(p_{z,z'}, \widehat{p}_{z,z'}) \, \forall z, z' \in \mathcal{Z}$. Then, the fairness measure for the clean groups will be satisfied within a slack of $\gamma_{z,z'}$ for each pairwise group, ensuring $g_{z,z'}(\theta) \leq \gamma_{z,z'} \, \forall z, z' \in \mathcal{Z}$.*

Relaxation. While Theorem 4.1 provides an upper bound for AUC fairness on the clean groups, our goal is to ensure that $g_{z,z'}(\theta) \leq 0, \forall z, z' \in \mathcal{Z}$. To achieve this, inspired by Wang et al. (2020), we employ a DRO approach. Specifically, any feasible solution to the following constrained optimization problem is guaranteed to satisfy the fairness constraints for the clean groups:

$$\min_{\theta} L_{AUC}(\theta), \quad \text{s.t.} \max_{\substack{\tilde{p}_{z,z'}: TV(\tilde{p}_{z,z'}, \widehat{p}_{z,z'}) \leq \gamma_{z,z'} \\ \tilde{p}_{z,z'} \ll p_{z,z'}}} \tilde{g}_{z,z'}(\theta) \leq 0, \forall z, z' \in \mathcal{Z}, \quad (8)$$

where $\tilde{g}_{z,z'}(\theta) = \mathbb{E}_{((X,Y),(X',Y')) \sim \tilde{p}_{z,z'}}[h(\theta)]$ and $\tilde{p}_{z,z'} \ll p_{z,z'}$ indicates that $\tilde{p}_{z,z'}$ is absolutely continuous w.r.t. $p_{z,z'}$.

Reformulation. To simplify the constrained optimization problem in Eq. (8) and demonstrate its practical application when the true distributions are unknown, we reformulate it as a minimax problem using a Lagrangian formulation. Additionally, we replace all expectations with those over the empirical distribution derived from a dataset $\mathcal{S} := \{(X_i, Y_i, \widehat{Z}_i)\}_{i=1}^n$ of n samples. For convenience, we denote X_i as X_i^+ or X_i^- if it has a positive or negative label, respectively. Let $n^+ = |\{X_i^+ \mid i \in [n]\}|$ and $n^- = |\{X_i^- \mid i \in [n]\}|$ represent the total number of positive and negative samples, respectively, where $[n] = \{1, \ldots, n\}$. Similarly, we denote X_i as X_i^{z+} or X_i^{z-} if it belongs to group z with a positive or negative label, respectively. Let n^{z+} and n^{z-} represent the total number of positive and negative samples from group z, respectively. In practice, we let the empirical distribution $\widehat{p}_{z,z'} \in \mathbb{R}^{n^+ \times n^-}$ be a matrix where the (i,j)-th entry is given by: $\widehat{p}_{z,z'}^{i,j} = \frac{1}{n^{z+} n^{z'-}}$ if the i-th positive sample belongs to group z and the j-th negative sample belongs to group z'; otherwise, it is set to 0.

The TV distance constraint can then be reformulated to identify an empirical distribution $\tilde{p}_{z,z'} \in \mathbb{R}^{n^+ \times n^-}$ within a ball defined by: $\mathbb{B}_{\gamma_{z,z'}}(\widehat{p}_{z,z'}) := \{\tilde{p}_{z,z'} : \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} |\tilde{p}_{z,z'}^{i,j} - \widehat{p}_{z,z'}^{i,j}| \leq 2\gamma_{z,z'}, \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \tilde{p}_{z,z'}^{i,j} = 1, \tilde{p}_{z,z'}^{i,j} \geq 0 \; \forall i \in [n^+], j \in [n^-]\}$. Denote $\lambda_{z,z'}$ as the Lagrangian multiplier associated with the pair $z, z' \in \mathcal{Z}$. Then, the empirical version of Eq. (8) can be rewritten as:

$$\min_{\theta} \max_{\lambda_{z,z'} \geq 0, \tilde{p}_{z,z'}^{i,j} \geq 0} \overline{L}_{AUC}(\theta) + \sum_{z=1}^{m} \sum_{z'=1}^{m} \lambda_{z,z'} \overline{g}_{z,z'}(\theta), \quad \text{s.t.} \; \|\tilde{p}_{z,z'} - \widehat{p}_{z,z'}\|_{1,1} \leq 2\gamma_{z,z'}, \|\tilde{p}_{z,z'}\|_{1,1} = 1, \forall z, z' \in \mathcal{Z}, \quad (9)$$

where $\overline{L}_{AUC}(\theta) = \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \mathbb{I}_{[\phi_\theta(X_i^+) \leq \phi_\theta(X_j^-)]}$ is the empirical form of $L_{AUC}(\theta)$, $\overline{g}_{z,z'}(\theta) = \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \tilde{p}_{z,z'}^{i,j} \mathbb{I}_{[\phi_\theta(X_i^+) > \phi_\theta(X_j^-)]} - \frac{1}{n^+ n^-} \sum_{i=1}^{n^+} \sum_{j=1}^{n^-} \mathbb{I}_{[\phi_\theta(X_i^+) > \phi_\theta(X_j^-)]}$ is the empirical form of $\tilde{g}_{z,z'}(\theta)$, and $\|\cdot\|_{1,1}$ is the $L_{1,1}$-norm. In practice, the non-differentiable indicator function $\mathbb{I}$ can be replaced with a (sub)-differentiable and non-increasing surrogate loss function ℓ. For instance, in our experiments, we replace $\mathbb{I}_{[a \leq 0]}$ with the logistic loss $\log(1 + \exp(-a))$.

4.3. Noisy Ratio Estimation

Theoretical Estimation. The learning objective in Eq. (9) relies on the upper bound $\gamma_{z,z'}$ of the TV distance between $p_{z,z'}$ and $\widehat{p}_{z,z'}$, as demonstrated in Theorem 4.1. However, in practice, $p_{z,z'}$ is typically unknown. To address this, we present a theoretical result to facilitate the estimation of $\gamma_{z,z'}$, as detailed below (proof provided in Appendix A.2):

Lemma 4.2. *Given a positive-negative pair group (z, z'), suppose the prior pairwise clean group probability $Pr[(z, z')]$ unaffected by the noise,* i.e.*, $Pr[(Z = z, Z' = z')] = Pr[(\widehat{Z} = z, \widehat{Z}' = z')]$. Then $TV(p_{z,z'}, \widehat{p}_{z,z'}) \leq Pr[(Z, Z') \neq (\widehat{Z}, \widehat{Z}') | (\widehat{Z} = z, \widehat{Z}' = z')]$.*

According to the above Lemma, the estimation of $\gamma_{z,z'}$ can be reduced to estimating the probability $Pr[(Z, Z') \neq (\widehat{Z}, \widehat{Z}') \mid (Z = z, Z' = z')]$. In robust machine learning, various methods can be employed for this estimation. For example, an auxiliary network can be trained from scratch to estimate this probability (Jiang et al., 2018; Yu et al., 2019), or an auxiliary clean dataset can be utilized for estimation (Kallus et al., 2022; Wang et al., 2020). However, training an auxiliary network introduces additional complexity to the target model, potentially reducing its generalizability across diverse scenarios. Furthermore, many existing approaches rely on a single data modality (e.g., images) for estimation, which can limit accuracy and effectiveness. Additionally, obtaining auxiliary datasets that are well-suited to the target problem can be challenging in practice.

Empirical Estimation. To address these challenges, we leverage the capabilities of pre-trained multi-modal foundation models (Li et al., 2024; Gardner et al., 2024) to construct a noisy label detector, drawing inspiration from recent works (Hu et al., 2023; Wei et al., 2024), without requiring additional training. While we use image data as an illustrative example, our method can be easily adapted to other data

modalities, such as tabular data, which we leave as future work. Specifically, for each image, we utilize its protected group label to design a pair of label-specific prompts: a positive prompt ($\mathcal{P}$) and a negative prompt ($\mathcal{N}$). For example, $\mathcal{P}$ is designed as "a photo of a {group_name}" and $\mathcal{N}$ as "a photo without a {group_name}," where {group_name} corresponds to the protected group label ($\widehat{Z}$) of the image. These prompts are fed into the text encoder of the CLIP model (Radford et al., 2021) to generate text feature representations $T^{\mathcal{P}}$ and $T^{\mathcal{N}}$. Simultaneously, the image is processed through CLIP's visual encoder to extract its visual feature representation $\mathcal{V}$. We then compute the cosine similarity between the visual features and the text features, resulting in $\texttt{sim}(\mathcal{V}, T^{\mathcal{P}})$ and $\texttt{sim}(\mathcal{V}, T^{\mathcal{N}})$. Finally, we regard the group label of the test image as clean if $\texttt{sim}(\mathcal{V}, T^{\mathcal{P}}) > \texttt{sim}(\mathcal{V}, T^{\mathcal{N}})$, and noisy otherwise.

We stress that CLIP is not used for relabeling, classification, or decision-making. Instead, its strong representational power is utilized to estimate the mismatch rate between protected group labels and corresponding semantic features. This design avoids additional training (e.g., adding MLP heads), thereby maintaining the theoretical guarantees of our fairness framework. Since CLIP predictions are not perfectly reliable, using them for classification would undermine the provable robustness that our method offers. Based on this estimation, we approximate $\gamma_{z,z'}$ as follows:

$$
\begin{aligned}
&Pr[(Z,Z') \neq (\widehat{Z},\widehat{Z}')|(\widehat{Z}=z,\widehat{Z}'=z')] \approx \\
&\Bigg(\Big|\Big\{(i,j)\Big|^{\widehat{Z}_i=z,\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{P}})>\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{N}}),}_{\widehat{Z}'_j=z',\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{P}})\leq\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{N}})}\Big\}\Big| \\
&+\Big|\Big\{(i,j)\Big|^{\widehat{Z}_i=z,\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{P}})\leq\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{N}}),}_{\widehat{Z}'_j=z',\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{P}})>\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{N}})}\Big\}\Big| \\
&+\Big|\Big\{(i,j)\Big|^{\widehat{Z}_i=z,\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{P}})\leq\texttt{sim}(\mathcal{V}_i^+,T_i^{+\mathcal{N}}),}_{\widehat{Z}'_j=z',\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{P}})\leq\texttt{sim}(\mathcal{V}_j^-,T_j^{-\mathcal{N}})}\Big\}\Big|\Bigg) \\
&/|\{(i,j)|\widehat{Z}_i=z,\widehat{Z}'_j=z'\}|.
\end{aligned}
\tag{10}
$$

Here, $i \in [n^+]$ and $j \in [n^-]$, with $\mathcal{V}_i^+$ and $\mathcal{V}_j^-$ as visual feature representations, and $\mathcal{T}_i^+$ and $\mathcal{T}_j^-$ as text feature representations for positive and negative samples, respectively.

4.4. Optimization

Finally, we develop a stochastic gradient descent-ascent (SGDA) method to solve the minimax optimization problem in Eq. (9). To avoid the model becoming stuck in sharp and narrow minima during training, we incorporate the sharpness-aware minimization (SAM) technique (Foret et al., 2020) to flatten the loss landscape. This flattening is achieved by finding the optimal perturbation ϵ^* to the model parameters θ, which maximizes the loss. The process is:

$$
\begin{aligned}
\epsilon^* &= \arg\max_{\|\epsilon\|_2\leq\nu} \underbrace{\Big(\overline{L}_{AUC} + \sum_{z=1}^{m}\sum_{z'=1}^{m}\lambda_{z,z'}\overline{g}_{z,z'}\Big)}_{\mathcal{L}}(\theta+\epsilon) \\
&\approx \arg\max_{\|\epsilon\|_2\leq\nu} \epsilon^\top\nabla_\theta\mathcal{L} = \nu\frac{\nabla_\theta\mathcal{L}}{||\nabla_\theta\mathcal{L}||_2},
\end{aligned}
\tag{11}
$$

Here, ν controls the perturbation magnitude, and the approximation is derived using a first-order Taylor expansion, assuming ϵ is small. The final equation is obtained by solving a dual norm problem, where sign represents the sign function, and $\nabla_\theta\mathcal{L}$ is the gradient of $\mathcal{L}$ with respect to θ. Consequently, the model weights are updated by solving the following optimization problem: $\min_\theta \mathcal{L}(\theta+\epsilon^*)$. The intuition is that perturbing the model parameters in the direction of the gradient norm maximizes the loss value, which in turn encourages the model to explore a flatter loss landscape, enhancing its generalizability.

To ensure each mini-batch contains both positive and negative samples from all possible groups, we follow Yang et al. (2023) and design a sampling operator (*i.e.*, $\texttt{Sampler}(\cdot,\cdot)$) that randomly selects a mini-batch of size b by stratifying the training set $\mathcal{S}$ based on the label and group attribute. The details of the sampling operator are outlined in Appendix B.

The overall optimization procedure is as follows: we first initialize the model parameters θ, $\lambda_{z,z'}$, and $\tilde{p}_{z,z'}$, and use the approach developed in the previous section with Eq. (10) to estimate $\gamma_{z,z'}$ for all possible positive-negative group

Algorithm 1 Robust AUC Fairness

Input: A training dataset $\mathcal{S}$ of size n, number of iterations T, batch size b, learning rates η_θ; η_λ; η_p, and $\gamma_{z,z'}$ estimated by Eq. (10)
Initialize: $\theta^{(1)}$, $\lambda_{z,z'}^{(1)}$, $\tilde{p}_{z,z'}^{(1)}$ for all pairs (z,z')
for $t=1$ to T **do**
  $B = \texttt{Sampler}(\mathcal{S}, b)$
  Compute ϵ^* based on Eq. (11)
  Compute perturbed parameters: $\overline{\theta}^{(t)} = \theta^{(t)} + \epsilon^*$
  Update θ: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_\theta\nabla_\theta\mathcal{L}|_{\overline{\theta}^{(t)}}$
  for each $(z,z') \in \{(1,1),(1,2),\ldots,(m,m)\}$ **do**
    Update $\lambda_{z,z'}$: $\lambda_{z,z'}^{(t+1)} \leftarrow \lambda_{z,z'}^{(t)} + \eta_\lambda\overline{g}_{z,z'}(\overline{\theta}^{(t)})$
    Update $\tilde{p}_{z,z'}$: $\tilde{p}_{z,z'}^{(t+1)} \leftarrow \tilde{p}_{z,z'}^{(t)} + \eta_p\lambda_{z,z'}^{(t)}\nabla_{\tilde{p}_{z,z'}}\overline{g}_{z,z'}(\overline{\theta}^{(t)})$
    Project $\tilde{p}_{z,z'}^{(t+1)}$ onto $\ell_{1,1}$-norm constraints: $\|\tilde{p}_{z,z'}^{(t+1)} - \widehat{p}_{z,z'}\|_{1,1} \leq 2\gamma_{z,z'}$, $\|\tilde{p}_{z,z'}^{(t+1)}\|_{1,1} = 1$
  end for
end for
return $\theta^{(t^*)}$, where t^* is the best iterate satisfying the constraints in Eq. (9) with the lowest objective.

pairs (z, z'). Next, we randomly select a mini-batch B using our `Sampler` operator and perform the following steps for each iteration on B (For more details, refer to Algorithm 1):

- Compute ϵ^* based on Eq. (11).
- Update θ based on the gradient descent: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}|_{\theta+\epsilon^*}$, where η_θ is the learning rate.
- For each $(z, z') \in \{(1,1),(1,2),...,(m,m)\}$, use gradient ascent to update $\lambda_{z,z'}$: $\lambda_{z,z'} \leftarrow \lambda_{z,z'} + \eta_\lambda \overline{g}_{z,z'}(\theta + \epsilon^*)$ and $\tilde{p}_{z,z'}$: $\tilde{p}_{z,z'} \leftarrow \tilde{p}_{z,z'} + \eta_p \lambda_{z,z'} \nabla_{\tilde{p}_{z,z'}} \overline{g}_{z,z'}(\theta + \epsilon^*)$, where η_λ and η_p are learning rates. Next, we project $\tilde{p}_{z,z'}$ onto $\ell_{1,1}$-norm constraints: $\|\tilde{p}_{z,z'} - \widehat{p}_{z,z'}\|_{1,1} \leq 2\gamma_{z,z'}$, $\|\tilde{p}_{z,z'}\|_{1,1} = 1$.

The project can be done efficiently with Duchi et al. (2008).

5. Experiments

5.1. Settings

Datasets. In experiments, we train models with different methods on both tabular and image datasets. For tabular data, we conduct socioeconomic analysis on three widely used datasets in fair machine learning research (Donini et al., 2018): Adult (protected attribute: gender), Bank (protected attribute: age), and Default (protected attribute: gender). Each dataset is randomly split into training, validation, and test sets in a 60%/20%/20% ratio. For image data, we focus on the deepfake detection task using datasets from Lin et al. (2024). Specifically, we train models on the FF++ (Rossler et al., 2019) training set (protected attribute: gender) and evaluate them on the test sets of FF++, DFDC (dee), DFD (Google & Jigsaw, 2019), and Celeb-DF (Li et al., 2020). Further details are provided in Appendix C.

Evaluation Metrics. For utility, we use overall AUC as the primary model performance metric. For fairness, we measure AUC fairness violation ('Violation', lower is better), defined as the maximum absolute difference between any group-level AUC score the overall AUC score. Additionally, following Yang et al. (2023), we use the Min/Max fairness metric (higher is better), defined as the ratio of the minimum to the maximum group-level AUC score. Their formulations can be found in Appendix D.

Baselines. For socioeconomic analysis, we compare our method against AUCMax (without fairness constraint) (Yang et al., 2023), InterFairAUC (Vogel et al., 2021), and MinimaxFairAUC (Yang et al., 2023). For deepfake detection, we compare our method with the latest fairness methods, including DAG-FDD, DAW-FDD (Ju et al., 2024), and PG-FDD (Lin et al., 2024). The comparison also includes 'Original' (a backbone with cross-entropy loss). More details can be found in Appendix E.

Implementation Details. All experiments are implemented in PyTorch and trained on an NVIDIA RTX A6000. For training, we set the batch size to 10,000 for socioeconomic analysis and 32 for deepfake detection, with 1,000 and 100 training epochs, respectively. We use the SGD optimizer. For socioeconomic analysis, we use 3-layer multilayer perceptron (MLP) as the model. γ is selected from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. For deepfake detection, we use noisy group labels, which are common in datasets like FF++ where demographic attributes are inferred. Evaluating fairness under label noise is practical and follows prior work (Celis et al., 2021; Mehrotra & Vishnoi, 2022). We use Xception (Chollet, 2017) and EfficientNet-B4 (Tan & Le, 2019) as the detector backbones. $\gamma = 0.02$ is estimated using Eq. (10). See Appendix F.1 for details.

5.2. Results

Performance on Tabluar Data. For each of the three tabular datasets, we introduce noise into the protected group labels by randomly selecting a fraction γ of data points and flipping their labels to another group. We evaluate the performance of each method under different noise levels, conducting three random runs per dataset and reporting the mean and standard deviation in Table 1. It is clear that our approach consistently achieves the lowest AUC fairness violation and the highest Min/Max AUC score across all datasets and noise levels, demonstrating its effectiveness in preserving fairness under noisy protected group labels. For instance, at a noise level of 0.5 on the Adult dataset, our method achieves a fairness violation of 0.0316, significantly lower than AUCMax (0.0766), InterFairAUC (0.0374), and MinimaxFairAUC (0.0375). While AUCMax attains a higher overall AUC due to the absence of fairness constraints, its fairness violation remains substantially higher. Similarly, at a noise level of 0.1 on the Default dataset, our method achieves 0.0187 fairness violation, reducing it by 5.14% compared to AUCMax. Moreover, as the noise level increases, baseline methods exhibit worsening fairness violations, whereas our approach maintains superior fairness performance, underscoring its robustness in handling noisy protected group labels.

Performance on Image Data. Deepfake detection datasets are inherently noisy due to inaccuracies in demographic annotations, as the protected groups of generated faces cannot be verified in practice. Thus, our proposed noisy estimation method is well-suited for estimating the noise ratio, which we determine to be 0.02 in our experiments. We fix this value for subsequent experiments and report the results in Table 2. From the table, we find our method achieves the lowest group AUC fairness violation across all datasets, demonstrating superior fairness preservation under noisy protected groups. For instance, on FF++ (Xception backbone), our method achieves a fairness violation of

Noise Level	Method	Adult			Bank			Default		
		AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑
0.1	AUCMax	**0.9159±0.0001**	0.0787±0.0018	0.9115±0.0012	**0.9288±0.0003**	0.1565±0.0016	0.8559±0.0014	**0.7762±0.0008**	0.0701±0.0009	0.9076±0.0011
	InterFairAUC	0.9031±0.0002	0.0380±0.0013	0.9479±0.0007	0.9068±0.0005	0.0975±0.0025	0.9048±0.0022	0.7399±0.0031	0.0242±0.0003	0.9649±0.0009
	MinimaxFairAUC	0.9059±0.0002	0.0350±0.0003	0.9513±0.0006	0.9148±0.0006	0.1246±0.0035	0.8795±0.0034	0.7536±0.0037	0.0302±0.0080	0.9520±0.0098
	Ours	0.9060±0.0002	**0.0332±0.0007**	**0.9536±0.0005**	0.9039±0.0007	**0.0876±0.0033**	**0.9143±0.0029**	0.7645±0.0017	**0.0187±0.0012**	**0.9691±0.0019**
0.2	AUCMax	**0.9112±0.0002**	0.0885±0.0011	0.9051±0.0008	**0.9264±0.0003**	0.1592±0.0026	0.8543±0.0022	**0.7890±0.0005**	0.0565±0.0011	0.9254±0.0013
	InterFairAUC	0.9035±0.0004	0.0377±0.0003	0.9477±0.0005	0.9037±0.0003	0.1071±0.0028	0.8960±0.0024	0.7527±0.0037	0.0256±0.0026	0.9560±0.0062
	MinimaxFairAUC	0.9045±0.0003	0.0369±0.0003	0.9489±0.0005	0.9142±0.0006	0.1226±0.0046	0.8814±0.0044	0.7526±0.0039	0.0267±0.0057	0.9558±0.0093
	Ours	0.9039±0.0004	**0.0328±0.0008**	**0.9539±0.0011**	0.9163±0.0005	**0.1053±0.0026**	**0.8988±0.0024**	0.7624±0.0016	**0.0206±0.0014**	**0.9660±0.0022**
0.3	AUCMax	**0.9150±0.0001**	0.0800±0.0005	0.9117±0.0004	**0.9293±0.0005**	0.1565±0.0026	0.8559±0.0023	**0.7768±0.0005**	0.0679±0.0013	0.9100±0.0015
	InterFairAUC	0.8990±0.0008	0.0379±0.0012	0.9482±0.0032	0.9078±0.0005	0.0994±0.0020	0.9031±0.0018	0.7405±0.0027	0.0217±0.0009	0.9684±0.0019
	MinimaxFairAUC	0.8977±0.0006	0.0367±0.0013	0.9490±0.0025	0.9142±0.0006	0.1207±0.0050	0.8833±0.0049	0.7533±0.0036	0.0290±0.0077	0.9526±0.0097
	Ours	0.9059±0.0005	**0.0356±0.0006**	**0.9513±0.0007**	0.9093±0.0004	**0.0949±0.0026**	**0.9080±0.0024**	0.7687±0.0011	**0.0129±0.0011**	**0.9785±0.0014**
0.4	AUCMax	**0.9131±0.0002**	0.0783±0.0007	0.9123±0.0006	**0.9275±0.0005**	0.1655±0.0025	0.8465±0.0023	**0.7795±0.0005**	0.0717±0.0007	0.9057±0.0008
	InterFairAUC	0.8987±0.0008	0.0367±0.0022	0.9488±0.0042	0.9060±0.0005	0.0944±0.0030	0.9078±0.0029	0.7524±0.0035	0.0256±0.0039	0.9559±0.0068
	MinimaxFairAUC	0.9056±0.0002	0.0373±0.0007	0.9492±0.0003	0.9146±0.0005	0.1246±0.0036	0.8795±0.0035	0.7538±0.0036	0.0301±0.0080	0.9520±0.0098
	Ours	0.9008±0.0002	**0.0341±0.0005**	**0.9530±0.0004**	0.9054±0.0004	**0.0937±0.0034**	**0.9086±0.0028**	0.7552±0.0024	**0.0246±0.0036**	**0.9580±0.0066**
0.5	AUCMax	**0.9127±0.0003**	0.0766±0.0009	0.9141±0.0008	**0.9290±0.0003**	0.1572±0.0017	0.8553±0.0014	**0.7767±0.0007**	0.0715±0.0015	0.9061±0.0017
	InterFairAUC	0.9012±0.0012	0.0374±0.0013	0.9482±0.0026	0.9067±0.0005	0.0947±0.0017	0.9076±0.0016	0.7488±0.0037	0.0278±0.0026	0.9562±0.0047
	MinimaxFairAUC	0.9012±0.0008	0.0375±0.0016	0.9482±0.0030	0.9145±0.0006	0.1229±0.0042	0.8811±0.0046	0.7526±0.0037	0.0271±0.0069	0.9553±0.0098
	Ours	0.8984±0.0006	**0.0316±0.0002**	**0.9554±0.0004**	0.9044±0.0004	**0.0876±0.0020**	**0.9145±0.0015**	0.7447±0.0047	**0.0243±0.0014**	**0.9571±0.0010**

*Table 1. Performance comparison across different noise levels (0.1–0.5). The numbers are reported as 'Mean ± Standard Deviation.' ↑ means higher is better and ↓ means lower is better. The best results are shown in **Bold**.*

Backbone	Method	FF++			DFDC			DFD			Celeb-DF		
		AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑
Xception	Original	0.9384	0.0303	0.9652	0.5953	0.0319	0.9492	0.7574	0.0326	0.9529	0.6660	0.1479	0.8101
	DAG-FDD	0.9628	0.0147	0.9776	0.6058	0.0229	0.9608	0.7770	0.0216	0.9633	0.7059	0.1730	0.7906
	DAW-FDD	0.9650	0.0288	0.9702	0.6037	0.0201	0.9684	0.7825	0.0312	0.9479	0.7092	0.1818	0.7792
	PG-FDD	**0.9708**	0.0111	0.9714	**0.6207**	0.0184	0.9594	**0.8025**	0.0113	0.9846	**0.7214**	0.1412	0.8350
	Ours	0.9644	**0.0090**	**0.9857**	0.6086	**0.0048**	**0.9930**	0.7847	**0.0069**	**0.9881**	0.7108	**0.0729**	**0.9117**
EfficientNet -B4	Original	0.9332	0.0209	0.9684	0.5982	0.0306	0.9320	0.7593	0.0445	0.9382	0.6692	0.2453	0.6962
	DAG-FDD	0.9563	0.0100	0.9869	0.6030	0.0372	0.9210	0.7706	0.0216	0.9626	0.7102	0.2196	0.7415
	DAW-FDD	0.9694	0.0169	0.9764	0.5941	0.0254	0.9449	0.7756	0.0380	0.9440	0.7345	0.2873	0.6792
	PG-FDD	0.9721	0.0144	0.9784	0.6043	0.0235	0.9476	0.8033	0.0260	0.9650	**0.7366**	0.1356	0.8373
	Ours	**0.9766**	**0.0061**	**0.9907**	**0.6172**	**0.0136**	**0.9771**	**0.8184**	**0.0135**	**0.9760**	0.7351	**0.0928**	**0.8876**

Table 2. Performance comparison on deepfake detection task. ↑ means higher is better and ↓ means lower is better.

0.0090, significantly lower than PG-FDD (0.0111), DAW-FDD (0.0288), and DAG-FDD (0.0147). Similarly, in the DFDC cross-domain scenario, our method attains a violation of 0.0048, reducing fairness violation by 1.36% compared to PG-FDD. This trend persists across DFD and Celeb-DF, where our method consistently maintains the lowest fairness violation. While PG-FDD achieves higher AUC as it is the state-of-the-art method for fairness generalization in deepfake detection, its fairness violation remains higher than ours across all datasets. Overall, our results demonstrate superior fairness preservation under noisy groups across different datasets and model backbones, indicating the robustness of our method in real-world image analysis applications.

5.3. Sensitivity Analysis

Inter-/Intra- Group Performance Gap. We examine the inter-group and intra-group AUC gaps across multiple datasets to assess the capability of our method in preserving fairness under noisy groups. As shown in the left of Fig. 3, our method reduces the intra-group AUC gap by 3.40% compared to MinimaxFairAUC on the Bank dataset and by 1.70% compared to PG-FDD on the Celeb-DF dataset. These results demonstrate that our method effectively mitigates both intra-group and inter-group AUC disparities simultaneously in the presence of noisy labels, as evidenced by its performance across both tabular and image datasets.

Correctness of Noise Ratio Estimation. To evaluate the accuracy of noise ratio estimation in the image-data scenario, we experiment with multiple noise levels ($\gamma \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$), as the exact noise ratio is unknown. We apply our method to train deepfake detectors using an EfficientNet-B4 backbone and test on the FF++ dataset. As shown in Fig. 3 (Right), our method achieves the lowest fairness violation at $\gamma = 0.02$, which aligns with the estimated γ derived from Eq. (10). This consistency validates the correctness of our noise ratio estimation.

5.4. Ablation Study

With vs. Without SAM. We conduct experiments on image datasets using the EfficientNet-B4 backbone without applying SAM during training and report the results in Table 3. The results show a performance decline (*e.g.*, 0.9656 Min/Max on DFDC) compared to our full method (*e.g.*, 0.9771), yet it still outperforms the baseline approaches in Table 2. This highlights the importance of SAM in enhancing AUC fairness generalization in our approach.

Robust vs. Non-robust. Under the same experimental setting, we compare our full method with a version that excludes robustness by using Eq. (6). As shown in Table

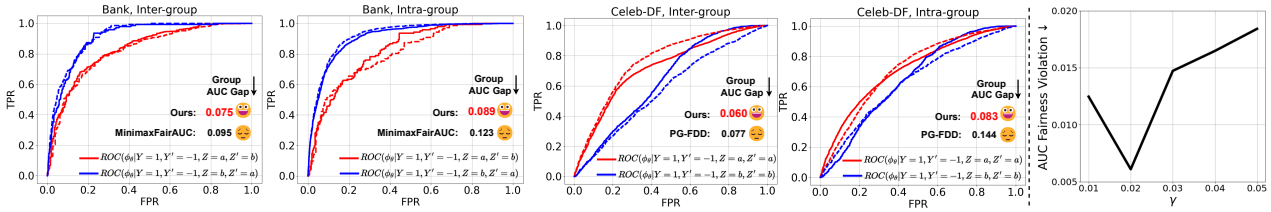

*Figure 3. (**Left**) Comparison of AUC gap on inter-group and intra-group across different datasets. For the tabular dataset, we compare our method with MinimaxFairAUC on the Bank dataset under a noise level of 0.1. For the image dataset, we compare ours with PG-FDD on Celeb-DF. (**Right**) AUC fairness violation across different γ values. γ has been set manually from $\{0.01, 0.02, 0.03, 0.04, 0.05\}$.*

Backbone	Method		DFDC			DFD		
	Robust	SAM	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑
EfficientNet -B4	✓		0.6099	0.0155	0.9656	0.7965	0.0148	0.9721
		✓	0.6090	0.0168	0.9636	0.8058	0.0289	0.9611
	✓	✓	**0.6172**	**0.0136**	**0.9771**	**0.8184**	**0.0135**	**0.9760**

*Table 3. Performance comparison between with and without SAM on our method. ↑ means higher is better and ↓ means lower is better. The best results are shown in **Bold**.*

2, the absence of robustness leads to a significant performance drop. For instance, on the DFDC dataset, the fairness violation increases by approximately 0.32%, highlighting the effectiveness of our proposed robust approach. More experiments are in Appendix F.2.

6. Conclusion

Existing methods for enhancing AUC fairness perform well with clean protected groups but fail under noisy groups. To address this, we propose a novel DRO-based approach with theoretical fairness guarantees, achieved by bounding the TV distance between clean and noisy group distributions. We estimate this bound through theoretical analysis and then develop an empirical method leveraging pre-trained multi-modal foundation models. Finally, we design an efficient SGDA algorithm to optimize the proposed learning objective, improving both AUC fairness and model generalization. Experimental results in diverse datasets highlight the superior AUC fairness maintenance capabilities of our method in three application scenarios.

Limitation. Although our approach guarantees AUC fairness, ensuring both utility and fairness simultaneously when training with noisy groups remains an open challenge. Additionally, while we use CLIP as an example of a multi-modal foundation model for noise estimation in image data, we do not imply that CLIP is directly applicable to tabular data. Rather, our statement refers to the potential to leverage domain-specific foundation models, such as those for tabular data, for analogous noise estimation tasks. Since we do not instantiate this component for tabular data in our current experiments, we acknowledge the lack of a concrete noise estimation strategy for tabular modalities as a limitation of this work and leave its exploration to future research.

Future work. In addition to addressing the aforementioned limitation, we aim to extend our approach to other pairwise ranking metrics (Yang, 2022) (*e.g.*, partial AUC, average precision) to enhance their group-level fairness.

Acknowledgements

We thank anonymous reviewers for constructive comments. This work is supported by the U.S. National Science Foundation (NSF) under grant IIS-2434967 and the National Artificial Intelligence Research Resource (NAIRR) Pilot and TACC Lonestar6. The views, opinions and/or fndings expressed are those of the author and should not be interpreted as representing the offcial views or policies of NSF and NAIRR Pilot.

Impact Statement

Our work addresses a critical and underexplored challenge in machine learning fairness: how to ensure AUC-based fairness when protected group labels are noisy or unreliable, a condition common in real-world datasets such as survey data and demographically annotated synthetic media. By proposing the first theoretically grounded, DRO framework tailored to AUC fairness under noisy protected groups, our research expands the practical reliability and applicability of fairness-aware machine learning.

The potential societal benefits are substantial. In domains like healthcare, finance, and digital media, where both performance and fairness are paramount, our method offers a robust solution that can help prevent algorithmic discrimination, especially under imperfect data labeling conditions. For example, it can reduce harm in risk-sensitive applications such as medical image analysis or deepfake detection by ensuring more equitable outcomes across groups.

Ethically, our approach prioritizes fairness even when data quality is compromised, an important step toward responsible AI. However, the broader implications of deploying fairness-aware systems should still be evaluated in context, especially when group identities are inferred or estimated.

We believe our contribution helps move the field toward more trustworthy and fair machine learning models, even under real-world constraints.

References

Deepfake detection challenge. https://www.kaggle.com/c/deepfake-detection-challenge. Accessed: 2021-04-24.

Asuncion, A., Newman, D., et al. Uci machine learning repository, 2007.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2212–2220, 2019.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pp. 1349–1361. PMLR, 2021.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pp. 272–279, 2008.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Edwards, D. A. On the kantorovich–rubinstein theorem. *Expositiones Mathematicae*, 29(4):387–398, 2011.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Gardner, J., Perdomo, J. C., and Schmidt, L. Large scale transfer learning for tabular data via language modeling. *arXiv preprint arXiv:2406.12031*, 2024.

Ghazimatin, A., Kleindessner, M., Russell, C., Abedjan, Z., and Golebiowski, J. Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2263–2279, 2022.

Ghosh, A., Kvitca, P., and Wilson, C. When fair classification meets noisy protected attributes. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 679–690, 2023.

Google and Jigsaw. Deepfakes dataset by google & jigsaw. In *https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html*, 2019.

Guo, H., Hu, S., Wang, X., Chang, M.-C., and Lyu, S. Robust attentive deep neural network for detecting gan-generated faces. *IEEE Access*, 10:32574–32583, 2022.

Gupta, M., Cotter, A., Fard, M. M., and Wang, S. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018. URL https://arxiv.org/abs/1806.11212.

Hanley, J. A. and McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Hu, P., Sun, X., Sclaroff, S., and Saenko, K. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Hu, S. and Chen, G. H. Fairness in survival analysis with distributionally robust optimization. *Journal of Machine Learning Research*, 25(246):1–85, 2024.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.

Ju, Y., Hu, S., Jia, S., Chen, G. H., and Lyu, S. Improving fairness in deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4655–4665, 2024.

Kallus, N. and Zhou, A. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

Kallus, N., Mao, X., and Zhou, A. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Kenfack, P. J., Kahou, S. E., and Aïvodji, U. A survey on fairness without demographics. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=3HE4vPNIfX.

Kollias, D., Arsenos, A., and Kollias, S. Domain adaptation explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4907–4914, 2024.

Krumpal, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & quantity*, 47(4):2025–2047, 2013.

Kumagai, A., Iwata, T., Takahashi, H., Nishiyama, T., and Fujiwara, Y. Auc maximization under positive distribution shift. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Levy, D., Carmon, Y., Duchi, J. C., and Sidford, A. Large-scale methods for distributionally robust optimization. *Advances in Neural Information Processing Systems*, 33: 8847–8860, 2020.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. Celeb-df: A new dataset for deepfake forensics. In *CVPR*, pp. 6,7, 2020.

Lin, L., He, X., Ju, Y., Wang, X., Ding, F., and Hu, S. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16815–16825, 2024.

Lin, L., Santosh, Wu, M., Wang, X., and Hu, S. Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Mehrotra, A. and Vishnoi, N. Fair ranking with noisy protected attributes. *Advances in Neural Information Processing Systems*, 35:31711–31725, 2022.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5248–5255, 2020.

Pu, W., Hu, J., Wang, X., Li, Y., Hu, S., Zhu, B., Song, R., Song, Q., Wu, X., and Lyu, S. Learning a deep dual-level network for robust deepfake detection. *Pattern Recognition*, 130:108832, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rawls, J. Justice as fairness: A restatement. *Erin Kelly/Harvard University*, 2001.

Rockafellar, R. T., Uryasev, S., et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Tian, Y., Wen, C., Shi, M., Afzal, M. M., Huang, H., Khan, M. O., Luo, Y., Fang, Y., and Wang, M. Fairdomain: Achieving fairness in cross-domain medical image segmentation and classification. In *European Conference on Computer Vision*, pp. 251–271. Springer, 2025.

Vogel, R., Bellet, A., and Clémençon, S. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International conference on artificial intelligence and statistics*, pp. 784–792. PMLR, 2021.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33:5190–5203, 2020.

Wei, T., Li, H.-T., Li, C., Shi, J.-X., Li, Y.-F., and Zhang, M.-L. Vision-language models are strong noisy label

detectors. *Advances in Neural Information Processing Systems*, 37:58154–58173, 2024.

Williamson, R. and Menon, A. Fairness risk measures. In *International conference on machine learning*, pp. 6786–6797. PMLR, 2019.

Yang, T. Deep auc maximization for medical image classification: Challenges and opportunities. *arXiv preprint arXiv:2111.02400*, 2021.

Yang, T. Algorithmic foundation of deep x-risk optimization. *arXiv preprint arXiv:2206.00439*, 2022.

Yang, Z., Ko, Y. L., Varshney, K. R., and Ying, Y. Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11909–11917, 2023.

Yao, Y., Lin, Q., and Yang, T. Stochastic methods for auc optimization subject to auc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 10324–10342. PMLR, 2023.

Yeh, I.-C. and Lien, C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., and Sugiyama, M. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pp. 7164–7173. PMLR, 2019.

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3040–3049, 2021.

Zhang, C., Shi, W., Luo, L., and Gu, B. Doubly robust auc optimization against noisy and adversarial samples. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3195–3205, 2023.

A. Proofs

A.1. Proof of Theorem 4.1

Our proof is inspired by Wang et al. (2020); however, the key distinction lies in the focus. While Wang et al. (2020) addresses non-ranking-based fairness measures, we derive results specifically for pairwise ranking-based AUC fairness measures. First, we define TV distance between $p_{z,z'}$ and $\widehat{p}_{z,z'}$ as follows,

Definition A.1. Let $m(x,y) = \mathbb{I}_{x\neq y}$ be a metric, and let π represent a coupling between the probability distributions $p_{z,z'}$ and $\widehat{p}_{z,z'}$. The TV distance is defined as $TV(p_{z,z'}, \widehat{p}_{z,z'}) = \inf_\pi \mathbb{E}_{X,Y\sim\pi}[m(X,Y)]$ s.t. $\int \pi(x,y)dy = p_{z,z'}(x), \int \pi(x,y)dx = \widehat{p}_{z,z'}(y)$.

Then, we introduce a Lemma as follows,

Lemma A.2. *(Edwards, 2011). A function h is called Lipschitz with respect to m if $|h(x) - h(y)| \leq m(x,y)$ for all x, y, and let $\mathcal{H}(m)$ denote the space of such functions. If m is a metric, the Wasserstein distance can be expressed as:*

$$W_c(p_{z,z'}, \widehat{p}_{z,z'}) = \sup_{h\in\mathcal{H}(m)} \mathbb{E}_{X\sim p_{z,z'}}[h(X)] - \mathbb{E}_{X\sim\widehat{p}_{z,z'}}[h(X)].$$

Now, let $m(x,y) = \mathbb{I}_{x\neq y}$. In this case, the Total Variation (TV) distance becomes:

$$TV(p_{z,z'}, \widehat{p}_{z,z'}) = \sup_{h:\mathcal{X}\to[0,1]} \mathbb{E}_{X\sim p_{z,z'}}[h(X)] - \mathbb{E}_{X\sim\widehat{p}_{z,z'}}[h(X)].$$

Finally, we prove Theorem 4.1 as follows,

Proof. For any pairwise group labels z, z',

$$g_{z,z'}(\theta) = g_{z,z'}(\theta) - \widehat{g}_{z,z'}(\theta) + \widehat{g}_{z,z'}(\theta) \leq |g_{z,z'}(\theta) - \widehat{g}_{z,z'}(\theta)| + \widehat{g}_{z,z'}(\theta).$$

By Lemma A.2, we have the following result.

$$|g_{z,z'}(\theta) - \hat{g}_{z,z'}(\theta)| = \left|\mathbb{E}[h(\theta)|Z=z, Z'=z'] - \mathbb{E}[h(\theta)|\widehat{Z}=z, \widehat{Z}'=z']\right| \leq TV(p_{z,z'}, \hat{p}_{z,z'}).$$

Given the assumption that θ satisfies the fairness constraints with respect to the noisy groups, $\hat{g}_{z,z'}(\theta) \leq 0$. Therefore, we derive the desired result:

$$g_{z,z'}(\theta) \leq TV(p_{z,z'}, \hat{p}_{z,z'}) \leq \gamma_{z,z'}.$$

□

A.2. Proof of Lemma 4.2

Proof. The Total Variation (TV) distance between the probability measures $p_{z,z'}$ and $\hat{p}_{z,z'}$ is defined as:

$$TV(p_{z,z'}, \hat{p}_{z,z'}) = \sup\{|p_{z,z'}(A) - \hat{p}_{z,z'}(A)| : A \text{ is a measurable event}\}.$$

Let A be any measurable event under both $p_{z,z'}$ and $\hat{p}_{z,z'}$. By the definition of $p_{z,z'}$, we have $p_{z,z'}(A) = Pr[A \mid (Z, Z') =$

Algorithm 2 `Sampler`(Dataset: $\mathcal{S}$, batch_size: b)

1: **for** $z \in \mathcal{Z}$ and $Y \in \mathcal{Y}$ **do**
2: Uniformly sample without replacement B^{zY} from $\mathcal{S}^{zY}$ with size $b^{zY} = \lceil b \cdot (|\mathcal{S}^{zY}|/|S|) \rceil$
3: **end for**
4: **return** $B = \cup_{z,Y} B^{zY}$

$(z, z')]$. In the context of the pairwise problem between $p_{z,z'}$ and $\hat{p}_{z,z'}$, it follows that:

$$
\begin{aligned}
&|p_{z,z'}(A) - \widehat{p}_{z,z'}(A)| \\
&= |Pr[A \mid (Z, Z') = (z, z')] - Pr[A \mid (\widehat{Z}, \widehat{Z}') = (z, z')]| \\
&= |Pr[A \mid (Z, Z') = (z, z'), (\widehat{Z}, \widehat{Z}') = (z, z')] Pr[(\widehat{Z}, \widehat{Z}') = (z, z') \mid (Z, Z') = (z, z')] \\
&\quad + Pr[A \mid (Z, Z') = (z, z'), (\widehat{Z}, \widehat{Z}') \neq (z, z')] Pr[((\widehat{Z}, \widehat{Z}') \neq (z, z') \mid (Z, Z') = (z, z')] \\
&\quad - Pr[A \mid (\widehat{Z}, \widehat{Z}') = (z, z'), (Z, Z') = (z, z')] Pr[(Z, Z') = (z, z') \mid (\widehat{Z}, \widehat{Z}') = (z, z')] \\
&\quad - Pr[A \mid (\widehat{Z}, \widehat{Z}') = (z, z'), (Z, Z') \neq (z, z')] Pr[(Z, Z') \neq (z, z') \mid (\widehat{Z}, \widehat{Z}') = (z, z')]| \\
&= |Pr[A \mid (Z, Z') = (z, z'), (\widehat{Z}, \widehat{Z}') = (z, z')] \cdot \\
&\qquad \big(Pr[(\widehat{Z}, \widehat{Z}') = (z, z') \mid (Z, Z') = (z, z')] - Pr[(Z, Z') = (z, z') \mid (\widehat{Z}, \widehat{Z}') = (z, z')]\big) \\
&\quad - Pr[(\widehat{Z}, \widehat{Z}') \neq (Z, Z') \mid (Z, Z') = (z, z')] \cdot \\
&\qquad \big(Pr[A \mid (Z, Z') = (z, z'), (\widehat{Z}, \widehat{Z}') \neq (z, z')] - Pr[A \mid (\widehat{Z}, \widehat{Z}') = (z, z'), (Z, Z') \neq (z, z')]\big)| \\
&= |0 - Pr[(\widehat{Z}, \widehat{Z}') \neq (Z, Z') \mid (Z, Z') = (z, z')] \cdot \\
&\qquad \big(Pr[A \mid (Z, Z') = (z, z'), (\widehat{Z}, \widehat{Z}') \neq (z, z')] - Pr[A \mid (\widehat{Z}, \widehat{Z}') = (z, z'), (Z, Z') \neq (z, z')]\big)| \\
&\leq Pr[(\widehat{Z}, \widehat{Z}') \neq (Z, Z') \mid (Z, Z') = (z, z')] = Pr[(Z, Z') \neq (\widehat{Z}, \widehat{Z}') \mid (\widehat{Z}, \widehat{Z}') = (z, z')].
\end{aligned}
$$

The second equality follows from the law of total probability. The third and the fourth equalities follow from the assumption that $Pr[(Z, Z') = (z, z')] = Pr[(\widehat{Z}, \widehat{Z}') = (z, z')]$, which implies that $Pr[(\widehat{Z}, \widehat{Z}') = (Z, Z') \mid (Z, Z') = (z, z')] = Pr[(Z, Z') = (\widehat{Z}, \widehat{Z}') \mid (\widehat{Z}, \widehat{Z}') = (z, z')]$ since

$$
\begin{aligned}
Pr[(\widehat{Z}, \widehat{Z}') = (Z, Z') \mid (Z, Z') = (z, z')] &= \frac{Pr[(\widehat{Z}, \widehat{Z}') = (Z, Z'); (Z, Z') = (z, z')]}{Pr[(Z, Z') = (z, z')]} \\
&= \frac{Pr[(\widehat{Z}, \widehat{Z}') = (Z, Z'); (\widehat{Z}, \widehat{Z}') = (z, z')]}{Pr[(\widehat{Z}, \widehat{Z}') = (z, z')]} \\
&= Pr[(\widehat{Z}, \widehat{Z}') = (Z, Z') \mid (\widehat{Z}, \widehat{Z}') = (z, z')]
\end{aligned}
$$

This further implies that $Pr[(\widehat{Z}, \widehat{Z}') \neq (Z, Z') \mid (Z, Z') = (z, z')] = Pr[(\widehat{Z}, \widehat{Z}') \neq (Z, Z') \mid (\widehat{Z}, \widehat{Z}') = (z, z')]$, which is the reason why the last equation holds.

B. Sampling Method

We denote strata of data as $\mathcal{S}^{zY}$, where $z \in \mathcal{Z}$ and $Y \in \mathcal{Y}$. Then, the sampling algorithm is shown in Algorithm 2.

C. Datasets Details

For tabular datasets, we do socioeconomic analysis on three datasets that have been commonly used in the fair machine learning literature (Donini et al., 2018). In `Adult` dataset, the sensitive attribute is the gender of the individual, i.e. female ($Z = a$) or male ($Z = b$). In `Bank` dataset , the sensitive attribute is the age of the individual: $Z = a$ when the age is less than 25 or over 60 and $Z = b$ otherwise. In `Default` dataset (Yeh & Lien, 2009) , the sensitive attribute is the gender of the individual, i.e. female ($Z = a$) or male ($Z = b$).

For image datasets, we do deepfake detection on the most widely used benchmark FaceForensics++(FF++) (Rossler et al., 2019). DFDC (dee), DFD (Google & Jigsaw, 2019), and Celeb-DF (Li et al., 2020). We only used the test set of the later

Type	Name	# Instances	Group ratio	Class ratio
Tabular	Adult	48,842	0.48:1	3.03:1
	Bank	41,188	0.05:1	7.55:1
	Default	30,000	1.52:1	3.52:1
Image	FF++	76,139	1.27:1	1:4.89
	DFDC	22,857	1.05:1	1.80:1
	DFD	9,386	1.22:1	1:2.04
	Celeb-DF	28,458	8.79:1	1:3.98

Table 4. Dataset Statistics. The group ratio is given by the protective attribute $Z = a$ vs. $Z = b$. The class ratio is given by negative vs. positive class.

three deepfake dataset. Since the original datasets do not have the demographic information of each video or image, we follow Ju et al. (2024) for data processing and data annotation. The sensitive attribute is the gender of the individual, i.e. female ($Z = a$) or male ($Z = b$). The summary statistics of the datasets are given in Table 4.

D. Fairness Metrics

For fairness, we measure AUC fairness violation ('Violation', lower is better), defined as the maximum absolute difference between any group-level AUC score the overall AUC score:

$$\max_{z,z' \in \mathcal{Z}} |g_{z,z'}(\theta)|.$$

Additionally, following Yang et al. (2023), we use the Min/Max fairness metric (higher is better), defined as the ratio of the minimum to the maximum group-level AUC score:

$$\min_{z,z' \in \mathcal{Z}} AUC_{z,z'}(\theta) / \max_{z,z' \in \mathcal{Z}} AUC_{z,z'}(\theta).$$

E. Baselines

- The AUCMax algorithm maximizes AUC across the entire dataset without distinguishing between groups. It updates the model parameters using mini-batch SGD.
- We select the method by Vogel et al. (2021) as a representative since they considered the same datasets. Their approach, which we refer to as InterFairAUC, ensures fair AUC scores by regularizing the difference between inter-group AUCs.
- MinimaxFairAUC introduced by Yang et al. (2023) is a minimax fairness framework that simultaneously addresses intra-group and inter-group AUC disparities using a Rawlsian approach, sup- ported by an efficient optimization algorithm with proven convergence guarantees.
- DAG-FDD (Ju et al., 2024), a demographic-aware Fair Deepfake Detection (DAW-FDD) method leverages demographic information and employs an existing fairness risk measure (Williamson & Menon, 2019). At a high level, DAW-FDD aims to ensure that the losses achieved by different user-specified groups of interest (*e.g.*, different races or genders) are similar to each other (so that the AI face detector is not more accurate on one group vs another) and, moreover, that the losses across all groups are low. Specifically, DAW-FDD uses a CVaR (Levy et al., 2020; Rockafellar et al., 2000) loss function across groups (to address imbalance in demographic groups) and, per group, DAW-FDD uses another CVaR loss function (to address imbalance in real vs AI-generated training examples).
- DAW-FDD (Ju et al., 2024), a demographic-agnostic Fair Deepfake Detection (DAG-FDD) method, which is based on the distributionally robust optimization (DRO) (Hashimoto et al., 2018; Duchi & Namkoong, 2021). To use DAG-FDD, the user does not have to specify which attributes to treat as sensitive such as race and gender, only need to specify a probability threshold for a minority group without explicitly identifying all possible groups.
- PG-FDD (Lin et al., 2024) (Preserving Generalization Fair Deepfake Detection) employs disentanglement learning to extract demographic and domain-agnostic forgery features, promoting fair learning across a flattened loss landscape.

Noise Level	Method	Adult			Bank			Default		
		AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑
0.1	Ours(Non-robust)	0.9064±0.0008	0.0370±0.0006	0.9535±0.0006	0.9059±0.0001	0.0920±0.0029	0.9102±0.0025	0.7538±0.0024	0.0212±0.0016	0.9650±0.0019
	Ours (Robust)	0.9060±0.0002	**0.0332±0.0007**	**0.9536±0.0005**	0.9039±0.0007	**0.0876±0.0033**	**0.9143±0.0029**	0.7645±0.0017	**0.0187±0.0012**	**0.9691±0.0019**
0.2	Ours(Non-robust)	0.9108±0.0009	0.0388±0.0006	0.9532±0.0004	0.9107±0.0008	**0.0868±0.0029**	**0.9111±0.0024**	0.7511±0.0016	0.0230±0.0018	0.9626±0.0019
	Ours (Robust)	0.9039±0.0004	**0.0328±0.0008**	**0.9539±0.0011**	0.9163±0.0005	0.1053±0.0026	0.8988±0.0024	0.7624±0.0016	**0.0206±0.0014**	**0.9660±0.0022**
0.3	Ours(Non-robust)	0.8971±0.0072	0.0370±0.0037	**0.9569±0.0007**	0.9060±0.0012	0.0973±0.0026	0.9053±0.0025	0.7540±0.0025	0.0213±0.0015	0.9647±0.0019
	Ours (Robust)	0.9059±0.0005	**0.0356±0.0006**	0.9513±0.0007	0.9093±0.0004	**0.0949±0.0026**	**0.9080±0.0024**	0.7687±0.0011	**0.0129±0.0011**	**0.9785±0.0014**
0.4	Ours(Non-robust)	0.9075±0.0012	0.0377±0.0006	0.9522±0.0001	0.9109±0.0011	0.0943±0.0010	0.9085±0.0008	0.7429±0.0018	0.0263±0.0007	**0.9594±0.0015**
	Ours (Robust)	0.9008±0.0002	**0.0341±0.0005**	**0.9530±0.0004**	0.9054±0.0004	**0.0937±0.0034**	**0.9086±0.0028**	0.7552±0.0024	**0.0246±0.0036**	0.9580±0.0066
0.5	Ours(Non-robust)	0.8989±0.0011	0.0372±0.0007	0.9553±0.0002	0.9108±0.0012	0.0895±0.0029	0.9126±0.0024	0.7505±0.0014	0.0261±0.0001	0.9564±0.0001
	Ours (Robust)	0.8984±0.0006	**0.0316±0.0002**	**0.9554±0.0004**	0.9044±0.0004	**0.0876±0.0020**	**0.9145±0.0015**	0.7447±0.0047	**0.0243±0.0014**	**0.9571±0.0010**

Table 5. Performance comparison across different noise levels (0.1–0.5) between robust and non-robust method of ours. The numbers are reported as 'Mean ± Standard Deviation.' ↑ means higher is better and ↓ means lower is better.

Backbone	Method	FF++			DFDC			DFD			Celeb-DF		
		AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑	AUC↑	Violation↓	Min/Max↑
Xception	Ours(Non-robust)	0.9546	0.0103	0.9842	0.6014	0.0191	0.9573	0.7826	0.0098	0.9863	0.7105	0.0994	0.8807
	Ours	0.9644	**0.0090**	**0.9857**	0.6086	**0.0048**	**0.9930**	0.7847	**0.0069**	**0.9881**	0.7108	**0.0729**	**0.9117**
EfficientNet-B4	Ours(Non-robust)	0.9729	0.0100	0.9842	0.6090	0.0168	0.9636	0.8058	0.0289	0.9611	0.7330	0.1046	0.8715
	Ours	**0.9766**	**0.0061**	**0.9907**	**0.6172**	**0.0136**	**0.9771**	**0.8184**	**0.0135**	**0.9760**	0.7351	**0.0928**	**0.8876**

Table 6. Performance comparison of Ours (Non-robust) and Ours on deepfake detection task. ↑ means higher is better and ↓ means lower is better.

Its framework combines disentanglement learning, fairness learning, and optimization modules. The disentanglement module introduces a loss to expose demographic and domain-agnostic features that enhance fairness generalization. The fairness learning module combines these features to promote fair learning, guided by generalization principles. The optimization module flattens the loss landscape, helping the model escape suboptimal solutions and strengthen fairness generalization.

F. More Implementation Details and results

F.1. Additional experimental setup details

In Algorithm 1, we explored the following hyperparameters:

- Tabular data:
  1. $\eta_\theta \in \{0.001, 0.01, 0.1\}$
  2. $\eta_\lambda \in \{0.001, 0.25, 0.5\}$
  3. $\eta_p \in \{0.001, 0.01, 0.1\}$
  4. ν in Eq. (11) is selected from $\{0.0005, 0.001, 0.005\}$
- Image data:
  1. $\eta_\theta \in \{0.0001, 0.0005, 0.001\}$
  2. $\eta_\lambda \in \{0.0001, 0.0005, 0.005\}$
  3. $\eta_p \in \{0.0001, 0.0005, 0.001\}$
  4. ν in Eq. (11) is selected from $\{0.7, 0.5, 0.3\}$

F.2. Additional experimental results

Table 5 and table 6 show the results between robust and non-robust methods of ours. In general, our method with using robust approach shows more robust performance than our method without the robust approach.

Label setting. In Table 1, fairness metrics use underlying group labels. In Table 2, we use noisy group labels, which are common in datasets like FF++ where demographic attributes are inferred. Evaluating fairness under label noise is practical and follows prior work ((Celis et al., 2021; Mehrotra & Vishnoi, 2022)). We also test the model on a human-corrected FF++ test set from the Lin et al. (2025), which contains relatively clean labels. As shown in Table 7, our method still maintains the best performance across all fairness metrics.

Backbone	Method	FF++ clean		
		AUC↑	Violation↓	Min/Max↓
Xception	Ori	0.9384	0.0289	0.9691
	DAG-FDD	0.9628	0.0128	0.9808
	DAW-FDD	0.965	0.0256	0.9751
	PG-FDD	**0.9708**	0.0105	0.9840
	Ours	0.9644	**0.0070**	**0.9894**

Table 7. Performance comparison on clean version of FF++ using Xception backbone. ↑ means higher is better and ↓ means lower is better.

Backbone	Method	Training Time per Epoch (minutes)
Xception	Original	8
	DAG-FDD	6
	DAW-FDD	9
	PG-FDD	28
	Ours	15

Table 8. Training time comparison on deepfake detection task on FF++ dataset.

Computational overhead. To evaluate the practicality of our approach at scale, we benchmarked training time on the FaceForensics++ dataset, a widely used, large-scale benchmark for deepfake detection. As shown in table 8, our method introduces moderate overhead compared to some baselines but remains significantly more efficient than others, such as PG-FDD. Specifically, our method requires 15 minutes per epoch, which is faster than PG-FDD (28 min) and reasonably close to other baselines like DAW-FDD and the Original model. This demonstrates that our approach remains computationally feasible and scalable in practice, even for large image datasets and backbone models like Xception.

Experiments with extreme settings. As Table 9 shows, even in extremely high noise levels (60%, 70%, 80%, and 90%), our method consistently achieves the lowest AUC fairness violation and the highest Min/Max AUC score. These results strongly align with our claims in the paper, demonstrating that our approach maintains robust fairness guarantees and balanced group performance in highly noisy settings.

Results on more groups. We add one more experiment about evaluating our method on a dataset with more than two protected groups. Specifically, in Table 10, we present results on the FaceForensics++ (FF++) dataset, where we consider race as the protected attribute, comprising four groups: White, Black, Asian, and Other. This multi-group setting is more challenging than the binary group setting commonly seen in fairness literature. Nonetheless, our method achieves the lowest fairness violation (0.0161) and the highest Min/Max AUC score (0.9850) among all baselines, demonstrating its effectiveness in ensuring AUC fairness across multiple protected groups. These results confirm that our approach generalizes well to settings involving complex, non-binary group structures, such as race.

CLIP for label prediction. We added experiments with one more baseline to address your concern. Specifically, as shown in Table 11, we include a baseline called Ours (CLIP-labeled), where we directly use CLIP to predict the protected group labels and then train the model based on those labels. In contrast, our full method, Ours (robust), uses CLIP only to estimate the group label noise level, not for prediction or relabeling. It is clear that Ours (robust) outperforms Ours (CLIP-labeled) in all metrics, achieving higher AUC (0.9766 vs. 0.9725), lower violation (0.0061 vs. 0.0089), and better Min/Max fairness (0.9907 vs. 0.9858). This indicates that directly using CLIP-predicted labels to mitigate the impact of noise during training does not yield optimal performance. More importantly, this baseline approach does not guarantee fairness under the noisy label setting.

Additional Visualizations. To better communicate the fairness–performance tradeoff across all methods in Table 1, we include two efficiency frontier plots in Fig 4 and Fig 5. We used the results in Table 1, which include performance on three tabular datasets: Adult, Bank, and Default. For each fairness-enhanced method, we compute the average AUC, average fairness violation, and average Min/Max AUC ratio across the three datasets for each noise level (0.1–0.3). This gives three points per method, each representing the average values across datasets at a given noise level. Fig 4 visualizes Average AUC vs. Average Fairness Violation, demonstrating the tradeoff of performance and fairness. Our method ("Ours") consistently occupies the top-left region, achieving lower violation than all baselines while maintaining competitive or superior AUC. Fig

Noise Level	Method	Default		
		AUC↑	Violation↓	Min/Max↑
0.6	AUCMax	**0.7771±0.0005**	0.0672±0.0004	0.9108±0.0005
	InterFairAUC	0.7548±0.0022	0.0238±0.0039	0.9586±0.0062
	MinimaxFairAUC	0.7550±0.0021	0.0249±0.0054	0.9577±0.0078
	Ours	0.7540±0.0024	**0.0186±0.0027**	**0.9676±0.0045**
0.7	AUCMax	**0.7788±0.0007**	0.0696±0.0010	0.9081±0.0012
	InterFairAUC	0.7532±0.0020	0.0239±0.0018	0.9594±0.0052
	MinimaxFairAUC	0.7545±0.0021	0.0265±0.0062	0.9557±0.0079
	Ours	0.7541±0.0024	**0.0183±0.0033**	**0.9687±0.0052**
0.8	AUCMax	**0.7760±0.0006**	0.0687±0.0014	0.9092±0.0017
	InterFairAUC	0.7548±0.0022	0.0251±0.0043	0.9566±0.0062
	MinimaxFairAUC	0.7532±0.0021	0.0246±0.0039	0.9588±0.0072
	Ours	0.7546±0.0025	**0.0183±0.0035**	**0.9688±0.0057**
0.9	AUCMax	**0.7795±0.0006**	0.0675±0.0007	0.9108±0.0008
	InterFairAUC	0.7545±0.0022	0.0252±0.0045	0.9567±0.0062
	MinimaxFairAUC	0.7544±0.0022	0.0267±0.0061	0.9555±0.0078
	Ours	0.7527±0.0022	**0.0201±0.0031**	**0.9661±0.0049**

*Table 9. Performance comparison across high noise levels (0.6–0.9). The numbers are reported as 'Mean ± Standard Deviation.' ↑ means higher is better and ↓ means lower is better. The best results are shown in **Bold**.*

Backbone	Method	FF++		
		AUC↑	Violation↓	Min/Max↓
Xception	Ori	0.9445	0.0335	0.9633
	DAG-FDD	0.9647	0.0226	0.9798
	DAW-FDD	0.9670	0.0215	0.9753
	PG-FDD	**0.9751**	0.0172	0.9837
	Ours	0.9625	**0.0161**	**0.9850**

Table 10. Performance comparison on the FF++ dataset with race as the protected attribute. ↑ means higher is better and ↓ means lower is better.

5 shows Average AUC vs. Average Min/Max AUC Ratio, highlighting group-wise fairness consistency. Our method ranks in the top-right region, attains the best Min/Max ratio with strong AUC, reflecting better fairness stability across groups.

To complement the tabular results in Table 2, we include two efficiency frontier plots based on the average values across all four benchmark datasets (FF++, DFDC, DFD, and Celeb-DF). Fig 6 reveals the trade-off between detection performance and fairness violation. Our method ("Ours") achieves the lowest average violation while maintaining a higher AUC than all other methods on both Xception and EfficientNet-B4 backbones. These results position our method at the top-left corner, indicating Pareto efficiency and demonstrating strong performance–fairness trade-offs. Fig 7 captures performance versus group-wise fairness consistency. Our method again ranks at the top-right region, with both the highest min/max ratio and competitive or superior AUC. This confirms that our method not only performs well but also offers greater fairness stability across demographic groups. We note that while the table provides fine-grained per-dataset results, these plots offer a complementary view by highlighting global efficiency across multiple objectives. The strong position of our method in both plots further supports its overall effectiveness.

Backbone	Method	FF++		
		AUC↑	Violation↓	Min/Max↓
EfficientNet-B4	Ori	0.9332	0.0209	0.9684
	DAG-FDD	0.9563	0.0100	0.9869
	DAW-FDD	0.9694	0.0169	0.9764
	PG-FDD	0.9721	0.0144	0.9784
	Ours(CLIP-labeled)	0.9725	0.0089	0.9858
	Ours(robust)	**0.9766**	**0.0061**	**0.9907**

Table 11. Performance comparison on FF++ using EfficientNet-B4 backbone. ↑ means higher is better and ↓ means lower is better.

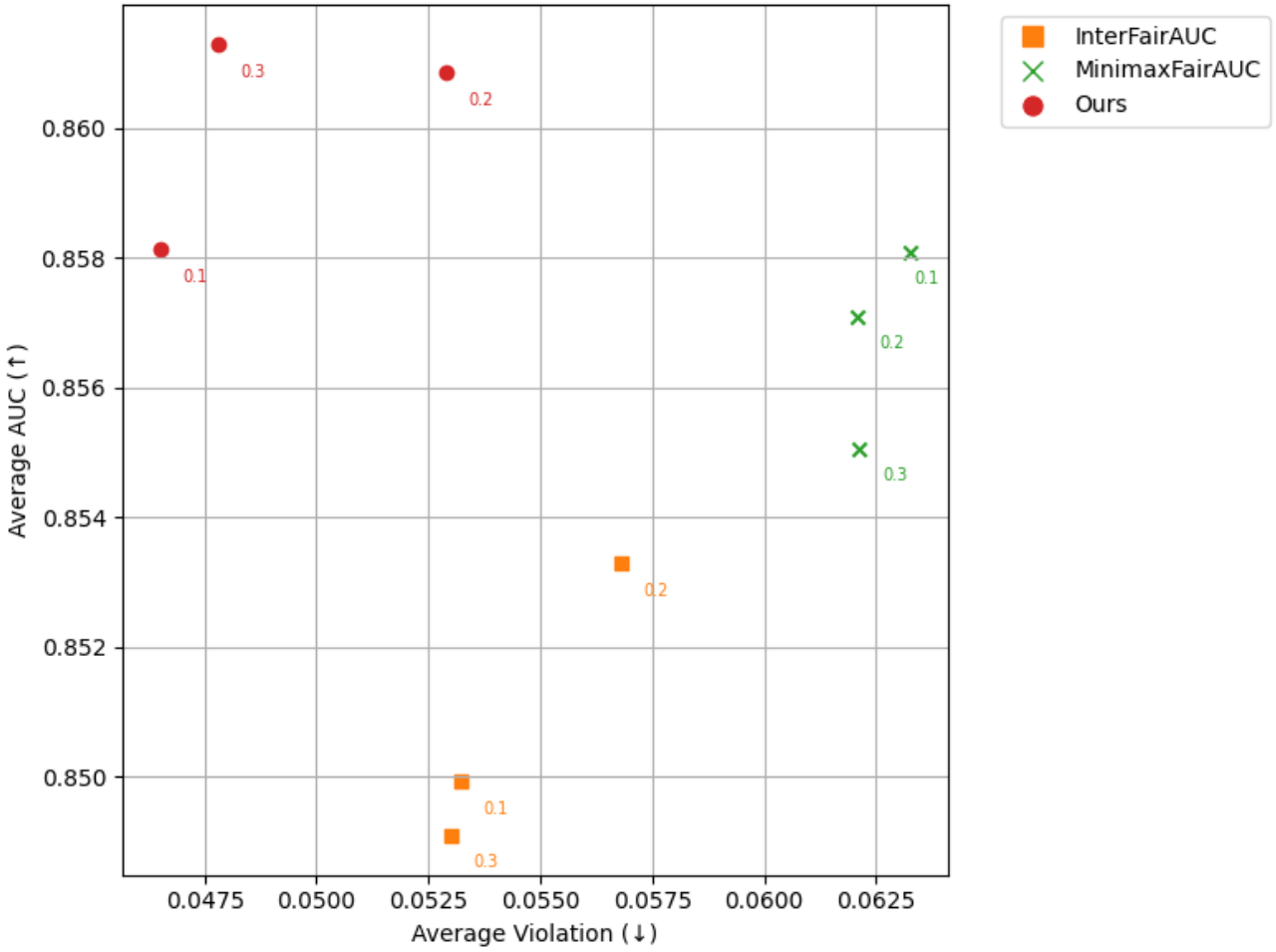

Figure 4. Efficiency frontier showing the trade-off between Average AUC and Average Fairness Violation across three tabular datasets (Adult, Bank, Default) at varying noise levels (0.1–0.3).

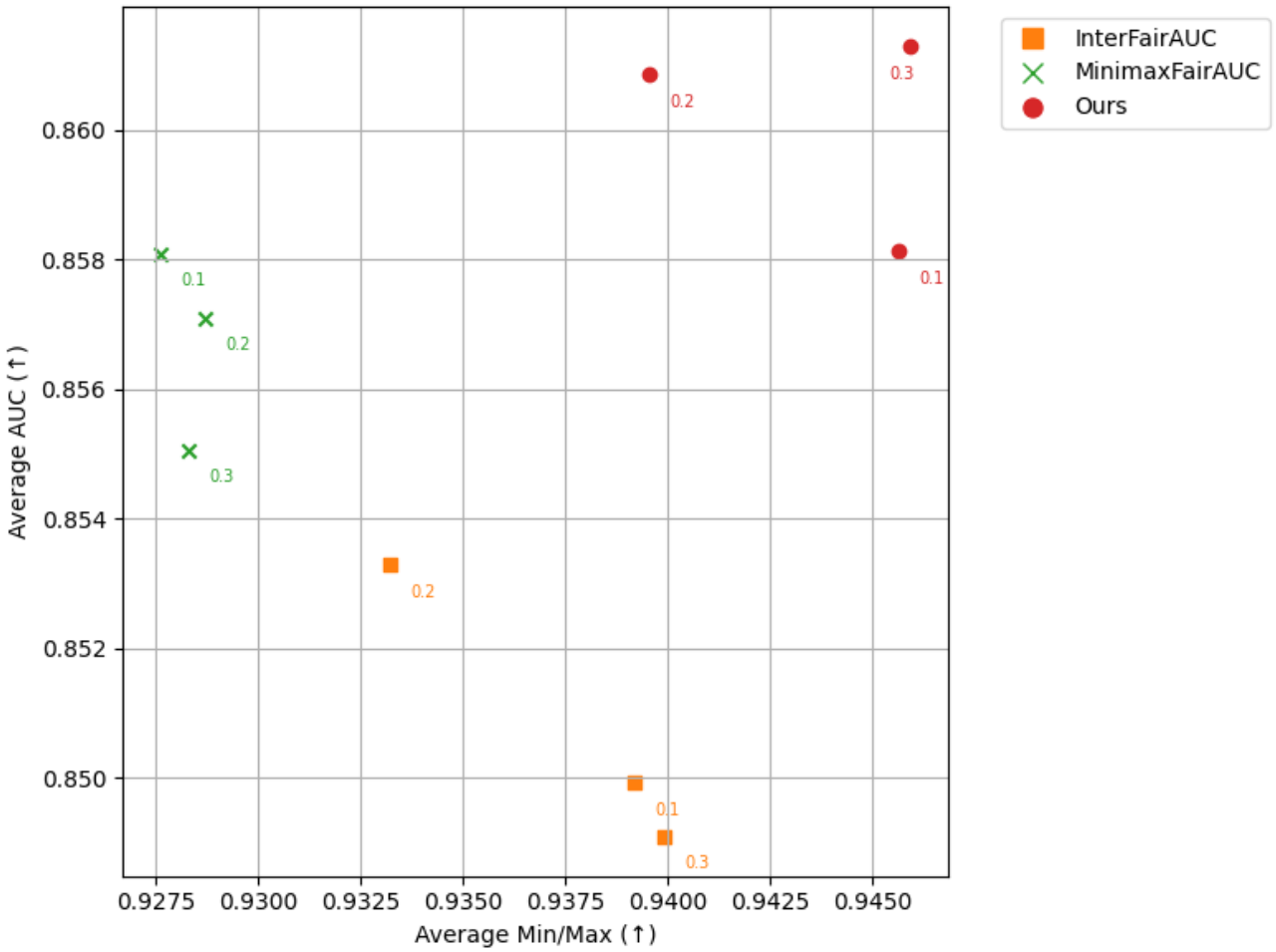

Figure 5. Efficiency frontier showing the trade-off between Average AUC and Average Min/Max AUC across three tabular datasets (Adult, Bank, Default) at varying noise levels (0.1–0.3).

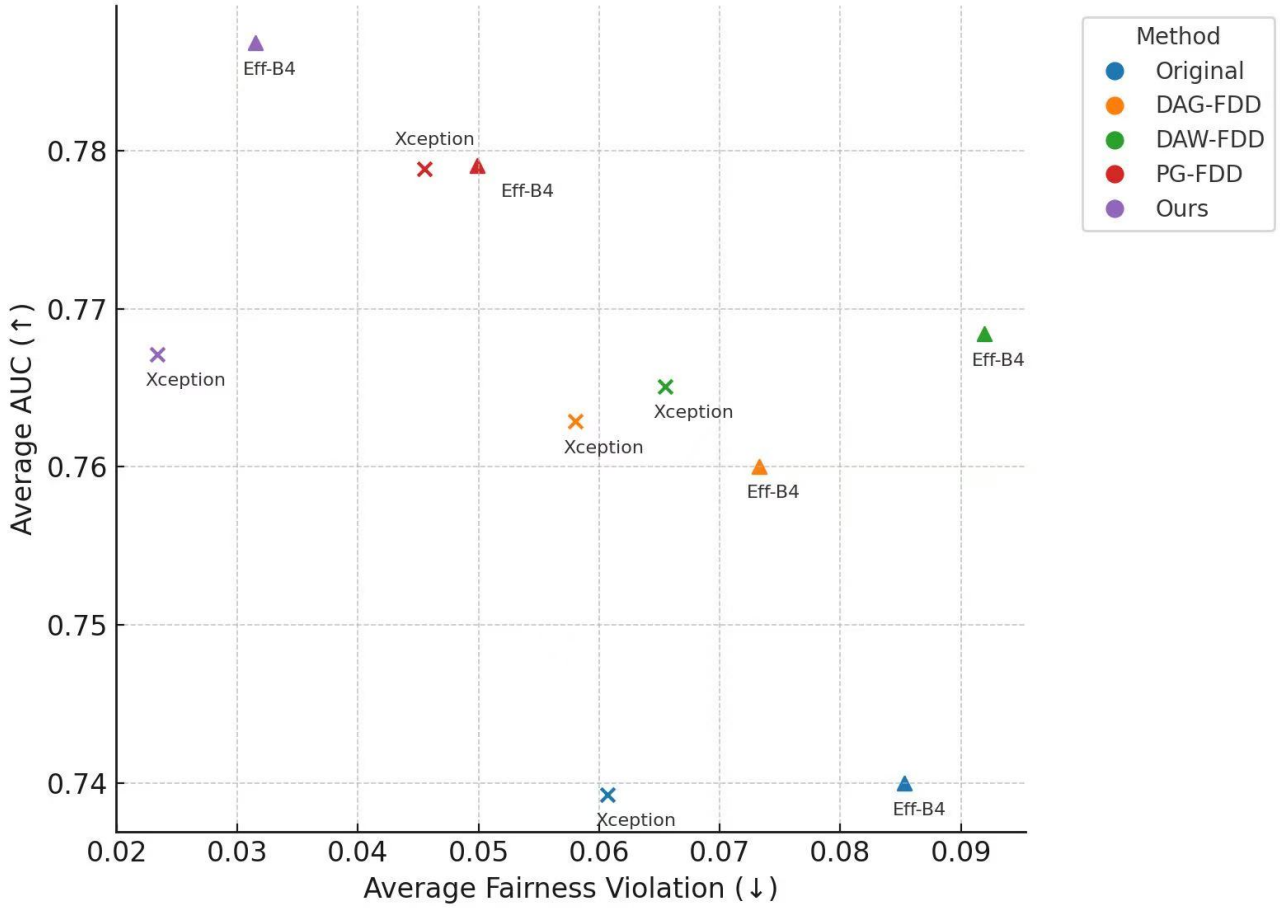

Figure 6. Efficiency frontier showing the trade-off between detection performance (AUC) and fairness violation across four benchmark datasets (FF++, DFDC, DFD, Celeb-DF).

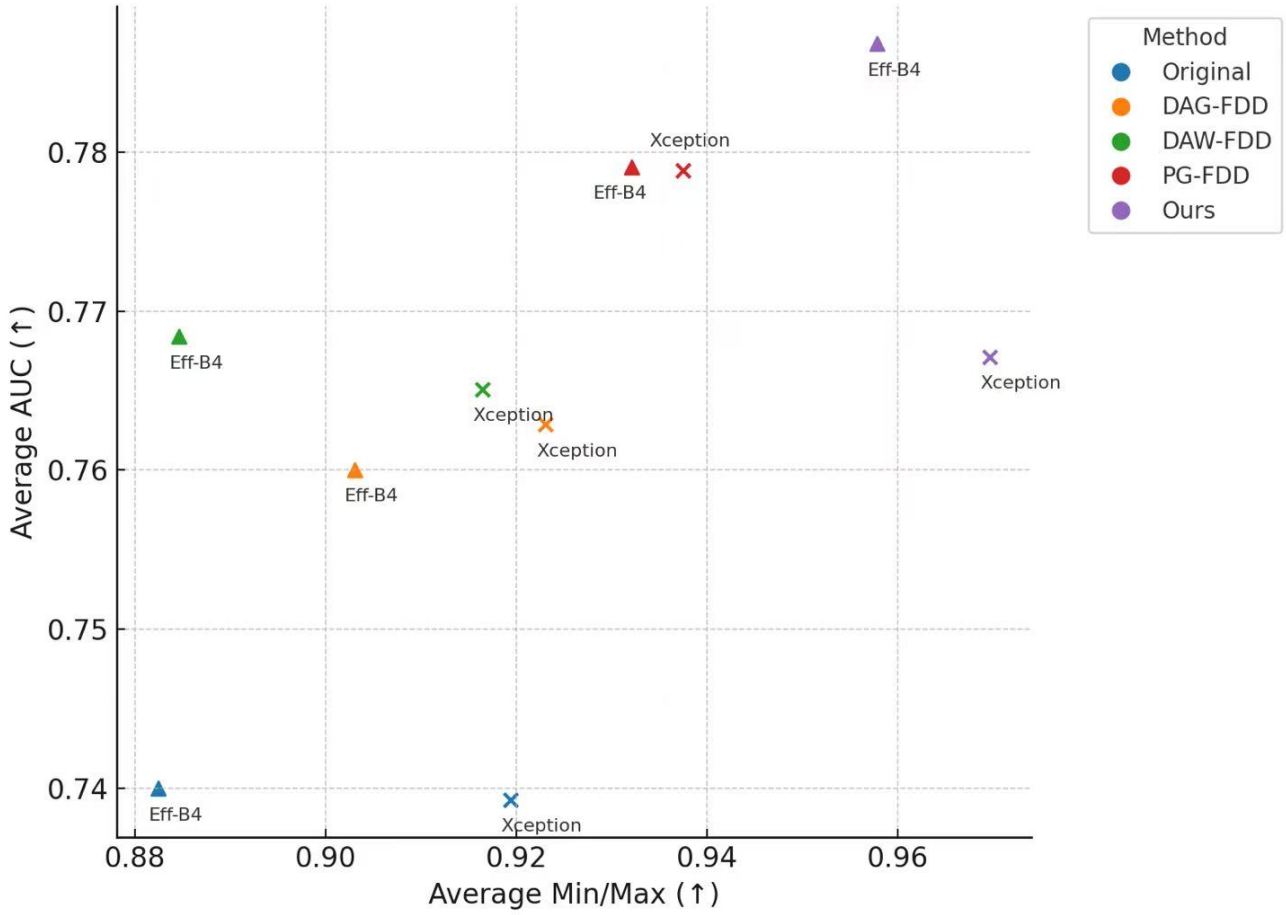

Figure 7. Efficiency frontier showing the trade-off between detection performance (AUC) and Min/Max AUC across four benchmark datasets (FF++, DFDC, DFD, Celeb-DF).