Improving Fairness in Deepfake Detection

Yan Ju[1*], Shu Hu[2*†], Shan Jia[1], George H. Chen[3], Siwei Lyu[1†]
[1] University at Buffalo, State University of New York {yanju, shanjia, siweilyu}@buffalo.edu
[2] Indiana University–Purdue University Indianapolis hu968@purdue.edu
[3] Carnegie Mellon University georgechen@cmu.edu

Abstract

*Despite the development of effective deepfake detectors in recent years, recent studies have demonstrated that biases in the data used to train these detectors can lead to disparities in detection accuracy across different races and genders. This can result in different groups being unfairly targeted or excluded from detection, allowing undetected deepfakes to manipulate public opinion and erode trust in a deepfake detection model. While existing studies have focused on evaluating fairness of deepfake detectors, to the best of our knowledge, no method has been developed to encourage fairness in deepfake detection at the algorithm level. In this work, we make the first attempt to improve deepfake detection fairness by proposing novel loss functions that handle both the setting where demographic information (*e.g.*, annotations of race and gender) is available as well as the case where this information is absent. Fundamentally, both approaches can be used to convert many existing deepfake detectors into ones that encourages fairness. Extensive experiments on four deepfake datasets and five deepfake detectors demonstrate the effectiveness and flexibility of our approach in improving deepfake detection fairness. Our code is available at https://github.com/littlejuyan/DF_Fairness.*

1. Introduction

"Deepfakes" refer to realistic images and videos where a person's likeness has been replaced by that of another with the help of deep learning technologies. Concerns have arisen regarding deepfakes being used for malicious purposes, such as in political propaganda or cyberattacks. For example, a deepfake video can depict a world leader making statements or taking actions that never occurred in reality [1], which could deceive the public. To mitigate the impact of deepfakes, a variety of deepfake detectors have been developed with promising detection accuracy [2–13].

However, recent studies [2, 14–18] have shown that current deepfake detectors are unfair: their detection accuracy is not consistent across gender, age, and ethnicity [15]. For example, several state-of-the-art detectors have higher detection accuracy for deepfakes with lighter skin tones than deepfakes with darker skin tones [14, 19]. A key reason for this disparity is that how often different demographic groups appear in the training data is imbalanced [16]. Collecting a larger "balanced" dataset can be costly and labor-intensive [20]. While conventional fairness methods can be applied (*e.g.*, by adding a fairness regularization term to the overall loss function [21]), deepfake detection poses an additional level of complexity. Specifically, we need to account for the imbalance in real vs training deepfake examples in addition to the usual imbalance in demographic groups.

In this paper, we propose two Fair Deepfake Detection (FDD) methods, both of which can be used to modify an existing deep-learning-based deepfake detector that does not account for fairness into one that does:

1. Our first method DAG-FDD (demographic-agnostic FDD) does not rely on demographic details (the user does not have to specify which attributes to treat as sensitive such as race and gender) and can be applied when, for instance, these demographic details have not been collected for the dataset. To use DAG-FDD, the user needs to specify a probability threshold for a minority group without explicitly identifying all possible groups. The goal is to ensure that all groups with at least a specified occurrence probability have low error.
2. The second method DAW-FDD (demographic-aware FDD) leverages demographic information and employs an existing fairness risk measure [22]. 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 deepfake detector is not more accurate on one group vs another) and, moreover, that the losses across all groups are low. This approach requires a way to estimate the loss of each group, for which we use a ranking-based estimator that addresses the imbalance in real vs deepfake examples per group.

[*] Equal contribution
[†] Corresponding authors

From a technical viewpoint, both of our methods are based on a distributionally robust optimization (DRO) technique called *Conditional Value-at-Risk* (CVaR) [23–25]. Whereas our first method DAG-FDD is a straightforward application of CVaR to the fair deepfake detection problem (so that the novelty is not in the method itself but in applying the method to a problem that we do not believe has previously been explored by DRO literature), our second method DAW-FDD uses the CVaR in a hierarchical manner that, to the best of our knowledge, is novel. Specifically, DAW-FDD uses a CVAR 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 deepfake training examples). We also show how several existing fairness approaches are special cases of DAW-FDD.

Our main contributions are as follows:

1. We propose two methods for achieving fair deepfake detection in ways that are either agnostic to or, separately, aware of demographic factors. Both methods convert an existing deep-learning-based deepfake detector that does not encourage fairness into one that does. Moreover, both use training procedures that alternate between minibatch gradient descent (to update neural network model parameters) and solving specific convex optimization problems related to data imbalance.
2. We demonstrate the effectiveness of our methods in improving fairness of several state-of-the-art deepfake detectors (while retaining strong detection performance) on four large-scale datasets (FaceForensics++ [26], Celeb-DF [27], DeepFakeDetection (DFD) [28], and Deepfake Detection Challenge (DFDC) [29]).

To the best of our knowledge, our paper is the first to propose novel algorithms for fair deepfake detection.

2. Related Work

2.1. The Categories of Fairness Approaches

Many approaches have been proposed to encourage fairness in general machine learning settings. These methods fall into two major categories: demographic-agnostic and demographic-aware. Typically, there is a tradeoff between encouraging fairness and achieving high prediction accuracy.

Demographic-agnostic. When demographic information is inaccessible (*e.g*., either it was not collected, or we do not have an exhaustive list of all groups that we want to be "fair" across), there are methods that achieve fairness without any prior knowledge of which attributes to treat as sensitive. Examples of such approaches include distributionally robust optimization (DRO) [30], adversarial learning [31], using input features to find surrogate group information [32], cluster-based balancing for input data [33], knowledge distillation [34], and causal variational autoencoders [35]. Among these, our work builds on existing DRO literature.

Demographic-aware. A large number of fairness definitions have been proposed in the literature for generating regularization terms to add to a training loss. There are two key types of fairness measures using demographic information that we highlight: group fairness [36] and intersectional fairness [37]. Specifically, group fairness considers a model fair across a user-specified set of groups if these different groups satisfy a condition such as demographic parity [36] or equalized odds [38]. Intersectional fairness accounts for multiple sensitive attributes (*e.g*., intersections of race and gender taking on specific combinations). More notions of fairness can be found in [21, 39]. A drawback of these approaches is that precisely which notion of fairness and which attributes to treat as sensitive (or an exhaustive list of groups to encourage fairness across) must be specified as part of the overall training loss function. If at a later time, we realize that we want to use a different notion of fairness or we want to account for different sensitive attributes or demographic groups, then model re-training may be required. One of our proposed approaches requires an exhaustive list of all groups that we want similar accuracy for.

Method	Year	#Detector	#Dataset	Fairness Solution	Require Demographics
Trinh *et al*. [14]	2021	3	1	×	-
Hazirbas *et al*. [19]	2021	5	1	×	-
Pu *et al*. [44]	2022	1	1	×	-
GBDF [16]	2022	5	4	Data-level	✓
Xu *et al*. [15]	2022	3	4	×	-
DAG-FDD (ours)	2023	5	4	Algorithm-level	×
DAW-FDD (ours)	2023	5	4	Algorithm-level	✓

Table 1. *Summary of previous studies and our work. '-' means not applicable.*

2.2. Fairness in Deepfake Detection

Despite considerable efforts [6, 40–43] dedicated to enhancing the generalization capability of deepfake detection to out-of-distribution (OOD) data, there is still limited progress in addressing the biased performance during testing within known domains. In contrast to previous studies, our research uniquely prioritizes fairness as its primary goal. Specifically, we address the issue of biased performance among groups under in-domain testing, aiming to achieve equal accuracy across user-specified demographic groups.

Extending the investigation of fairness that was originally for face recognition [45–48], several recent studies have examined fairness concerns in deepfake detection, as shown in Table 1. The work in [14] is the first to evaluate biases in existing deepfake datasets and detection models across protected groups. They examined three popular deepfake detectors and observed large disparities in prediction accuracy across races, with up to 10.7% difference in error rate between groups. Similar observations are found in [19]. Pu *et al*. [44] evaluated the reliability of one popular deepfake detection model (MesoInception-4) on FF++ and showed that the MesoInception-4 model is generally more effective for female subjects. A more comprehensive analysis of deepfake detection bias with regards to demographic and non-demographic attributes is presented in [15]. The authors collected comprehensive annotations for 5 widely-used deepfake detection datasets to facilitate future research. The work in [16] showed significant bias in both datasets and detection models and they tried to reduce the performance bias across genders by providing a gender-balanced dataset. This leads to limited improvement at the cost of highly time-consuming data annotation, which does not extend to other possible non-gender attributes that we might want to treat as sensitive. Developing more effective bias-mitigating deepfakes detection solutions remains an open challenge [2].

3. Method

In this work, we propose two deep-learning-based deepfake detection methods that encourage fairness. The first method, termed DAG-FDD, is applicable when we have training data without demographic annotations. This approach works with most existing deepfake datasets. The second method, termed DAW-FDD, works when the dataset contains additional demographic annotations (specifically so that we know which group each training point belongs to among some user-specified exhaustive list of groups we aim to ensure fairness over).

Both approaches are meant to modify an existing deep-learning-based deepfake detector into one that encourages fairness. To this end, in what follows, we assume that $\mathcal{S} := \{(X_i, Y_i)\}_{i=1}^n$ is the training set that consists of i.i.d. samples from a joint distribution $\mathbb{P}$, where X_i is the i-th data point's raw features (*e.g.*, an image or video) and $Y_i \in \{0, 1\}$ is the i-th point's label (0 means real, 1 means deepfake). We assume that the underlying deepfake detector aims to minimize a risk of the form

$$\mathcal{R}_{\text{avg}}(\theta) := \mathbb{E}_{(X,Y)\sim\mathbb{P}}[\ell(\theta; X, Y)] \quad \text{for } \theta \in \Theta, \tag{1}$$

where ℓ is the loss function (*e.g.*, cross entropy loss) of the deepfake detector model, which is assumed to have parameters θ that belong to a set Θ; the loss function is evaluated for a specific input X with target label Y. As is standard in machine learning, instead of minimizing the true unknown risk $\mathcal{R}_{\text{avg}}(\theta)$, in practice we use some variant of minibatch gradient descent to minimize the empirical risk given by the loss function $\mathcal{L}_{\text{avg}}(\theta) := \frac{1}{n}\sum_{i=1}^n \ell(\theta, X_i, Y_i)$.

3.1. Demographic-agnostic FDD (DAG-FDD)

We first present DAG-FDD, which is based on the distributionally robust optimization (DRO) [30, 49]. Roughly, the idea is that there are K unknown underlying groups of individuals. We assume that each group occurs with probability at least $\alpha \in (0, 1)$. Then by a standard result of DRO, there is a loss function that we can minimize that aims to ensure that all K latent groups have low error despite us not explicitly knowing what these latent groups are. We formalize this high level idea in the rest of this section.

The worst-case risk $\mathcal{R}_{\max}(\theta)$. We assume that there are K true unknown groups that comprise the joint distribution $\mathbb{P}$. In other words, $\mathbb{P}$ can be represented as a mixture of K distributions $\mathbb{P} := \sum_{m=1}^K \pi_m \mathbb{P}_m$, where the m-th group occurs with probability $\pi_m \in (0, 1)$ and has distribution $\mathbb{P}_m$, and $\sum_{m=1}^K \pi_m = 1$. Instead of minimizing the average risk (1), we seek to minimize the following worst-case risk:

$$\mathcal{R}_{\max}(\theta) := \max_{m=1,\ldots,K} \mathbb{E}_{(X,Y)\sim\mathbb{P}_m}[\ell(\theta; X, Y)], \tag{2}$$

where ℓ is a loss function for an individual data point used by the original deepfake detector that we are modifying (*i.e.*, ℓ is same function used in equation (1)). Directly minimizing $\mathcal{R}_{\max}$ is intractable since we do not know the K latent groups; in fact, we assume that we do not know the value of K either. However, it turns out that we can minimize an empirical version of its upper bound.

Upper bound on $\mathcal{R}_{\max}(\theta)$. We use the well-established risk function called the *Conditional Value-at-Risk* (CVaR) [25]:

$$\text{CVaR}_\alpha(\theta) := \inf_{\lambda\in\mathbb{R}} \left\{ \lambda + \frac{1}{\alpha}\mathbb{E}_{(X,Y)\sim\mathbb{P}}\big[[\ell(\theta; X, Y) - \lambda]_+\big] \right\}, \tag{3}$$

where $[a]_+ = \max\{0, a\}$ is the hinge function (also called the ReLU function), and we assume that each of the K latent groups occurs with probability at least $\alpha \in (0, 1)$. The following result shows that the risk $\text{CVaR}_\alpha(\theta)$ is an upper bound of $\mathcal{R}_{\max}(\theta)$, so that by minimizing $\text{CVaR}_\alpha(\theta)$, we are minimizing an upper bound on the worst-case risk (2).

Proposition 1. *Suppose that* $\alpha \le \min_{m=1,\ldots,K} \pi_m$*. Then* $\text{CVaR}_\alpha(\theta) \ge \mathcal{R}_{\max}(\theta)$.

Note that this result is not new [50]. However, to the best of our knowledge, we are the first to apply it to learning fair deepfake detectors in a demographic-agnostic way. We include the proof of Proposition 1 in Appendix A.1.

In practice, α is a user-specified hyperparameter that says how rare of a group we want to ensure low risk for. As $\alpha \to 0$, we are asking for low risk even for an extremely rare group. In contrast, as $\alpha \to 1$ (*i.e.*, the rarest group occurs with probability 1), then for Proposition 1 to hold, it means that we would have $K = 1$ and $\pi_1 = 1$, in which case the worst-case risk (2) would simply become the standard average risk (1). By tuning $\alpha \in (0, 1)$, we effectively say how "fine-grain" of groups we aim to encourage fairness over, which naturally leads to a tradeoff between fairness and population-level average accuracy.

DAG-FDD. In practice, we minimize an empirical version of $\text{CVaR}_\alpha(\theta)$. This gives us the following optimization problem, which we refer to as our first method DAG-FDD:

$$\min_{\theta\in\Theta,\lambda\in\mathbb{R}} \mathcal{L}_{\text{DAG-FDD}}(\theta, \lambda) := \lambda + \frac{1}{\alpha n}\sum_{i=1}^n [\ell(\theta; X_i, Y_i) - \lambda]_+. \tag{4}$$

As a reminder, Θ is the set of possible model parameters of the original deepfake detector (see equation (1)).

To provide some intuition for the loss function $\mathcal{L}_{\text{DAG-FDD}}(\theta, \lambda)$, suppose for a moment that we have obtained

Algorithm 1: DAG-FDD

```
Input: A training dataset $\mathcal{S}$ of size $n$, $\alpha$, max_iterations,
       num_batch, learning rate $\eta$
Output: A fair deepfake detection model with parameters $\theta^*$
Initialization: $\theta_0, l = 0$
for $e = 1$ to max_iterations do
    for $b = 1$ to num_batch do
        Sample a mini-batch $\mathcal{S}_b$ from $\mathcal{S}$
        Compute $\ell(\theta_l; X_i, Y_i), \forall (X_i, Y_i) \in \mathcal{S}_b$
        Using binary search to find $\lambda$ that minimizes (4) on $\mathcal{S}_b$
        Compute $\theta_{l+1}$ with equation (5)
        $l \leftarrow l + 1$
    end
end
return $\theta^* \leftarrow \theta_l$
```

the optimal value of λ^* in (4). Then the only training points that contribute to the loss are the "hard" ones with a loss value greater than λ^*. In other words, the loss function always focuses on "hard" training points with large enough loss values whereas the "easy" training points with low loss values are ignored. Which training points are "easy" vs "hard" can vary as a function of model parameters $\theta \in \Theta$.

Solving the optimization problem in (4) can be done through an iterative gradient descent approach [51–54]. In practice, we first initialize model parameters θ and then randomly select a mini-batch set $\mathcal{S}_b$ from the training set $\mathcal{S}$, performing the following two steps for each iteration on $\mathcal{S}_b$ (see Algorithm 1):

- We fix θ and use binary search to find the global optimum of λ since $\mathcal{L}_{\text{DAG-FDD}}(\theta, \lambda)$ is convex w.r.t. λ.
- We fix λ and update θ using (stochastic) gradient descent with a user-specified learning rate $\eta > 0$:

$$\theta_{l+1} = \theta_l - \frac{\eta}{\alpha|\mathcal{S}_b|} \sum_{i \in \mathcal{S}_b} \partial_\theta \ell(\theta_l; X_i, Y_i) \cdot \mathbb{1}_{[\ell(\theta_l; X_i, Y_i) > \lambda]}, \tag{5}$$

  where $\mathbb{1}_{[a]}$ is an indicator function (that equals 1 if a is true, and 0 otherwise), $\partial_\theta \ell$ represents the (sub)gradient of ℓ w.r.t. θ, and η is the learning rate.

We stop iterating after reaching some user-specified stopping criteria (*e.g.*, maximum number of iterations). Note that this optimization process is similar to how one would train the original deepfake detector being modified except that we now have an additional binary search to update λ; thus the training time complexity is comparable.

3.2. Demographic-aware FDD (DAW-FDD)

We now turn to the setting where within the training data, we have demographic labels available so that we know for each training point which group it belongs to among some user-specified exhaustive listing of all possible groups that we want to ensure fairness across. Specifically, we let $\mathcal{G}$ denote the user-specified set of groups (*e.g.*, if we aim to encourage fairness across gender, then $\mathcal{G}$ would consist of the different genders). Then for the i-th training point (with raw features X_i and target label Y_i) we assume that we also know its group $G_i \in \mathcal{G}$.

We first state the group-level risk that we aim to minimize that encourages fairness across groups (*i.e.*, this risk aims to address imbalance in user-specified demographic groups within the data). When it comes to empirically estimating this risk, we then discuss how we account for imbalance in real vs deepfake examples.

Group-level risk (addresses demographic imbalance). Each group $g \in \mathcal{G}$ has a group-specific risk defined as $\mathcal{R}_g(\theta) := \mathbb{E}_{(X,Y)|G=g}[\ell(\theta; X, Y)]$, where random variable G denotes the group corresponding to a generic data point with raw features X and target Y. To treat the different groups to be "equally weighted", the risk we use intentionally views G to be sampled uniformly at random from $\mathcal{G}$ (even if in the actual data, G may not be uniformly distributed so that different groups could occur with different probabilities). Specifically, we use the "group CVaR" risk

$$\text{CVaR}_\alpha^{\mathcal{G}}(\theta) := \inf_{\lambda \in \mathbb{R}} \left\{ \lambda + \frac{1}{\alpha} \mathbb{E}_{G \sim \text{Uniform}(\mathcal{G})} \left[[\mathcal{R}_G(\theta) - \lambda]_+ \right] \right\}. \tag{6}$$

To provide some intuition for this risk, recall that the non-group-level CVaR risk from earlier (equation (3)) focuses on individual data points that have "large enough" loss values (specifically, if λ^* achieves the infimum value in equation (3), then $\text{CVaR}_\alpha(\theta)$ only focuses on data points with loss values above λ^*). In the group-level version of CVaR presented in equation (6), we instead focus on groups that have "large enough" risk values.

The group CVaR risk as a fairness risk measure. In fact, the group CVaR risk can be directly interpreted as a fairness risk measure, as shown by [22].

Proposition 2. *(Equation (21) of [22]) Let $\alpha \in (0, 1)$,*

$$\min_{\theta \in \Theta} \text{CVaR}_\alpha^{\mathcal{G}}(\theta) = \min_{\theta \in \Theta} \left\{ \mathbb{E}_{G \sim \textit{Uniform}(\mathcal{G})}[\mathcal{R}_G(\theta)] + \mathbb{D}(\{\mathcal{R}_g(\theta) : g \in \mathcal{G}\}) \right\},$$

where $\mathbb{D}$ is a "deviation measure" that looks at how different the different groups' losses are (if they are all the same, then the deviation measure would be 0). Specifically,

$$\mathbb{D}(\{\mathcal{R}_g(\theta) : g \in \mathcal{G}\}) := \inf_{\lambda \in \mathbb{R}} \left\{ \lambda + \frac{1}{\alpha|\mathcal{G}|} \sum_{g \in \mathcal{G}} [\mathcal{R}_g(\theta) - \overline{\mathcal{R}}(\theta) - \lambda]_+ \right\},$$

where $\overline{\mathcal{R}}(\theta) := \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathcal{R}_g(\theta)$.

This proposition states that minimizing the group-level CVaR risk is equivalent to minimizing a risk that is the sum of two terms: the first term $\mathbb{E}_{G \sim \text{Uniform}(\mathcal{G})}[\mathcal{R}_G(\theta)] = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathcal{R}_g(\theta)$ is the equally weighted average of the different groups' losses, and the second term $\mathbb{D}(\{\mathcal{R}_g(\theta) : g \in \mathcal{G}\})$ asks that the different groups' losses are close to each other (*i.e.*, the deepfake detector learned should not be more accurate for one group vs another).

DAW-FDD (empirical estimation of group-level risk that accounts for imbalance in real vs deepfake examples). Recall that previously when we presented our demographic agnostic approach DAG-FDD, we empirically estimate the CVAR risk from equation (3) in a straightforward manner with the loss function $\mathcal{L}_{\text{DAG-FDD}}(\theta, \lambda)$ from equation (4). Now that we use a group-level CVAR risk instead (given in equation (6)), we have to be more careful with empirically estimating the risk. The issue is that we need an accurate estimate of each group's risk $\mathcal{R}_g(\theta) = \mathbb{E}_{(X,Y)|G=g}[\ell(\theta; X, Y)]$. A naive approach would take an equally weighted average across examples belonging to group g. However, the imbalance in the number of real vs deepfake examples in group g could bias the estimate of $\mathcal{R}_g(\theta)$.

To address this issue, we use the average top-k operator [55] to estimate group risks instead of the average operator. In more detail, denote the training points in group $g \in \mathcal{G}$ as $\mathcal{I}_g := \{i = 1, \ldots, n : G_i = g\}$, the number of points in group g as $n_g := |\mathcal{I}_g|$, and the set of individual losses in group g as $\ell^g(\theta) := \{\ell(\theta; X_i, X_j) : i \in \mathcal{I}_g\}$. We further denote the j-th largest loss in $\ell^g(\theta)$ as $\ell^g_{[j]}$ (ties can be broken in any consistent manner). Then we empirically estimate group g's risk $\mathcal{R}_g(\theta)$ with the loss function

$$\mathcal{L}_g(\theta) := \frac{1}{k_g} \sum_{j=1}^{k_g} \ell^g_{[j]}(\theta), \tag{7}$$

where $k_g \in \{1, \ldots, n_g\}$ is a user-specified integer. This choice of empirical estimate can enhance the influence of the minority class while reducing the influence of the majority class in each group as samples with small loss values are most likely from the majority class per group. Since k_g may vary across groups, we set $k_g = \alpha_g n_g$, where $\alpha_g \in [1/n_g, 1]$. In practice, we can set α_g to be the same for all groups and tune it on a predefined grid.

Finally, we solve the following optimization problem which minimizes an empirical estimate of $\text{CVaR}^{\mathcal{G}}_{\alpha}(\theta)$:

$$\min_{\theta\in\Theta, \lambda\in\mathbb{R}} \mathcal{L}_{\text{DAW-FDD}}(\theta, \lambda) := \lambda + \frac{1}{\alpha|\mathcal{G}|} \sum_{g\in\mathcal{G}} [\mathcal{L}_g(\theta) - \lambda]_+. \tag{8}$$

As it turns out, the group specific loss function $\mathcal{L}_g(\theta)$ in equation (7) can itself be written as a CVaR loss, as we establish in the following theorem.

Theorem 1. *For a set of real numbers $\bar{\ell} = \{\ell_1, ..., \ell_q\}$, let $\ell_{[k]}$ denote the k-th largest value in $\bar{\ell}$ for $k \in \{1, \ldots q\}$. Then we have $\frac{1}{k}\sum_{i=1}^{k} \ell_{[i]} = \min_{\lambda\in\mathbb{R}}\{\lambda + \frac{1}{k}\sum_{i=1}^{q}[\ell_i - \lambda]_+\}$. Using this result, optimization problem (8) is equivalent to*

$$\min_{\theta\in\Theta, \lambda\in\mathbb{R}} \mathcal{L}_{\text{DAW-FDD}}(\theta, \lambda) := \lambda + \frac{1}{\alpha|\mathcal{G}|} \sum_{g\in\mathcal{G}} [\mathcal{L}_g(\theta) - \lambda]_+, \tag{9a}$$

$$\text{s.t. } \mathcal{L}_g(\theta) = \min_{\lambda_g\in\mathbb{R}} \mathcal{L}_g(\theta, \lambda_g) := \lambda_g + \frac{1}{\alpha_g n_g} \sum_{i\in\mathcal{I}_g} [\ell(\theta; X_i, Y_i) - \lambda_g]_+. \tag{9b}$$

We defer the proof to Appendix A.2. Theorem 1 tells us that optimization problem (8) is equivalent to a optimization problem with a hierarchical structure: across the demographic groups, we have a group-level CVaR loss (equation (9a)). To compute this group-level CVaR loss, we compute each group's loss function $\mathcal{L}_g(\theta)$ (equation (9b)), which in turn is of the form of a CVaR loss (*i.e.*, the top-k operator can be written as a CVaR loss). This CVaR loss per group is specifically meant for addressing the imbalance in real vs deepfake examples. We call this approach DAW-FDD.

To optimize (9), the iterative procedure in Algorithm 1 can still be applied except where each iteration now consists of three steps: updating $\{\lambda_g : g \in \mathcal{G}\}$, λ, and θ. The pseudocode is shown in Algorithm 2 in Appendix B. Note that the explicit form of $\partial_\theta \mathcal{L}_{\text{DAW-FDD}}(\theta_l, \lambda)$ (*i.e.*, the (sub) gradient of $\mathcal{L}_{\text{DAW-FDD}}(\theta, \lambda)$ w.r.t. θ) can be found in Appendix C.

Remark 1. *For our method* DAW-FDD*, by choosing values of α and α_g in specific ways, we recover several existing fairness methods. For example, if $\alpha \to 1$ and $\alpha_g \to 1$, we minimize the average of group risks, which aligns with the impartial observer principle [56]. If $\alpha \to 0$ and $\alpha_g \to 1$, we instead minimize the largest group risk (2) [30, 57], which aligns with the maximin principle [58]. If $\alpha_g \to 1$ (*i.e., *we replace the top-k operator with a simple average), our approach is just the empirical version of CVaR$^{\mathcal{G}}_{\alpha}(\theta)$ [22].*

4. Experiment

This section evaluates the effectiveness of the proposed methods in terms of fairness performance and deepfake detection performance. We present the most significant information and results of our experiments. More detailed information and additional results are provided in Appendix D and E, respectively.

4.1. Experimental Settings

Datasets. Our experiments are based on four popular large-scale benchmark deepfake datasets, namely FaceForensics++ (FF++) [26], Celeb-DF [27], DeepFakeDetection (DFD) [28], and Deepfake Detection Challenge Dataset (DFDC) [29]. Since all the original datasets do not have the demographic information of each video or image, we use the annotations from [15] which provides annotated demographic information for these four datasets, including Gender (Male and Female) and Race (Asian, White, Black, and Others) attributes. We also double-check the annotations for each dataset. In addition to the single attribute fairness, we also consider the combined attributes (Intersection) group, including Male-Asian (M-A), Male-White (M-W), Male-Black (M-B), Male-Others (M-O), Female-Asian (F-A), Female-White (F-W), Female-Black (F-B), and Female-Others (F-O). We use Dlib [61] for face extraction and alignment, and the cropped faces are resized to 380×380 for training and testing. Following the previous study [15],

Methods	Require Demo-graphics	Fairness Metrics (%) ↓									Detection Metrics (%)			
		Gender			Race			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Original	–	4.10	4.10	9.06	13.09	17.28	21.00	17.93	31.59	53.95	92.76	22.06	94.43	91.49
DRO$_{\chi^2}$ [30]	×	2.68	2.68	6.75	8.32	8.97	20.40	8.73	22.97	55.54	97.18	6.32	90.25	90.86
DAG-FDD (Ours)		1.63	1.63	6.21	8.23	9.53	11.49	9.65	21.21	48.10	97.13	9.54	94.32	93.63
Naive [16]	✓	11.98	11.98	18.20	16.57	22.01	25.97	28.90	72.19	93.59	83.17	50.77	92.62	84.87
FRM [22]		1.33	1.33	5.88	9.24	12.75	20.13	10.39	25.57	60.90	**97.81**	**4.76**	90.85	91.63
Group DRO [59]		8.20	8.20	12.87	14.37	20.97	23.79	21.86	44.98	65.24	91.13	27.83	95.15	91.04
$Cons.$ EFPR [60]		4.24	4.24	7.91	7.09	7.49	12.41	14.95	27.80	46.62	94.30	22.61	94.94	91.80
$Cons.$ EO [60]		1.77	1.77	4.79	10.92	12.61	16.50	17.25	26.95	44.68	95.74	16.28	**95.89**	93.72
DAW-FDD (Ours)		**0.32**	**0.32**	**3.99**	**2.49**	**3.88**	**6.29**	**6.61**	**14.06**	**33.84**	97.46	11.46	95.40	**94.17**

Table 2. *Comparison results with different fairness solutions using Xception detector on FF++ testing set across Gender, Race, and Intersection groups. The best results are shown in **Bold**. ↑ means higher is better and ↓ means lower is better. Gray highlights mean our methods outperform the baselines in the group* (i.e., *DAG-FDD vs. Original/DRO$_{\chi^2}$, DAW-FDD vs. Original/Naive/FRM/Group DRO/$Cons.$ EFPR/$Cons.$ EO).*

Methods	Fairness Metrics (%) ↓			Detection Metrics (%)				Training Time (mins)/Epoch	Binary Search Time (mins)/Epoch
	Intersection			Overall					
	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑		
Original	24.00 (9.00)	45.50 (16.39)	63.24 (12.96)	95.00 (2.96)	19.28 (7.14)	**95.94** (1.41)	93.29 (1.60)	2.6	N/A
DAG-FDD (Ours)	13.83 (11.86)	**24.38** (17.04)	48.52 (15.37)	96.81 (1.68)	13.43 (7.00)	95.33 (1.40)	93.81 (1.13)	3.0	0.59
DAW-FDD (Ours)	**11.53** (3.43)	26.55 (7.97)	**47.50** (10.99)	**97.40** (0.30)	**12.21** (4.05)	95.45 (1.37)	**94.11** (0.66)	3.0	0.66

Table 3. *Detection mean and standard deviation (in parentheses) of Xception detector on FF++ testing set across 5 experimental repeats, in the same format as Table 2. Training time and binary search time per epoch for each method are also reported.*

we split the annotated datasets into training/validation/test sets with a ratio of approximately 60%/20%/20%, without identity overlapping. In particular, the validation set is used for hyperparameter tuning. More details of the datasets, including attribute groups and number of training samples are provided in Tables E.5 and E.6 of the Appendix E.7.

Evaluation metrics. Fairness measures are selected considering the practical use of deepfake detection systems in social media. Given that real cases outnumber (deep)fake ones, we prioritized metrics related to False Positives (misclassifying real as fake) to prevent potential consequences such as suspicion, distrust, legal, or social repercussions, especially for users from specific ethnic groups. Three fairness metrics are used to report the fairness performance of methods. Specifically, we report the maximum differences in False Positive Rate (FPR) Gap (G_{FPR}) for Gender, Race, and Intersection groups. We also consider the Equal False Positive Rate (F_{FPR}) and Equal Odds (F_{EO}) metrics as used in [60]. These metrics are defined as follows (for ease of notation, we write this for the training data but it is of course evaluated on test data):

$$
\begin{aligned}
G_{\text{FPR}} &:= \max_{g,g'\in\mathcal{G}} \left|\text{FPR}_g - \text{FPR}_{g'}\right|,\\
F_{\text{FPR}} &:= \sum_{g\in\mathcal{G}} \left| \frac{\sum_{i=1}^n \mathbb{1}_{[\hat{Y}_i=1,G_i=g,Y_i=0]}}{\sum_{i=1}^n \mathbb{1}_{[G_i=g,Y_i=0]}} - \frac{\sum_{i=1}^n \mathbb{1}_{[\hat{Y}_i=1,Y_i=0]}}{\sum_{i=1}^n \mathbb{1}_{[Y_i=0]}} \right|,\\
F_{\text{EO}} &:= \sum_{g\in\mathcal{G}} \sum_{q=0}^{1} \left| \frac{\sum_{i=1}^n \mathbb{1}_{[\hat{Y}_i=1,G_i=g,Y_i=q]}}{\sum_{i=1}^n \mathbb{1}_{[G_i=g,Y_i=q]}} - \frac{\sum_{i=1}^n \mathbb{1}_{[\hat{Y}_i=1,Y_i=q]}}{\sum_{i=1}^n \mathbb{1}_{[Y_i=q]}} \right|,
\end{aligned}
\tag{10}
$$

where FPR_g represents the FPR scores of group g. Y_i and $\hat{Y}_i$ respectively represent the true and predicted labels of the sample X_i. Their values are binary, where 0 means real and 1 means fake. For all fairness metrics here, lower is better.

Since there is usually a trade-off between fairness and detection performance [21, 39], we also include detection metrics to assess the balance between fairness and detection performance. Four widely-used deepfake detection metrics are reported: 1) the area under the curve (AUC), 2) FPR, which is essential in real-world use as it indicates the count of incorrect fake classifications, 3) True Positive Rate (TPR), which measures the number of correct fake classification, and 4) Accuracy (ACC). We calculate the FPR, TPR, and ACC with a fixed threshold of 0.5 [62, 63].

Baseline methods. We apply our proposed methods DAG-FDD and DAW-FDD in Section 3 to popular deepfake detectors to show their effectiveness. Five deepfake detection models are considered, including three widely-used CNN architectures in deepfake detection [16, 27, 64] (*i.e.*, Xception [26], ResNet-50 [65], and EfficientNet-B3 [66]) and two well-designed deepfake detectors with outstanding performance, namely DSP-FWA [27] and RECCE [67]. We denote the detectors with their original loss functions (*e.g.*, binary cross-entropy) as "Original".

In terms of comparison in fairness detection, we consider the method [16] based on balancing the number of training samples in each group for deepfake fairness improvement; we take this method as a baseline for comparison denoted as "Naive". Specifically, we use an intersectional group with the smallest number of training samples and then randomly select the same number of training samples from the other groups to create such a balanced training dataset. Moreover, we compare our two loss functions with the χ^2-divergence based DRO (DRO$_{\chi^2}$) [30], the fairness risk measure (FRM) [22], and a popular Group DRO method [59] in fairness research although they have not been applied to deepfake detection. Besides, we modify F_{FPR} and F_{EO} as regularization terms [60], and incorporate them with binary cross-entropy loss as baselines: $Cons.$ EFPR and $Cons.$ EO.

Implementation details. All experiments are conducted on the PyTorch platform [68] using 4 NVIDIA RTX A6000

Datasets	Methods	Fairness Metrics (%) ↓									Detection Metrics (%)			
		Gender			Race			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
Celeb-DF	Original	4.93	4.93	22.04	3.31	4.77	**26.06**	11.81	15.66	39.95	97.17	13.01	**95.83**	**94.05**
	DAG-FDD (Ours)	**2.02**	**2.02**	**16.77**	**1.20**	**1.22**	28.56	**2.54**	**3.09**	**30.43**	98.00	2.42	87.40	89.44
	DAW-FDD (Ours)	3.81	3.81	18.93	3.14	3.34	33.91	3.80	4.91	35.48	**98.03**	**2.10**	84.53	87.21
DFD	Original	2.95	2.95	5.52	7.35	7.35	7.72	8.67	15.81	24.31	92.94	**25.00**	96.01	**89.09**
	DAG-FDD (Ours)	2.92	2.92	4.79	6.08	6.08	7.05	8.30	13.52	19.57	**93.40**	28.07	**96.31**	88.28
	DAW-FDD (Ours)	**1.40**	**1.40**	**3.14**	**2.36**	**2.36**	**3.35**	**7.20**	**8.74**	**14.70**	93.17	27.75	95.95	88.14
DFDC	Original	1.64	1.64	4.36	4.02	5.85	**38.84**	20.17	38.68	119.71	92.40	7.28	**76.32**	86.87
	DAG-FDD (Ours)	**1.30**	**1.30**	5.38	4.50	5.78	46.56	14.65	33.79	**113.93**	92.69	6.61	74.41	86.61
	DAW-FDD (Ours)	1.73	1.73	**3.39**	**3.48**	**4.14**	42.87	**11.19**	**23.63**	115.15	**94.88**	**4.27**	75.10	**88.37**

Table 4. *Results of Xception detector on Celeb-DF, DFD, and DFDC testing sets, in the same format as Table 2.*

Models	Methods	Fairness Metrics (%) ↓									Detection Metrics (%)			
		Gender			Race			Intersection			Overall			
		G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	G_{FPR}	F_{FPR}	F_{EO}	AUC ↑	FPR ↓	TPR ↑	ACC ↑
ResNet-50	Original	2.58	2.58	6.64	12.80	15.45	17.62	20.06	43.18	60.66	94.32	25.96	**96.36**	**92.37**
	DAG-FDD (Ours)	**2.21**	**2.21**	**6.37**	**6.44**	11.42	14.17	16.68	37.50	**52.79**	**94.51**	**22.52**	95.34	92.15
	DAW-FDD (Ours)	3.78	3.78	10.21	7.01	**8.27**	**13.09**	**13.28**	**35.08**	60.07	93.70	23.56	93.65	90.58
EfficientNet-B3	Original	1.97	1.97	**4.15**	9.05	10.86	14.12	13.38	22.65	**40.13**	95.91	20.25	**97.21**	**94.09**
	DAG-FDD (Ours)	0.47	0.47	5.36	9.48	9.58	13.50	10.87	19.34	46.08	**97.20**	8.40	92.87	92.65
	DAW-FDD (Ours)	**0.04**	**0.04**	5.53	**3.79**	**4.67**	**12.63**	**6.43**	**12.57**	43.72	96.30	**8.22**	91.43	91.49
DSP-FWA	Original	5.90	5.90	11.81	11.07	14.58	21.98	21.38	48.20	75.91	**91.79**	31.64	93.17	88.74
	DAG-FDD (Ours)	4.64	4.64	**9.77**	12.52	18.04	25.03	15.61	40.57	**74.54**	91.47	32.35	**93.70**	**89.05**
	DAW-FDD (Ours)	**3.02**	**3.02**	11.30	**5.75**	**10.52**	**19.34**	**12.84**	**36.05**	75.73	90.84	**30.43**	91.97	87.97
RECCE	Original	0.87	0.87	**3.14**	18.81	27.65	30.07	30.26	67.38	80.34	98.05	21.20	**98.21**	94.74
	DAG-FDD (Ours)	0.55	0.55	3.71	12.68	17.41	20.33	15.40	36.17	54.24	98.33	12.01	96.80	**95.23**
	DAW-FDD (Ours)	**0.25**	**0.25**	4.75	**6.99**	**7.96**	**11.95**	**13.54**	**23.44**	**52.95**	**98.35**	**8.15**	94.59	94.10

Table 5. *Results of ResNet-50, EfficientNet-B3, DSP-FWA, and RECCE detectors on FF++ testing set, in the same format as Table 2.*

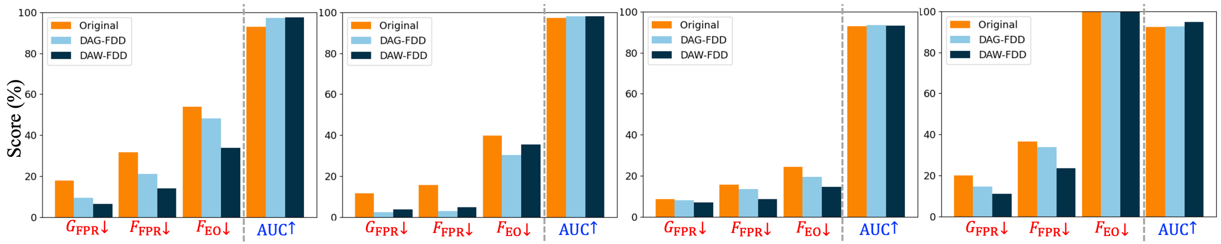
GPU cards. We train all methods by using a (mini-batch) stochastic gradient descent optimizer with batch size 640, epochs 200, and learning rate as 5×10^{-4}. We build our loss functions on the binary cross-entropy loss for the binary deepfake classification task. Since the DAW-FDD method needs to pre-define a set of groups, we use the Intersection group in experiments and also report the performance on single attributes. The hyperparameters α and α_g are tuned on the grid $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Following [69], the final hyperparameter setting per dataset and per method is determined based on a preset rule that allows up to a 5% degradation of overall AUC in the validation set from the corresponding "Original" method while minimizing the intersectional F_{FPR}. More details and the evaluation of the influence of different parameter settings on detection performance are provided in Appendix D.1, E.2.

4.2. Results

Performance on FF++ dataset. We first report results of our methods compared with several baselines on the FF++ dataset using the Xception deepfake detector in Table 2. These results show that in the cases where demographic information is unavailable from the training data, our DAG-FDD method achieves superior fairness performance to the Original method across most metrics for all three sensitive attribute groups, as shown in gray highlights. For example, we enhance the G_{FPR} of Gender, F_{EO} of Race, and F_{FPR} of Intersection by 2.47%, 9.51%, and 10.38%, respectively, compared to the Original. These results indicate our method has strong applicability in scenarios where demographic data is unavailable. In addition, it is clear that our method outperforms the DRO$_{\chi^2}$ method on most fairness metrics. This is benefited from the tighter upper bound (see Proposition 1) on the risk $\mathcal{R}_{\max}(\theta)$ in our DAG-FDD method than the DRO$_{\chi^2}$ method as mentioned in [50].

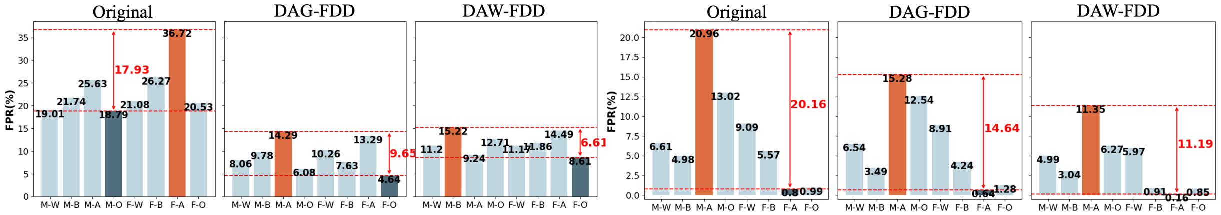
With demographic information, we see that the Naive method trained on a balanced dataset does not guarantee an improvement in fairness metrics on test data. For example, on the intersectional groups, all fairness scores of the Naive method (*e.g.*, F_{FPR}: 72.19%) are worse than the Original method (*e.g.*, F_{FPR}: 31.59%). This can be attributed to the fact that a naive balancing strategy will reduce the number of available training samples, resulting in a significant decrease in detection performance. The same trends can be found in AUC scores, which decrease from 92.76% (Original) to 83.17% (Naive), and in FPR scores, which increase from 22.06% (Original) to 50.77% (Naive). Thus, a poorly trained model on balanced data could result in worse fairness scores.

Our DAW-FDD method outperforms all methods on the most fairness metrics (as shown in Bold). The reason is that DAW-FDD uses the additional demographic information to guide training to achieve fairness without reducing the dataset size. In particular, the DAW-FDD method achieves the best fairness performance on all metrics in the Intersection group thanks to the guidance of intersection group information in the design of DAW-FDD. The superiority of DAW-FDD over Group DRO is evident. Specifically, Group DRO does not show any improvement, possibly because it places greater emphasis on improving the worst-group generalization performance and less on ensuring overall fairness. Furthermore, in our comparison between the DAW-FDD and FRM methods, we have found our method outperforms FRM. This result clearly demonstrates the effectiveness of our learning strategy that considers two types of imbalance



(a) FF++ (b) Celeb-DF (c) DFD (d) DFDC

Figure 1. *Comparison of Xception detector on Intersection group of four datasets: (a) FF++, (b) DFDC, (c) DFD, and (d) DFDC.*



(a) FF++ (b) DFDC

Figure 2. *FPR comparison of Xception detector on intersectional groups of (a) FF++ and (b) DFDC dataset.* *Orange* *and* *dark cyan* *bars show the groups with the highest and lowest FPR, respectively, while the double arrow indicates their gap (smaller is better).*

(demographic groups and, separately, real vs deepfake).

Most importantly, our DAG-FDD and DAW-FDD methods not only enhance fairness performance but also improve the detection performance of the detector. For example, we see improvements of approximately 4.7% in AUC and 12.52% in FPR when compared with the Original method. To show the stability of our methods, we run experimental repeats with 5 random seeds as shown in Table 3. It is clear that our methods robustly improve fairness. We also show the training time per epoch costs during training Xception on FF++ dataset in Table 3. Based on the presented table results, our methods show a slightly higher time requirement compared to the original method. However, the difference is minimal, mainly due to the incorporation of a binary search in the calculation of model training loss.

Performance on different datasets. Table 4 shows the evaluation performance of the Xception detector on three popular deepfake datasets. It is clear that our proposed DAG-FDD and DAW-FDD methods outperform the Original method on all three datasets across all groups and most fairness metrics, especially on the Intersection group (also as shown in Figure 1). Moreover, our methods achieve similar or better scores on most detection metrics. Note that our methods on the Celeb-DF dataset lead to a decrease in TPR. One possible reason is that our methods involve hyperparameter tuning based on F_{FPR}, as mentioned in experimental settings. To evaluate the effectiveness of our method on other metrics, we employ F_{EO} as an index to tune the hyperparameter and report the results in Appendix E.1. The results illustrate that optimizing hyperparameters using F_{EO} can improve TPR and F_{EO}. This demonstrates the good flexibility and applicability of our methods to different metrics and datasets. We further show the FPR comparison results on FF++ and DFDC datasets with detailed performance in groups in Figure 2. Our methods evidently narrow the disparity between groups and lower the FPR of each group.

Performance on various detection models. We further evaluate the effectiveness of our methods on four popular deepfake detection models on the FF++ dataset. The results are presented in Table 5. It is clear that our methods can improve the fairness performance of the detectors without significantly decreasing the detection performance. These results indicate that our methods exhibit high scalability and can be seamlessly integrated with different backbones and deepfake detection models.

5. Conclusion

In this work, we propose two methods, DAG-FDD and DAW-FDD, for training fair deepfake detection models in ways that are agnostic to or, separately, aware of demographic factors. Extensive experiments on four large-scale deepfake datasets and five deepfake detectors show the effectiveness of our methods in improving the fairness of existing deepfake detectors.

A limitation of our methods is that they rely on the assumption that loss functions can be decomposed into individual terms and that each instance is independent. Therefore, integrating our methods into graph learning-based detectors may not be straightforward.

In future work, we aim to extend this work in the following areas. First, we will examine the fairness generalization abilities of cross-dataset deepfake detection. Second, we will investigate fairness methods for managing non-decomposable loss-based detectors.

Acknowledgement. This work was supported in part by the US Defense Advanced Research Projects Agency (DARPA) Semantic Forensic (SemaFor) program, under Contract No. HR001120C0123. G. H. Chen is supported by NSF CAREER award #2047981.

References

[1] M. Boháček and H. Farid, "Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms," *Proceedings of the National Academy of Sciences*, vol. 119, no. 48, p. e2216035119, 2022. 1

[2] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, pp. 1–53, 2022. 1, 3

[3] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," *arXiv preprint arXiv:1910.12467*, 2019.

[4] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *Image Analysis and Processing*, pp. 219–229, Springer, 2022.

[5] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proceedings of the International Conference on Multimedia Retrieval*, pp. 615–623, 2022.

[6] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2185–2194, 2021. 2

[7] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, "Sharp multiple instance learning for deepfake video detection," in *Proceedings of the ACM International Conference on Multimedia*, pp. 1864–1872, 2020.

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

[9] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Open-eye: An open platform to study human performance on identifying ai-synthesized faces," in *IEEE International Conference on Multimedia Information Processing and Retrieval*, pp. 224–227, IEEE, 2022.

[10] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Eyes tell all: Irregular pupil shapes reveal GAN-generated faces," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2904–2908, IEEE, 2022.

[11] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Robust attentive deep neural network for detecting GAN-generated faces," *IEEE Access*, vol. 10, pp. 32574–32583, 2022.

[12] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "GAN-generated faces detection: A survey and new perspectives," in *European Conference on Artificial Intelligence*, 2023.

[13] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2500–2504, IEEE, 2021. 1

[14] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection," in *International Joint Conference on Artificial Intelligence*, 2021. 1, 2

[15] Y. Xu, P. Terhörst, K. Raja, and M. Pedersen, "A comprehensive analysis of AI biases in deepfake detection with massively annotated databases," *arXiv preprint arXiv:2208.05845*, 2022. 1, 2, 5

[16] A. V. Nadimpalli and A. Rattani, "GBDF: gender balanced deepfake dataset towards fair deepfake detection," *arXiv preprint arXiv:2207.10246*, 2022. 1, 2, 3, 6, 15

[17] L. Zhang, H. Chen, S. Hu, B. Zhu, X. Wu, J. Hu, and X. Wang, "X-transfer: A transfer learning-based framework for robust gan-generated fake image detection," *arXiv preprint arXiv:2310.04639*, 2023.

[18] S. Yang, S. Hu, B. Zhu, Y. Fu, S. Lyu, X. Wu, and X. Wang, "Improving cross-dataset deepfake detection with deep information decomposition," *arXiv preprint arXiv:2310.00359*, 2023. 1

[19] C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer, "Towards measuring fairness in ai: the casual conversations dataset," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 324–332, 2021. 1, 2

[20] L. Kim, "Fake pictures of people of color won't fix AI bias," in *WIRED*, `https://tinyurl.com/yc4dsyms`, 2023. 1

[21] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. 1, 2, 6

[22] R. Williamson and A. Menon, "Fairness risk measures," in *International Conference on Machine Learning*, pp. 6786–6797, PMLR, 2019. 1, 4, 5, 6, 15

[23] H. Rahimian and S. Mehrotra, "Distributionally robust optimization: A review," *arXiv preprint arXiv:1908.05659*, 2019. 2

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

[25] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, pp. 21–42, 2000. 2, 3

[26] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 5, 6

[27] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3207–3216, 2020. 2, 5, 6

[28] "Deepfakes dataset by Google & Jigsaw.," in `https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html`, 2019. 2, 5

[29] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (DFDC) dataset," *arXiv preprint arXiv:2006.07397*, 2020. 2, 5

[30] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*, pp. 1929–1938, PMLR, 2018. 2, 3, 5, 6, 15

[31] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 728–740, 2020. 2

[32] T. Zhao, E. Dai, K. Shu, and S. Wang, "Towards fair classifiers without sensitive attributes: Exploring biases in related features," in *Proceedings of the ACM International Conference on Web Search and Data Mining*, pp. 1433–1442, 2022. 2

[33] S. Yan, H.-t. Kao, and E. Ferrara, "Fair class balancing: Enhancing model fairness without observing sensitive attributes," in *Proceedings of the ACM International Conference on Information & Knowledge Management*, pp. 1715–1724, 2020. 2

[34] J. Chai, T. Jang, and X. Wang, "Fairness without demographics through knowledge distillation," in *Advances in Neural Information Processing Systems*. 2

[35] V. Grari, S. Lamprier, and M. Detyniecki, "Fairness without the sensitive attribute via causal variational autoencoder," in *International Joint Conference on Artificial Intelligence*, 2022. 2

[36] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012. 2

[37] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *IEEE International Conference on Data Engineering*, pp. 1918–1921, IEEE, 2020. 2

[38] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016. 2

[39] S. Caton and C. Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020. 2, 6

[40] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023. 2

[41] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9468–9478, 2022.

[42] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3994–4004, 2023.

[43] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, "AltFreezing for more general video face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4129–4138, 2023. 2

[44] M. Pu, M. Y. Kuan, N. T. Lim, C. Y. Chong, and M. K. Lim, "Fairness evaluation in deepfake detection models using metamorphic testing," *arXiv preprint arXiv:2203.06825*, 2022. 2

[45] J. Yu, X. Hao, H. Xie, and Y. Yu, "Fair face recognition using data balancing, enhancement and fusion," in *European Conference on Computer Vision*, pp. 492–505, Springer, 2020. 2

[46] K. Karkkainen and J. 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.

[47] M. Fang, W. Yang, A. Kuijper, V. Struc, and N. Damer, "Fairness in face presentation attack detection," *arXiv preprint arXiv:2209.09035*, 2022.

[48] R. Ramachandra, K. Raja, and C. Busch, "Algorithmic fairness in face morphing attack detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 410–418, 2022. 2

[49] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021. 3

[50] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar, "Doro: Distributional and outlier robust optimization," in *International Conference on Machine Learning*, pp. 12345–12355, PMLR, 2021. 3, 7

[51] S. Hu, L. Ke, X. Wang, and S. Lyu, "TkML-AP: Adversarial attacks to top-k multi-label learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7649–7657, 2021. 4

[52] S. Hu and G. H. Chen, "Distributionally robust survival analysis: A novel fairness loss without demographics," in *Machine Learning for Health*, pp. 62–87, PMLR, 2022.

[53] S. Hu, Y. Ying, X. Wang, and S. Lyu, "Sum of ranked range loss for supervised learning," *Journal of Machine Learning Research*, vol. 23, no. 1, pp. 4826–4869, 2022.

[54] S. Hu, Y. Ying, and S. Lyu, "Learning by minimizing the sum of ranked range," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 21013–21023, 2020. 4

[55] S. Hu, X. Wang, and S. Lyu, "Rank-based decomposable losses in machine learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5

[56] J. C. Harsanyi, *Rational behaviour and bargaining equilibrium in games and social situations*. Cambridge University Press, 1977. 5

[57] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *International Conference on Machine Learning*, pp. 4615–4625, PMLR, 2019. 5

[58] J. Rawls, *Justice as fairness: A restatement*. Harvard University Press, 2001. 5

[59] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *International Conference on Learning Representations*, 2020. 6, 15

[60] J. Wang, X. E. Wang, and Y. Liu, "Understanding instance-level impact of fairness constraints," in *International Conference on Machine Learning*, pp. 23114–23130, PMLR, 2022. 6, 15

[61] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009. 5, 15

[62] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14548–14556, 2023. 6

[63] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 943–952, 2023. 6

[64] A. V. Nadimpalli and A. Rattani, "On improving cross-dataset generalization of deepfake detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 91–99, 2022. 6

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. 6

[66] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019. 6

[67] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4122, 2022. 6

[68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019. 6

[69] K. N. Keya, R. Islam, S. Pan, I. Stockwell, and J. Foulds, "Equitable allocation of healthcare resources with fair survival models," in *Proceedings of the SIAM International Conference on Data Mining*, pp. 190–198, SIAM, 2021. 7