Preserving Fairness Generalization in Deepfake Detection

Li Lin[1], Xinan He[2], Yan Ju[3], Xin Wang[4], Feng Ding[2], Shu Hu[1]*
[1]Purdue University {lin1785, hu968}@purdue.edu
[2]Nanchang University {shahur, fengding}@ncu.edu.cn
[3]University at Buffalo, State University of New York yanju@buffalo.edu
[4]University at Albany, State University of New York xwang56@albany.edu

Abstract

Although effective deepfake detection models have been developed in recent years, recent studies have revealed that these models can result in unfair performance disparities among demographic groups, such as race and gender. This can lead to particular groups facing unfair targeting or exclusion from detection, potentially allowing misclassified deepfakes to manipulate public opinion and undermine trust in the model. The existing method for addressing this problem is providing a fair loss function. It shows good fairness performance for intra-domain evaluation but does not maintain fairness for cross-domain testing. This highlights the significance of fairness generalization in the fight against deepfakes. In this work, we propose the first method to address the fairness generalization problem in deepfake detection by simultaneously considering features, loss, and optimization aspects. Our method employs disentanglement learning to extract demographic and domain-agnostic forgery features, fusing them to encourage fair learning across a flattened loss landscape. Extensive experiments on prominent deepfake datasets demonstrate our method's effectiveness, surpassing state-of-the-art approaches in preserving fairness during cross-domain deepfake detection. The code is available at https://github.com/Purdue-M2/Fairness-Generalization.

1. Introduction

Deepfakes, a portmanteau of "deep learning" and "fake," have emerged as a captivating yet concerning facet of contemporary technology. These are AI-generated or manipulated media (*e.g.*, images, videos) through deep neural networks (*e.g.*, variational autoencoder [1], generative adversarial networks [2], diffusion models [3]) that appear startlingly genuine, often featuring individuals engaged in actions they never partook in or uttering words they never spoke. While deepfakes have opened doors to creative content and entertainment, malicious use of deepfakes can lead to misinformation, privacy breaches, and even political manipulation, eroding trust and generating confusion [4, 5].

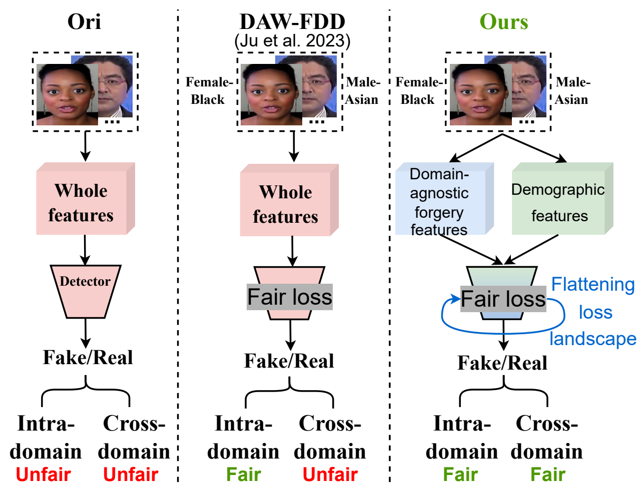

Figure 1. Comparison between our method and existing deepfake detection baselines. (**Left**) The Ori represents the conventional method without any fair characters. (**Middle**) The DAW-FDD [6] is an intra-domain fair deepfake detection method. However, this method fails in cross-domain fair detection. (**Right**) Our method succeeds in achieving both intra-domain and cross-domain fair detection by exposing domain-agnostic forgery features and demographic features and then fusing them for fair learning across a flattened loss landscape.

To counteract the spread of deceptive deepfakes, there is a burgeoning field of deepfake detection methods that are data-driven and deep-learning based [7–25]. However, recent research and reports [26–30] have brought to light fairness issues within current deepfake detection methods. One significant concern revolves around the inconsistency in performance when assessing different demographic groups, including gender, age, and ethnicity [27]. For example, some of the most advanced detectors exhibit higher accuracy when evaluating deepfakes featuring individuals with lighter skin tones compared to those with darker skin tones [26, 31].

*Corresponding author

This allows attackers to generate harmful deepfakes targeting specific populations in order to evade detection.

An initial algorithm-level approach to addressing fairness in deepfake detection has been presented by Ju et al. [6]. They showed that the proposed DAW-FDD model could exhibit the best fairness performance under the intra-domain evaluation scenario, *i.e.*, training and testing data are generated by the same forgery techniques. However, in practice, we found that their method does not preserve fairness for cross-domain evaluation, *i.e.*, when testing on data generated by unknown forgeries. Notably, achieving fairness generalization is critical. Without such generalization, the current fair deepfake detection methods are susceptible to obsolescence easily.

In this work, we experimentally and theoretically analyze the entanglement of demographic and forgery features, and the sharpness of loss landscapes could be the fuse to affect the fairness generalization in deepfake detection. To address these issues, we propose a novel framework to preserve fairness in deepfake detection generalization, consisting of three key modules: disentanglement learning, fairness learning, and optimization. Specifically, in the disentanglement learning module, we introduce a disentanglement loss to expose demographic and domain-agnostic forgery features — the feature-level factors directly affecting the fairness generalization capabilities of the detector. The fairness learning module combines these disentangled features to promote fair learning while guided by generalization principles. Additionally, we include a bi-level fairness loss to enhance fairness both across and within subgroups. The optimization module focuses on flattening the loss landscape, allowing the model to escape suboptimal solutions and fortify its fairness generalization capability. Fig. 1 illustrates how our method differs from existing ones. Our contributions are as follows:

- We experimentally and theoretically analyze the unfairness problem in deepfake detection generalization.
- We propose the first method to improve fairness generalization in deepfake detection by simultaneously addressing features, loss, and optimization. Specifically, we utilize disentanglement learning to extract demographic and domain-agnostic forgery features, which are then integrated to facilitate fair learning across a flattened loss landscape.
- Our method outperforms state-of-the-art approaches in preserving fairness during cross-domain deepfake detection, as demonstrated in extensive experiments on various leading deepfake datasets.

2. Related Work

Deepfake Detection. The largest portion of existing deepfake detection methods fall into the *data-driven* category, including [7–13]. These methods leverage various types of Deep Neural Networks (DNNs) trained on both authentic and deepfake videos to capture specific discernible artifacts. While these methods have achieved promising performance for the intra-domain evaluation, they suffer from sharp performance degradation on cross-domain testing. To address the generalization issue, disentanglement learning [32] is widely used for forgery detection by extracting relevant features while eliminating irrelevant ones. For instance, Hu et al. [14] introduced a disentanglement framework to automatically locate forgery-related regions, and Zhang et al. [15] enhanced generalization through auxiliary supervision. Liang et al. [16] proposed a framework that improves feature independence through content consistency and global representation contrastive constraints. Yan et al. [17] extended this framework by exclusively utilizing common forgery features, which are separated from forgery-related features.

Fairness in Deepfake Detection. Recent studies have mentioned fairness issues in deepfake detection [30]. Trinh et al. [26] identified biases in both deepfake datasets and detection models, revealing significant error rate differences across subgroups. Similar observations were reported in the study by Hazirbas et al. [31]. Pu et al. [33] assessed the fairness of the MesoInception-4 deepfake detection model on FF++ and found it to be unfair to both genders. Xu et al. [27] conducted a comprehensive analysis of bias in deepfake detection, enriching datasets with diverse annotations to support future research. Additionally, Nadimpalli et al. [29] highlighted substantial bias in datasets and detection models, introducing a gender-balanced dataset to mitigate gender-based performance bias. However, this approach yielded only modest improvements and required extensive data annotation. Ju et al. [6] focused on enhancing fairness within the same data domain but did not address fairness in cross-domain testing, which is the central focus of our paper.

3. Motivation

Unfairness in Cross-domain Detection. To assess the performance of existing fair deepfake detection methods in ensuring fairness across different testing domains, we utilized the DAW-FDD method [6] with an Xception backbone. For comparison, we employed a baseline detector with the same backbone and binary cross-entropy loss, and named it 'Ori'. To evaluate the effectiveness of incorporating fairness loss in generalized detectors, we examined the UCF baseline [17] and trained it with the DAW-FDD fair loss during training, denoted as DAW-FDD (UCF). All models were trained on the FF++ dataset [34] and were subsequently tested on both the FF++ and DFD [35] datasets. Fairness performance was assessed in terms of demographic group intersection using two fairness metrics: F_{MEO} [36] and F_{DP} [37] (details are provided in Appendix B).

The comparison results are presented in Fig. 2 (Left & Middle). The intra-domain testing results reveal that the fairness scores of DAW-FDD and DAW-FDD (UCF) are con-

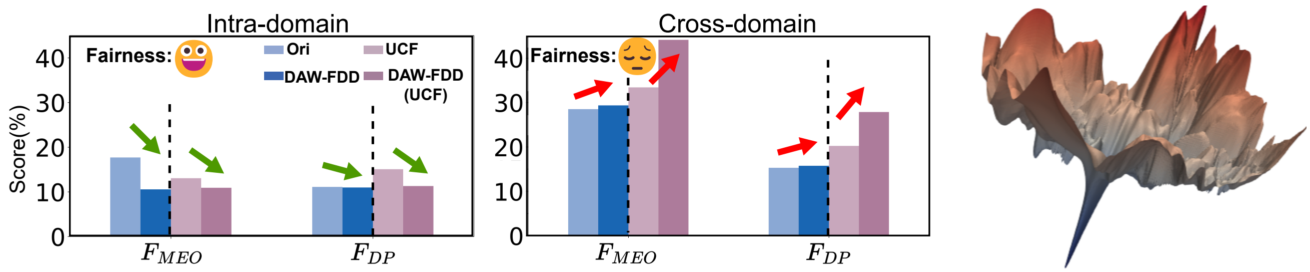

Figure 2. Experimental results for Motivation. Testing fairness results (lower is better for all metrics) of deepfake detectors in intra-domain (**Left**, train and test: FF++) and cross-domain (**Middle**, train: FF++, test: DFD) detection. (**Right**) Visualization of loss landscape for DAW-FDD. The numerous local and global minima could cause the model to have poor generalization.

sistently lower across all metrics when compared to Ori and UCF, respectively. However, in cross-domain testing, DAW-FDD's fairness scores are worse than those of Ori, highlighting the challenge of maintaining fairness when applied across different domains. Additionally, DAW-FDD (UCF) has fairness scores worse than UCF, indicating that merely integrating a fair loss into generalized deepfake detectors is insufficient to ensure successful fairness generalization in cross-domain scenarios.

Analysis. Next, we investigate why current methods fall short in preserving fairness in cross-domain detection, examining both features and optimization-related aspects. In this analysis, we use variables: X (*e.g.*, an image), Y (the corresponding target variable, *e.g.*, fake or real), $\hat{Y}$ (the classifier's prediction for X), and D (the demographic variable linked to X). Here, $D \in \mathcal{J}$, where $\mathcal{J}$ represents user-defined subgroups (*e.g.*, $\mathcal{J}$ ={male, female} for gender). For simplicity, we assume $\mathcal{J}$ contains two subgroups, $\mathcal{J}_1$ and $\mathcal{J}_2$.

Feature Aspect. We introduce a theorem as follows:

Theorem 1. *([38]) If X is entangled with Y and D, the use of a perfect classifier for $\hat{Y}$,* i.e.*, $P(\hat{Y}|X) = P(Y|X)$, does **not** imply demographic parity,* i.e.*, $P(\hat{Y} = y|D = \mathcal{J}_1) = P(\hat{Y} = y|D = \mathcal{J}_2)$, $\forall y \in \{0, 1\}$, where 0 means real and 1 means fake.*

Theorem 1 highlights the challenge of achieving fairness in a model that directly operates on entangled representations $r(X)$ (*i.e.*, $r(X) = X$ when the representations are the identity function), where these representations are a blend of target information $r(X)_Y$ (for identifying label Y) and demographic information $r(X)_D$ (for identifying D). This observation suggests a possible reason for the limited success of DAW-FDD [6] in fairness generalization.

Therefore, disentanglement could be an approach to enhance fairness by untangling the representations $r(X)_Y$ and $r(X)_D$ from $r(X)$, ensuring their independence, *i.e.*, $r(X)_Y \perp\!\!\!\perp r(X)_D$. Previous methods [14–17] have explored disentanglement learning, particularly in extracting forgery-related features to enhance the generalization of deepfake detection. However, none of these methods addresses the disentanglement of demographic representation $r(X)_D$. This omission explains why directly applying DAW-FDD to these existing generalization-based models does not preserve fairness in cross-dataset testing. Yet, isolating $r(X)_Y \perp\!\!\!\perp r(X)_D$ could compromise the detection performance of models that rely solely on $r(X)_Y$. This is because forgery and demographic features in deepfakes are often linked to facial characteristics. Removing $r(X)_D$ would result in the loss of facial information that could be related to forgery, potentially causing performance degradation. Hence, this presents a complex challenge that requires careful consideration.

Optimization Aspect. In addition, existing DNN-based deepfake detection models are highly overparameterized, enabling them to memorize both data and demographic patterns during training. Consequently, the straightforward minimization of commonly used fairness loss functions, such as in the DAW-FDD method, is insufficient to ensure robust fairness generalization. Training these models results in sharp loss landscapes characterized by multiple local and global minima [39], each leading to models with varying generalization capabilities due to being trapped into different suboptimal minima. Refer to Fig. 2 (Right) for an example of the DAW-FDD loss landscape. Hence, it becomes essential to flatten the loss landscape to enhance fairness generalization.

4. Method

4.1. Overview of Proposed Method

According to the insights from Section 3, we propose a new method to preserve fairness generalization in deepfake detection in this section. We first formulate the problem.

Problem Setup. Given a training dataset $\mathcal{S} = \{(X_i, D_i, A_i, Y_i)\}_{i=1}^n$ with size n. A_i represents the domain label, indicating the source of X_i. For example, in the FF++ dataset [34], $A_i \in$ {real, DeepFakes [40], Face2Face [41], FaceSwap [42], NeuralTextures [43], FaceShifter [44]}, which correspond to real and fake images generated by various face manipulation methods. Our objective is to train a fair deepfake detection model using $\mathcal{S}$ that can then generalize to an unseen deepfake dataset while maintaining both accuracy and fairness.

Framework. Fig. 3 depicts our framework, comprising three

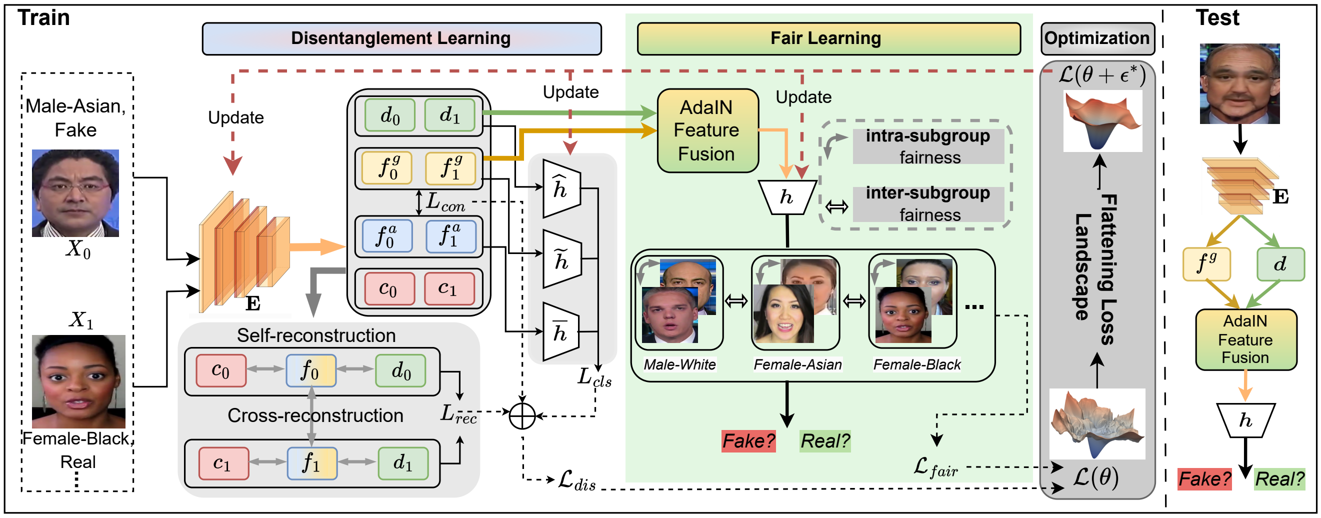

Figure 3. An overview of our proposed method. 1) For the disentanglement learning module, we utilize it to expose demographic and forgery features. 2) For the fair learning module, we fuse those two features for a fair classifier head h and obtain the fair prediction using two-level fairness loss $\mathcal{L}_{fair}$. 3) For the optimization module, we flatten the loss landscape to further enhance fairness generalization.

modules: disentanglement learning, fair learning, and optimization. The disentanglement learning module's purpose is to extract domain-agnostic forgery and demographic features from input images. The fair learning module leverages these two types of features to develop a fair classifier. Both learning modules are supervised by an optimization module, enhancing fairness generalization during model training. We will delve into each module's specifics in the following sections. The entire training process is end-to-end.

4.2. Exposing Demographic & Forgery Features

We propose a disentanglement learning module to extract both demographic features (for fairness) and domain-agnostic forgery features (for generalization). To achieve this, we use pairs of images $(X_i, X_{i'})$, where X_i is fake (or real), $X_{i'}$ is real (or fake), $i, i' \in \{1, \cdots, n\}$, and $i \neq i'$. Each image is processed by an encoder $\mathbf{E}(\cdot)$, which includes three distinct encoders[1] responsible for extracting content features c (*i.e.*, related to the image background), forgery features f, and demographic features d. Note that the forgery features encompass both domain-specific forgery features f^a (*i.e.*, specific to the forgery method) and domain-agnostic forgery features f^g (*i.e.*, common to various forgery methods). The procedure is formulated as follows,

$$c_i, f_i^a, f_i^g, d_i = \mathbf{E}(X_i).$$

Classification Loss. Disentangling domain-specific forgery, domain-agnostic forgery, and demographic features typically involves using cross-entropy (CE) loss for each of them. However, deepfake datasets often suffer from imbalances in demographic subgroup distributions, a fundamental issue in achieving fairness in detection [29, 45]. Additionally, conventional CE loss training tends to lead to overfitting on examples from the majority subgroups [46], making it unsuitable for learning fair demographic feature representations. To address these challenges, we propose a demographic distribution-aware margin loss inspired by [47] as follows:

$$M(\widehat{h}(d_i), D_i) = -\log \frac{e^{\widehat{h}^{D_i}(d_i) - \Delta^{D_i}}}{e^{\widehat{h}^{D_i}(d_i) - \Delta^{D_i}} + \sum_{p \neq D_i} e^{\widehat{h}^{p}(d_i)}},$$

where $\Delta^p = \frac{\delta}{n_p^{1/4}}$ is a demographic subgroup-dependent margin for $p \in \mathcal{J}$ and δ is a constant. n_p denotes the number of training data points from subgroup p. $\widehat{h}$ is the classification head for d_i and $\widehat{h}^p$ represents the output for p.

By incorporating this margin loss, we improve generalization for minority subgroups with small n_p by using larger margins Δ^p, promoting unbiased demographic feature representation. Hence, the total classification loss is:

$$L_{cls} = C(\widetilde{h}(f_i^g), Y_i) + \rho_1 C(\overline{h}(f_i^a), A_i) + \rho_2 M(\widehat{h}(d_i), D_i),$$

where $C(\cdot,\cdot)$ is the CE loss. $\overline{h}$ and $\widetilde{h}$ are the classification heads for f_i^a and f_i^g, respectively[2]. ρ_1 and ρ_2 are two trade-off hyperparameters. Training with the above classification loss enables the encoder to acquire specific feature information, enhancing the model's generalization capability.

Contrastive Loss. The classification loss, which focuses on individual images, overlooks the image correlations that play a crucial role in enhancing the encoder's representation capabilities. Inspired by contrastive learning [17, 48], we can introduce a contrastive loss to address this gap:

$$L_{con} = [b + \|f_{\text{anchor}} - f_{+}\|_2 - \|f_{\text{anchor}} - f_{-}\|_2]_{+},$$

[1]The three encoders share the same architecture but with different parameters, and the architecture details can be found in Appendix C.

[2]These classification heads share the same multilayer perceptron (MLP) architecture but with different parameters.

where f_{anchor} represents anchor forgery features of an image, and f_+ and f_- represent its positive counterpart from the same source and the negative counterpart from a different source, respectively. b is a hyperparameter and $[\cdot]_+ = \max\{0, \cdot\}$ is a hinge function. We employ L_{con} for both domain-specific and domain-agnostic forgery features in practice. For domain-specific forgery features, the source is considered the forgery domain, and the contrastive loss motivates the encoder to learn specific forgery representations. For domain-agnostic forgery features, the source can be either real or fake, and the loss encourages the encoder to learn a generalizable representation that is not tied to any specific forgery method.

Reconstruction Loss. To preserve the completeness of the extracted features and maintain consistency between the original and reconstructed images at the pixel level, we employ a reconstruction loss. It is formulated as:

$$L_{rec} = \|X_i - \mathbf{D}(c_i, f_i, d_i)\|_1 + \|X_i - \mathbf{D}(c_i, f_{i'}, d_i)\|_1,$$

where $\mathbf{D}(\cdot,\cdot,\cdot)$ is the decoder responsible for reconstructing an image using the disentangled feature representations (refer to Appendix C for architecture details). In L_{rec} loss, the first term is the self-reconstruction loss, which minimizes reconstruction errors using the latent features of the input image. The second term is the cross-reconstruction loss, which penalizes reconstruction errors by incorporating the partner's forgery feature. These two terms work together to improve feature disentanglement.

Disentanglement Loss. Therefore, the disentanglement loss for exposing demographic and forgery features is

$$\mathcal{L}_{dis} = \frac{1}{n}\sum_i [L_{cls} + \rho_3 L_{con} + \rho_4 L_{rec}], \tag{1}$$

where ρ_3 and ρ_4 are trade-off hyperparameters.

4.3. Fair Learning under Generalization

Once we acquire both the domain-agnostic forgery features and demographic features, we combine them for the purpose of fairness learning using Adaptive Instance Normalization (AdaIN) [49]. The fused feature I_i can be formed as follows,

$$I_i = \sigma(d_i)\Big(\frac{f_i^g - \mu(f_i^g)}{\sigma(f_i^g)}\Big) + \mu(d_i),$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ compute the mean and standard deviation of the input feature across spatial dimensions independently for each channel. The combination is necessary because deepfake forgery methods often modify the facial region of an image, which contains essential features for determining demographic information. Ignoring either of these features would significantly reduce fairness generalization performance. Our experiments in Section 5.3 confirm this.

Fairness Loss. Traditional approaches for achieving fair learning, such as [36, 37], often involve adding a fairness penalty to the learning objective. However, these methods can only ensure fairness on specific fairness measures, like demographic parity [50] or equalized odds [51], which limits the model's fairness scalability and its ability to work with new datasets. Additionally, even if the overall deepfake dataset has balanced fake and real examples, imbalances can still exist within demographic subgroups, potentially leading to biased learning within those subgroups.

To address these problems, inspired by [6, 52–57], we introduce a bi-level fairness loss as follows:

$$\mathcal{L}_{fair} = \min_{\eta\in\mathbb{R}} \eta + \frac{1}{\alpha|\mathcal{J}|}\sum_{j=1}^{|\mathcal{J}|}[L_j - \eta]_+, \tag{2a}$$

$$\text{s.t. } L_j = \min_{\eta_j\in\mathbb{R}} \eta_j + \frac{1}{\alpha'|\mathcal{J}_j|}\sum_{i:D_i=\mathcal{J}_j}[C(h(I_i), Y_i) - \eta_j]_+. \tag{2b}$$

Here, $|\mathcal{J}|$ represents the size of set $\mathcal{J}$, with each subgroup $\mathcal{J}_j \in \mathcal{J}$, and $|\mathcal{J}_j|$ represents the number of training examples in $\mathcal{J}_j$. h is the classification head for I_i, sharing the same MLP architecture as other heads, and $\alpha, \alpha' \in (0,1)$ are two hyperparameters. The outer-level formulation (Eq. (2a)) draws inspiration from the fairness risk measure [58], aiming to promote fairness among *inter-subgroups*. The inner-level formulation (Eq. (2b)) is inspired by distributionally robust optimization (*i.e.*, Conditional Value-at-Risk [59]), which enhances fairness across both real and fake examples within *intra-subgroup*, thereby bolstering model robustness.

4.4. Joint Optimization

Lastly, we jointly optimize the above two modules in a unified framework. To avoid numerous sharp and narrow minima described in Fig. 2, we utilize the sharpness-aware minimization method [39] to flatten the loss landscape. Specifically, denoting the model weights of the whole framework as θ, flattening is attained by determining an optimal ϵ^* for perturbing θ to maximize the loss, defined as:

$$\begin{aligned}\epsilon^* &= \arg\max_{\|\epsilon\|_2\le\gamma} \underbrace{(\mathcal{L}_{dis} + \lambda\mathcal{L}_{fair})}_{\mathcal{L}}(\theta+\epsilon) \\ &\approx \arg\max_{\|\epsilon\|_2\le\gamma} \epsilon^\top \nabla_\theta\mathcal{L} = \gamma\texttt{sign}(\nabla_\theta\mathcal{L}),\end{aligned} \tag{3}$$

where γ is a hyperparameter that controls the perturbation magnitude, and λ is a trade-off hyperparameter. The approximation is obtained using first-order Taylor expansion with the assumption that ϵ is small. The final equation is obtained by solving a dual norm problem, where $\texttt{sign}$ represents a sign function and $\nabla_\theta\mathcal{L}$ being the gradient of $\mathcal{L}$ with respect to θ. As a result, the model weights are updated by solving the following problem:

$$\min_\theta \mathcal{L}(\theta + \epsilon^*). \tag{4}$$

The intuition is that the perturbation along the gradient norm direction increases the loss value significantly and then makes the model more generalizable in terms of fairness.

End-to-end Training. In practice, we first initialize the model weights θ and then randomly select a mini-batch set

$\mathcal{S}_b$ from $\mathcal{S}$, performing the following steps for each iteration on $\mathcal{S}_b$ (see Appendix D for more details about Algorithm):

- Fix θ and use binary search to find the global optimum of η_j since (2b) is convex w.r.t. η_j.
- Take L_j into (2a) and use binary search to find the global optimum of η since (2a) is convex w.r.t. η.
- Fix η_j and η, compute ϵ^* based on Eq. (3).
- Update θ based on the gradient approximation for (4): $\theta \leftarrow \theta - \beta \nabla_\theta \mathcal{L}\big|_{\theta+\epsilon^*}$, where β is a learning rate.

5. Experiment

5.1. Experimental Settings

Datasets. To validate the fairness generalization ability of our proposed method, we train our model on the most widely used benchmark FaceForensics++(FF++) [34] and test it on FF++, DeepfakeDetection (DFD) [35], Deepfake Detection Challenge (DFDC) [60], and Celeb-DF [61]. The forged images we use in FF++ are generated by five face manipulation algorithms, including DeepFakes (DF) [40], Face2Face (F2F) [41], FaceSwap (FS) [42], NerualTexture (NT) [43], and FaceShifter (FST) [44]. Since the original datasets do not have the demographic information of each video or image, we follow Ju et al. [6] for data processing, data annotation, and sensitive attributes combination (Intersection). Therefore, the Intersection group contains 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). Details of each annotated dataset are in Appendix E.

Evaluation Metrics. For detection comparison, the Area Under Curve (AUC) is used to benchmark our approach against previous works, which aligns with the detection evaluation approach adopted in precedent works [17, 62]. Regarding fairness, we use four distinct fairness metrics to evaluate the effectiveness of our proposed method. Specifically, we report the Equal False Positive Rate (F_{FPR}) [6], Max Equalized Odds (F_{MEO}) [36], Demographic Parity (F_{DP}) [37] and Overall Accuracy Equality (F_{OAE}) [36]. The definition of those fairness metrics can be found in Appendix B.

Baseline Methods. We compare our method against the latest fairness method DAW-FDD [6] in deepfake detection. The comparison also includes 'Ori' (a backbone with cross-entropy loss) and UCF [17] (the latest disentanglement-based deepfake detector). Unless explicitly specified, all methods are employed on Xception [63] backbone.

Implementation Details. All experiments are based on the PyTorch and trained with NVIDIA RTX 3090Ti. For training, we fix the batch size 16, epochs 100, use SGD optimizer with learning rate β =5×10^{-4}. For the overall loss, we set the λ in Eq. (3) as 1.0, the γ (neighborhood size of perturbation in flattening loss) as 0.05, the ρ_1, ρ_2 in L_{cls} as 0.1, 0.1, the ρ_3, ρ_4 in $\mathcal{L}_{dis}$ as 0.05 and 0.3, b in L_{con} as 3.0, and δ in $M(\widehat{h}(d_i), D_i)$ as 2.89 based on the demographic sample distribution. The α and α' in Eq. (2) are tuned on the grid {0,1,0.3,0.5,0.7,0.9}. Following [6], the final α and α' are determined based on a preset rule that allows up to a 5% degradation of overall AUC in the validation set from the corresponding 'Ori' method while minimizing the F_{FPR} on Intersection group.

Testing Set	Method	Fairness Metrics(%)↓				Detection Metric(%)↑
		F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC
F2F [41]	DAW-FDD [6]	20.42	12.66	35.46	11.58	97.74
	Ours	**17.42**	**10.00**	**33.20**	**9.56**	**98.65**
FS [42]	DAW-FDD [6]	32.96	14.52	21.39	**3.95**	98.62
	Ours	**26.32**	**9.97**	**19.30**	6.70	**99.23**
NT [43]	DAW-FDD [6]	**23.64**	20.83	20.50	17.36	94.99
	Ours	23.98	**16.83**	**16.03**	**13.61**	**96.35**
DF [40]	DAW-FDD [6]	20.41	12.66	9.99	6.16	98.20
	Ours	**17.42**	**9.02**	**9.43**	**5.86**	**99.05**
FST [44]	DAW-FDD [6]	25.36	10.05	10.34	8.79	98.02
	Ours	**15.38**	**7.79**	**6.45**	**5.70**	**98.96**

Table 1. Intra-domain evaluation on FF++. DAW-FDD and our method are trained on FF++, tested on its test sub-datasets separated by five forgeries, *i.e.*, F2F is the sub-dataset in FF++ test set generated by Face2Face [41]. The best results are shown in **Bold**.

5.2. Results

Performance on Intra-domain sub-datasets. Intra-domain evaluation, conducted on individual forgery sub-dataset, assesses the model's proficiency in fitting the specific forgery sub-dataset. As illustrated in Table 1, our disentanglement learning approach, which separates domain-specific forgery, guides the model not to overfit to a particular forgery domain. In general, our method enhances fairness and consistently achieves a higher AUC on each sub-dataset compared to DAW-FDD. This result suggests the effectiveness of eliminating domain-specific biases.

Performance of Fairness Generalization. Taking Xception backbone as an example, Table 2 shows our method has superior fairness generalization ability compared to other methods, while simultaneously achieving the best detection results. Specifically, our method has an 8.90% improvement in F_{DP} on DFDC and enhances the F_{FPR} by 11.69% on Celeb-DF, 7.94% on DFD compared with DAW-FDD [6]. In addition, although DAW-FDD, as a fair detector, works well on FF++ compared to Ori, it underperforms Ori under certain cross-domain scenarios, with a notable 4.72% decrease in F_{DP} on DFDC and declines in F_{MEO} and F_{DP} on DFD. UCF [17], recognized as a state-of-the-art detector in improving detection generalization, surpasses Ori and DAW-FDD in detection. However, it fails to ensure fairness, as evidenced by its F_{DP} being 3.94% inferior to Ori's even in intra-domain testing, with all four fairness metrics on DFD performing worse than Ori. Overall, our method outperforms all compared methods across most fairness metrics, achieving the best in both fairness generalization and AUC.

Fairness Generalization of Different Backbones. To ex-

Dataset	Method	Xception [63]					ResNet-50 [64]					EfficientNet-B3 [65]				
		Fairness Metrics(%)↓				Detection Metric(%)↑	Fairness Metrics(%)↓				Detection Metric(%)↑	Fairness Metrics(%)↓				Detection Metric(%)↑
		F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC	F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC	F_{FPR}	F_{MEO}	F_{DP}	F_{OAE}	AUC
FF++	Ori [34]	31.31	17.69	11.12	10.08	92.77	34.69	17.29	9.83	8.85	94.83	18.78	33.21	31.36	26.01	93.55
	DAW-FDD [6]	14.06	10.55	10.97	8.72	97.46	30.36	9.74	8.89	7.42	93.23	23.33	26.15	24.74	21.23	94.92
	UCF [17]	21.52	13.06	15.06	10.58	97.10	35.13	10.87	10.81	8.05	95.92	20.92	33.08	30.01	24.56	94.21
	Ours	**10.63**	**8.15**	**10.41**	**7.60**	**98.28**	**22.70**	**9.28**	**8.72**	**5.74**	**97.72**	**11.19**	**20.61**	**18.40**	**16.18**	**95.39**
DFDC	Ori [34]	52.77	37.78	13.87	30.30	56.72	45.84	28.89	16.67	26.25	58.08	62.38	37.56	22.44	25.93	57.81
	DAW-FDD [6]	45.14	35.77	18.59	14.07	59.96	44.07	34.14	18.72	24.58	60.11	50.73	43.79	18.31	29.57	58.29
	UCF [17]	53.07	44.44	15.70	23.22	60.03	43.39	35.62	15.86	19.15	**61.06**	42.79	40.54	19.35	21.13	58.85
	Ours	**40.73**	**34.48**	**9.69**	**13.71**	**61.47**	**37.17**	**27.78**	**10.94**	**18.52**	59.76	**22.89**	**33.78**	**12.35**	**16.73**	**60.67**
Celeb-DF	Ori [34]	27.55	25.65	17.74	58.44	62.66	24.94	22.32	19.47	48.62	70.64	30.86	27.47	19.15	59.32	62.36
	DAW-FDD [6]	22.31	20.60	**11.65**	49.71	69.55	26.82	21.93	20.80	47.14	75.70	31.36	21.79	6.91	**50.86**	70.14
	UCF [17]	27.81	25.96	16.51	48.63	71.73	32.17	28.28	19.38	45.15	76.44	24.95	22.41	15.14	58.48	72.65
	Ours	**10.62**	**12.77**	15.04	**36.01**	**74.42**	**11.55**	**17.01**	**17.21**	**29.58**	**78.55**	**13.00**	**9.73**	**5.21**	55.74	**75.32**
DFD	Ori [34]	35.14	28.52	15.31	12.95	74.34	31.76	26.91	5.90	28.48	76.02	39.37	38.57	20.01	17.00	75.87
	DAW-FDD [6]	34.02	29.37	15.75	11.31	71.42	33.05	24.24	7.12	27.08	77.05	32.72	28.74	17.12	24.70	74.76
	UCF [17]	42.66	33.41	20.24	19.84	81.88	42.54	33.17	5.24	30.98	78.97	36.59	27.32	25.83	9.36	76.76
	Ours	**26.08**	**21.37**	**11.65**	**8.37**	**84.82**	**25.71**	**20.02**	**2.34**	**25.60**	**79.67**	**29.34**	**24.52**	**11.46**	**5.11**	**77.28**

Table 2. Comparison with different methods in terms of improving fairness and detection generalization under both intra-domain (FF++) and cross-domain (DFDC, Celeb-DF, and DFD) scenarios. ↑ means higher is better and ↓ means lower is better.

Effects	Method							Dataset							
								FF++		DFDC		Celeb-DF		DFD	
	Name	Cls(CE)	Cls	Rec	Con	Ff	Lf	F_{FPR}↓	AUC↑	F_{FPR}↓	AUC↑	F_{FPR}↓	AUC↑	F_{FPR}↓	AUC↑
Dl	VariantA		✓			✓	✓	17.62	98.06	43.24	58.14	19.08	68.38	27.81	81.98
	VariantB		✓	✓		✓	✓	17.40	98.24	41.44	59.84	13.61	71.07	26.52	82.08
	VariantC		✓		✓	✓	✓	15.96	97.93	44.01	60.91	12.76	72.41	26.36	84.19
	VariantD	✓		✓	✓	✓	✓	16.58	98.05	42.76	60.16	14.04	74.14	29.57	84.66
	Ours		✓	✓	✓	✓	✓	**10.63**	**98.28**	**40.73**	**61.47**	**10.62**	**74.42**	**26.08**	**84.82**
Ff&Lf	VariantE		✓	✓	✓	✓		13.93	97.98	44.91	60.10	18.56	73.47	31.34	81.44
	VariantF		✓	✓	✓		✓	18.67	98.04	41.17	61.03	14.72	71.43	30.08	82.46

Table 3. Ablation study of the loss constraints in our disentanglement learning (Dl) module, and the effectiveness of our feature fusion (Ff) and loss flattening (Lf). 'Cls', 'Rec', and 'Con' represent our classification loss, reconstruction loss, and contrastive loss, respectively. 'Cls(CE)' means we replace our demographic distribution-aware margin loss with cross-entropy loss. All methods are only trained on FF++.

amine the fairness generalization capability of our proposed method concerning backbone selection, we substitute the Xception backbone with ResNet-50 [64] and EfficientNet-B3 [65]. The results in Table 2 indicate that our method based on different backbones shows similar superior results. Such outcomes suggest that our proposed approach is not limited to backbone choice, but is effective and applicable to diverse backbone settings.

5.3. Ablation Study

Effects of Components in Disentanglement Learning. The results of VariantA/B/C/D in Table 3 demonstrate the contribution of each loss constraint in our disentanglement learning (Dl) module. Without reconstructive loss and contrastive loss, VariantA shows relatively lower performance on both F_{FPR} and AUC compared with other Variants and Ours. VariantB and VariantC underscore the value of our reconstructive loss (*e.g.*, F_{FPR} drops 5.47% and the AUC increases 2.69% on Celeb-DF) and contrastive loss (*e.g.*, F_{FPR} drops 6.32% and the AUC increases 4.03% on Celeb-DF), respectively. Comparing Ours with VariantD demonstrates the impact of our demographic distribution-aware margin loss. By replacing CE loss with the demographic distribution-aware margin loss, the F_{FPR} reduces 5.95% and the AUC improves 0.23% on FF++. The similar tread is also observed on three other datasets.

Effects of Feature Fusion (Ff) and Loss Flattening (Lf). The results of VariantE/F in Table 3 reveal the effects of our feature fusion (Ff) and loss flattening (Lf) methods. When comparing ours with VariantE (without Lf), the F_{FPR} is enhanced by 7.94% on Celeb-DF and 4.18% on DFDC. While ours against VariantF (without Ff), the F_{FPR} is improved by 4.10% and 0.44% on those two datasets. This indicates that Lf boosts the model's fairness generalization more than Ff. Overall, our method with both Ff and Lf yields the most substantial gains in fairness and AUC across all datasets.

Comparison on Intersectional Subgroups. We present detailed results of the False Positive Rate (FPR) on each subgroup across all datasets, as shown in Fig. 4 (left). The results clearly indicate that our approach significantly narrows the disparity between these subgroups, *e.g.*, the maximum FPR gap of DAW-FDD on Celeb-DF is 20.6, while our method lowers the gap to 9.3. Overall, ours leads to a consistent and marked reduction in the FPR across all test datasets.

5.4. Visualization

Visualization of Loss Landscape. Fig. 4 (right) visually illustrates our method's loss landscape. Without the flatten-

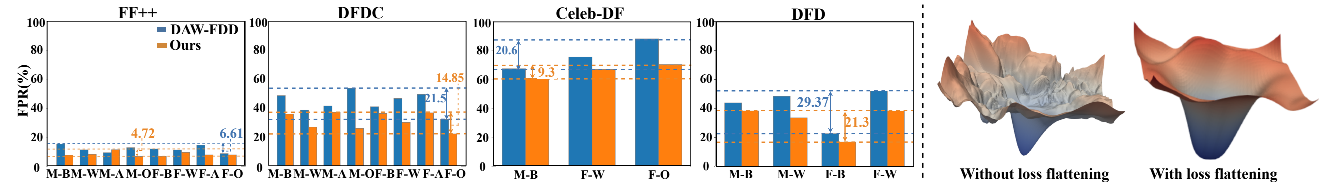

Figure 4. (**Left**) Comparison of FPR on Intersectional subgroups. Models are trained on FF++ and tested on FF++, DFDC, Celeb-DF, and DFD. The subgroups not represented in Celeb-DF and DFD are inapplicable. (**Right**) The loss landscape visualization of our proposed method with (right) and without (left) flattening the loss landscape.

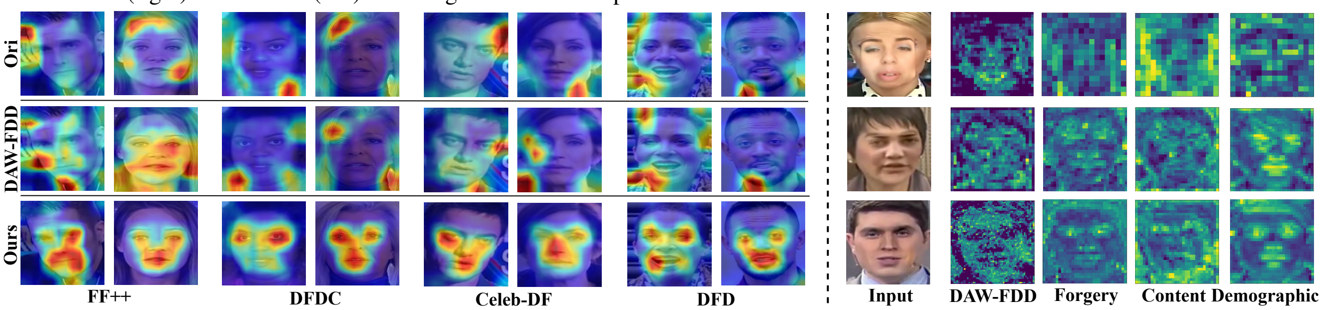

Figure 5. (**Left**) Grad-CAM visualization of Ori's (first row), DAW-FDD (second row), and ours (third row) on the intra-domain dataset (FF++), and cross-domain datasets (DFDC, Celeb-DF, and DFD). (**Right**) Visualization of the image (first column), DAW-FDD's features (second column), ours disentangled forgery (third column), content (fourth column), and demographic features (last column).

ing process, the landscape is sharp with numerous peaks and valleys. Such sharpness may trap the model into different suboptimal minima, leading to inconsistent generalization. However, after flattening, the landscape becomes smoother, suggesting an easier optimization path, potentially leading to better training and generalization. This visualization underscores the significance of Joint Optimization in our method for enhancing fairness generalization.

Visualization of the Saliency Map. To more intuitively demonstrate the effectiveness of our method, we visualize the Grad-CAM [66] of Ori, DAW-FDD [17], and our method, respectively, as shown in Fig. 5 (left). Grad-CAM shows that the Ori without any constraints, is prone to overfitting to small local regions or focusing on content noise outside the facial region. DAW-FDD has the fair loss as a constraint that performs well in intra-domain. Once the data is unseen, it loses fair detection ability and its Grad-CAM shows similar results as Ori's. On the contrary, our method's activation region demonstrates a consistent model focus on facial salient features, irrespective of the dataset.

Visualization of Features. The feature visualization in Fig. 5 (right) reveals key insights into the focus areas of DAW-FDD and our method. DAW-FDD's abstracted patterns and highlighted regions (second column) show a broad emphasis on facial features without specific targeting. In contrast, our disentangled features demonstrate distinct areas of focus: the forgery features (third column) and demographic features (last column) predominantly highlight facial areas, whereas the content features (fourth column) are oriented towards the background. This differentiation underscores the importance of integrating forgery and demographic features, and eliminating content features, to foster fairer learning.

6. Conclusion

While current methods for enhancing fairness in deepfake detection perform well within a specific domain, they struggle to maintain fairness when tested across different domains. Recognizing this limitation, we introduce an innovative framework designed to address the fairness generalization challenge in deepfake detection. By combining disentanglement learning and fair learning modules, our approach ensures both generalizability and fairness. Furthermore, we incorporate a loss flattening strategy to streamline the optimization process for these modules, resulting in robust fairness generalization. Experimental results on diverse deepfake datasets showcase the superior fairness maintenance capabilities of our method across various domains.

Limitation. One limitation of our method is its dependency on datasets including forged videos generated by multiple manipulation techniques. However, there exist few deepfake datasets that do not have such characteristics.

Future Work. We aim to design a method that can preserve fairness not rely on multi-forged data, but can directly detect images generated by diffusion or GANs. In addition, we plan to enhance fairness across not just video datasets, but also in a multi-modal context.

Acknowledgments. Shu Hu is supported by the Purdue University start-up grant.

References

[1] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19667–19679, 2020. 1

[2] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020. 1

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021. 1

[4] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, "Gan-generated faces detection: A survey and new perspectives," *ECAI*, 2023. 1

[5] 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. 1

[6] Y. Ju, S. Hu, S. Jia, G. H. Chen, and S. Lyu, "Improving fairness in deepfake detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4655–4665, 2024. 1, 2, 3, 5, 6, 7, 12

[7] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of gan-generated images," in *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6, IEEE, 2019. 1, 2, 12

[8] M. Goebel, L. Nataraj, T. Nanjundaswamy, T. M. Mohammed, S. Chandrasekaran, and B. Manjunath, "Detection, attribution and localization of gan generated images," *arXiv preprint arXiv:2007.10466*, 2020.

[9] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

[10] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.

[11] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting cnn-generated facial images in real-world scenarios," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 642–643, 2020.

[12] 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.

[13] 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. 2, 12

[14] J. Hu, S. Wang, and X. Li, "Improving the generalization ability of deepfake detection via disentangled representation learning," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3577–3581, IEEE, 2021. 2, 3, 12

[15] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 641–657, Springer, 2020. 2, 12

[16] J. Liang, H. Shi, and W. Deng, "Exploring disentangled content information for face forgery detection," in *European Conference on Computer Vision*, pp. 128–145, Springer, 2022. 2, 12

[17] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22412–22423, October 2023. 2, 3, 4, 6, 7, 8, 12

[18] P. Zheng, H. Chen, S. Hu, B. Zhu, J. Hu, C.-S. Lin, X. Wu, S. Lyu, G. Huang, and X. Wang, "Few-shot learning for misinformation detection based on contrastive models," *Electronics*, vol. 13, no. 4, p. 799, 2024.

[19] T. Chen, S. Yang, S. Hu, Z. Fang, Y. Fu, X. Wu, and X. Wang, "Masked conditional diffusion model for enhancing deepfake detection," *arXiv preprint arXiv:2402.00541*, 2024.

[20] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu, "Detecting multimedia generated by large ai models: A survey," *arXiv preprint arXiv:2402.00045*, 2024.

[21] B. Fan, S. Hu, and F. Ding, "Synthesizing black-box anti-forensics deepfakes with high visual quality," *ICASSP*, 2024.

[22] 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.

[23] 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.

[24] B. Fan, Z. Jiang, S. Hu, and F. Ding, "Attacking identity semantics in deepfakes via deep feature fusion," in *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 114–119, IEEE, 2023.

[25] H. Chen, P. Zheng, X. Wang, S. Hu, B. Zhu, J. Hu, X. Wu, and S. Lyu, "Harnessing the power of text-image contrastive models for automatic detection of online misinformation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 923–932, 2023. 1

[26] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection," *IJCAI*, 2021. 1, 2, 12

[27] 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, 12

[28] K. Wiggers, "Deepfake detectors and datasets exhibit racial and gender bias, usc study shows," in *VentureBeat*, `https://tinyurl.com/ms8zbu6f`, 2021.

[29] A. V. Nadimpalli and A. Rattani, "Gbdf: gender balanced deepfake dataset towards fair deepfake detection," *arXiv preprint arXiv:2207.10246*, 2022. 2, 4, 12

[30] 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, 2, 12

[31] 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, 12

[32] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *arXiv preprint arXiv:2211.11695*, 2022. 2, 12

[33] 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, 12

[34] 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*, pp. 1–11, 2019. 2, 3, 6, 7

[35] Google and Jigsaw, "Deepfakes dataset by google & jigsaw," in `https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html`, 2019. 2, 6

[36] H. Wang, L. He, R. Gao, and F. P. Calmon, "Aleatoric and epistemic discrimination in classification," *ICML*, 2023. 2, 5, 6

[37] 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. 2, 5, 6

[38] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," *Advances in neural information processing systems*, vol. 32, 2019. 3

[39] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2020. 3, 5, 14

[40] "Deepfakes," in `https://github.com/deepfakes/faceswap`, 2017. 3, 6

[41] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016. 3, 6

[42] M. Kowalski, "Faceswap," in `https://github.com/MarekKowalski/FaceSwap/`, 2018. 3, 6

[43] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019. 3, 6

[44] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, pp. 2, 5, 2019. 3, 6

[45] S. Mathews, S. Trivedi, A. House, S. Povolny, and C. Fralick, "An explainable deepfake detection framework on a novel unconstrained dataset," *Complex & Intelligent Systems*, pp. 1–13, 2023. 4

[46] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 322–330, 2019. 4

[47] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019. 4, 17

[48] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 4

[49] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017. 5, 13

[50] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012. 5

[51] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016. 5

[52] 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

[53] 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.

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

[55] 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.

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

[57] S. Hu, Z. Yang, X. Wang, Y. Ying, and S. Lyu, "Outlier robust adversarial training," *ACML*, 2023. 5

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

[59] D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford, "Large-scale methods for distributionally robust optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8847–8860, 2020. 5

[60] "Deepfake detection challenge." `https://www.kaggle.com/c/deepfake-detection-challenge`. Accessed: 2021-04-24. 6

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

[62] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*, 2021. 6

[63] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pp. 1251–1258, 2017. 6, 7, 13

[64] 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. 7

[65] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019. 7

[66] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. 8

[67] S. Hu, Y. Li, and S. Lyu, "Exposing gan-generated faces using inconsistent corneal specular highlights," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504, IEEE, 2021. 12

[68] H. Guo, S. Hu, X. Wang, M.-C. Chang, and S. Lyu, "Eyes tell all: Irregular pupil shapes reveal gan-generated faces," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2904–2908, IEEE, 2022. 12

[69] 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 *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 224–227, IEEE, 2022. 12

[70] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE International workshop on information forensics and security (WIFS)*, pp. 1–7, IEEE, 2018. 12

[71] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, IEEE, 2019. 12

[72] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing gan-synthesized faces using landmark locations," in *Proceedings of the ACM workshop on information hiding and multimedia security*, pp. 113–118, 2019. 12

[73] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *European conference on computer vision*, pp. 86–103, Springer, 2020. 12

[74] M. Khayatkhoei and A. Elgammal, "Spatial frequency bias in convolutional generative adversarial networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 7152–7159, 2022. 12

[75] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," *Advances in neural information processing systems*, vol. 33, pp. 3022–3032, 2020. 12

[76] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*, pp. 3247–3258, PMLR, 2020. 12

[77] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6, IEEE, 2019. 12

[78] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018. 17