AI-Face: A Million-Scale Demographically Annotated AI-Generated Face Dataset and Fairness Benchmark[*]

Li Lin[1], Santosh[1], Mingyang Wu[1], Xin Wang[2], Shu Hu[1†]
[1]Purdue University, West Lafayette, USA {lin1785, santosh2, wu2415, hu968}@purdue.edu
[2]University at Albany, State University of New York, New York, USA xwang56@albany.edu

Abstract

AI-generated faces have enriched human life, such as entertainment, education, and art. However, they also pose misuse risks. Therefore, detecting AI-generated faces becomes crucial, yet current detectors show biased performance across different demographic groups. Mitigating biases can be done by designing algorithmic fairness methods, which usually require demographically annotated face datasets for model training. However, ***no existing dataset encompasses both demographic attributes and diverse generative methods simultaneously****, which hinders the development of fair detectors for AI-generated faces. In this work, we introduce the* ***AI-Face*** *dataset, the first million-scale demographically annotated AI-generated face image dataset, including real faces, faces from deepfake videos, and faces generated by Generative Adversarial Networks and Diffusion Models. Based on this dataset, we conduct the first comprehensive fairness benchmark to assess various AI face detectors and provide valuable insights and findings to promote the future fair design of AI face detectors. Our AI-Face dataset and benchmark code are publicly available at* `https://github.com/Purdue-M2/AI-Face-FairnessBench`.

1. Introduction

AI-generated faces are created using sophisticated AI technologies that are visually difficult to discern from real ones [1]. They can be summarized into three categories: deepfake videos [2] created by typically using Variational Autoencoders (VAEs) [3, 4], faces generated from Generative Adversarial Networks (GANs) [5–8], and Diffusion Models (DMs) [9]. These technologies have significantly advanced the realism and controllability of synthetic facial representations. Generated faces can enrich media and increase creativity [10]. However, they also carry significant risks of misuse. For example, during the 2024 United States presidential election, fake face images of Donald Trump surrounded by groups of black people smiling and laughing to encourage African Americans to vote Republican are spreading online [11]. This could distort public opinion and erode people's trust in media [12, 13], necessitating the detection of AI-generated faces for their ethical use.

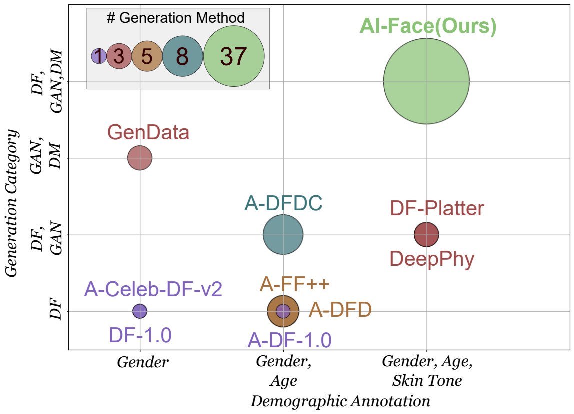


Figure 1. *Comparison between AI-Face and other datasets in terms of demographic annotation, generation category, and the number of generation methods. 'DF', 'GAN', and 'DM' stand for Deepfake Videos, Generative Adversarial Networks, and Diffusion Models.*

However, one major issue existing in current AI face detectors [24–27] is biased detection (*i.e*., unfair detection performance among demographic groups [19, 28–30]). Mitigating biases can be done by designing algorithmic fairness methods, but they usually require demographically annotated face datasets for model training. For example, works like [29, 30] have made efforts to enhance fairness in the detection based on A-FF++ [19] and A-DFD [19]. However, both datasets are limited to containing only faces from deepfake videos, which could cause the trained models not to be applicable for fairly detecting faces generated by GANs and DMs. While some datasets (*e.g*., GenData [17], DF40 [31]) include GAN and DM faces, they either lack demographic annotations or provide only limited demographic attributes. Most importantly, no existing dataset offers sufficient diversity in generation methods while also providing demographic labels. A comparison of existing datasets is shown in Fig. 1. *These limitations of existing datasets hamper the development of fair technologies for detecting AI-generated faces.*

[*] This paper has been accepted by CVPR 2025
[†] Corresponding author

Dataset	Year	Face Images		Generation Category			#Generation	Source of Real Images	Demographic Annotation		
		#Real	#Fake	Deepfake Videos	GAN	DM	Methods		Skin Tone	Gender	Age
DF-1.0 [14]	2020	2.9M	14.7M	✓			1	Self-Recording		✓	
DeePhy [15]	2022	1K	50.4K	✓	✓		3	YouTube	✓	✓	✓
DF-Platter [16]	2023	392.3K	653.4K	✓	✓		3	YouTube	✓	✓	✓
GenData [17]	2023	-	20K		✓	✓	3	CelebA [18]		✓	
A-FF++ [19]	2024	29.8K	149.1K	✓			5	YouTube		✓	✓
A-DFD [19]	2024	10.8K	89.6K	✓			5	Self-Recording		✓	✓
A-DFDC [19]	2024	54.5K	52.6K	✓	✓		8	Self-Recording		✓	✓
A-Celeb-DF-v2 [19]	2024	26.3K	166.5K	✓			1	Self-Recording		✓	
A-DF-1.0 [19]	2024	870.3K	321.5K	✓			1	Self-Recording		✓	✓
AI-Face	2025	400K	1.2M	✓	✓	✓	37	FFHQ [6], IMDB-WIKI [20], real from FF++ [2], DFDC [21], DFD [22],Celeb-DF-v2 [23]	✓	✓	✓

Table 1. *Quantitative comparison of existing datasets with ours on demographically annotated AI-generated faces.*

Moreover, benchmarking fairness provides a direct method to uncover prevalent and unique fairness issues in recent AI-generated face detection. However, there is a lack of a comprehensive benchmark to estimate the fairness of existing AI face detectors. Existing benchmarks [32–35] primarily assess utility, neglecting systematic fairness evaluation. Two studies [28, 36] do evaluate fairness in detection models, but their examination is based on a few outdated detectors. Furthermore, detectors' fairness reliability (*e.g.*, robustness with test set post-processing, fairness generalization) has not been assessed. *The absence of a comprehensive fairness benchmark impedes a thorough understanding of the fairness behaviors of recent AI face detectors and obscures the research path for detector fairness guarantees.*

In this work, we build the **first** million-scale demographically annotated AI-generated face image dataset: **AI-Face**. The face images are collected from various public datasets, including the real faces that are usually used to train AI face generators, faces from deepfake videos, and faces generated by GANs and DMs. Each face is demographically annotated by our designed measurement method and Contrastive Language-Image Pretraining (CLIP) [37]-based lightweight annotator. Next, we conduct the **first** comprehensive fairness benchmark on our dataset to estimate the fairness performance of 12 representative detectors coming from four model types. Our benchmark exposes common and unique fairness challenges in recent AI face detectors, providing essential insights that can guide and enhance the future design of fair AI face detectors. Our contributions are as follows:

- We build the first million-scale demographically annotated AI-generated face dataset by leveraging our designed measure and developed lightweight annotator.
- We conduct the first comprehensive fairness benchmark of AI-generated face detectors, providing an extensive fairness assessment of current representative detectors.
- Based on our experiments, we summarize the unsolved questions and offer valuable insights within this research field, setting the stage for future investigations.

2. Background and Motivation

AI-generated Faces and Biased Detection. AI-generated face images, created by advanced AI technologies, are visually difficult to discern from real ones. They can be summarized into three categories: *1) Deepfake Videos.* Initiated in 2017 [13], these use face-swapping and face-reenactment techniques with a variational autoencoder to replace a face in a target video with one from a source [3, 4]. Note that our paper focuses solely on images extracted from videos. *2) GAN-generated Faces.* Post-2017, Generative Adversarial Networks (GANs) [38] like StyleGANs [6–8] have significantly improved generated face realism. *3) DM-generated Faces.* Diffusion models (DMs), emerging in 2021, generate detailed faces from textual descriptions and offer greater controllability. Tools like Midjourney [39] and DALLE2 [40] facilitate customized face generation. While these AI-generated faces can enhance visual media and creativity [10], they also pose risks, such as being misused in social media profiles [41, 42]. Therefore, numerous studies focus on detecting AI-generated faces [24–27], but current detectors often show performance disparities among demographic groups [19, 28–30]. This bias can lead to unfair targeting or exclusion, undermining trust in detection models. Recent efforts [29, 30] aim to enhance fairness in deepfake detection but mainly address deepfake videos, overlooking biases in detecting GAN- and DM-generated faces.

The Existing Datasets. Current AI-generated facial datasets with demographic annotations are limited in *size*, *generation categories*, *methods*, and *annotations*, as illustrated in Table 1. For instance, A-FF++, A-DFD, A-DFDC, and A-Celeb-DF-v2 [19] are deepfake video datasets with fewer than one million images. Datasets like DF-1.0 [14] and DF-Platter [16] lack various demographic annotations. Additionally, existing datasets offer limited generation methods. These limitations hinder the development of fair AI face detectors, motivating us to build a million-scale demographically annotated AI-Face dataset.

Benchmark for Detecting AI-generated Faces. Benchmarks are essential for evaluating AI-generated face detectors under standardized conditions. Existing benchmarks, as shown in Table 2, mainly focus on detectors' utility, often overlooking fairness [31–35]. Loc et al. [28] and CCv1 [36] examined detector fairness. However, their study did not have an analysis on DM-generated faces and only measured bias between groups in basic scenarios without considering

Existing Benchmarks	Year	Category			Scope of Benchmark		
						Fairness	
		Deepfake Videos	GAN	DM	Utility	General	Reliability
Loc et al. [28]	2021	✓			✓	✓	
CCv1 [36]	2021	✓	✓		✓	✓	
DeepfakeBench [34]	2023	✓	✓		✓		
CDDB [32]	2023		✓		✓		
Lin et al. [33]	2024	✓	✓		✓		
Le et al. [35]	2024	✓	✓		✓		
DF40 [31]	2024	✓	✓	✓	✓		
Ours	2025	✓	✓	✓	✓	✓	✓

Table 2. *Comparison of existing AI-generated face detection benchmarks and ours. Fairness 'General' means fairness evaluation under default/basic settings. Fairness 'Reliability' measures fairness consistency across dynamic scenarios (*e.g.*, post-processing).*

fairness reliability under real-world variations and transformations. This motivates us to conduct a comprehensive benchmark to evaluate AI face detectors' fairness.

The Definition of Demographic Categories. Demography-related labels are highly salient to measuring bias. Following prior works [36, 43–47], we will focus on three key demographic categories: ***Skin Tone***, ***Gender***, *and* ***Age***, in this work. For skin tone, this vital attribute spans a range from pale to dark. We use the Monk Skin Tone scale [48], specifically designed for computer vision applications. For gender, we adopt binary categories (*i.e.*, Male and Female), following practices by many governments [49, 50] and facial recognition research [45, 51, 52], based on sex at birth. For age, using definitions from the United Nations [53] and Statistics Canada [54], we define five age groups: Child (0-14), Youth (15-24), Adult (25-44), Middle-age Adult (45-64), and Senior (65+). More discussion is in Appendix A.

3. AI-Face Dataset

This section outlines the process of building our demographically annotated AI-Face dataset (see Fig. 2), along with its statistics and annotation quality assessment.

3.1. Data Collection

We build our AI-Face dataset by collecting and integrating public real and AI-generated face images sourced from academic publications, GitHub repositories, and commercial tools. We strictly adhere to the license agreements of all datasets to ensure that they allow inclusion in our datasets and secondary use for training and testing. More details are in Appendix B.1. Specifically, the fake face images in our dataset originate from **4 Deepfake Video datasets** (*i.e.*, FF++ [2], DFDC [21], DFD [22], and Celeb-DF-v2 [23]), generated by **10 GAN** models (*i.e.*, AttGAN [55], MMDGAN [56], StarGAN [55], StyleGANs [55, 57, 58], MSGGAN [56], ProGAN [59], STGAN [56], and VQ-GAN [60]), and **8 DM** models (*i.e.*, DALLE2 [61], IF [61], Midjourney [61], DCFace [62], Latent Diffusion [63], Palette [64], Stable Diffusion v1.5 [65], Stable Diffusion Inpainting [65]). This constitutes a total of 1,245,660 fake face images in our dataset. We include **6 real** source datasets (*i.e.*, FFHQ [6], IMDB-WIKI [20], and real images from FF++ [2], DFDC [21], DFD [22], and Celeb-DF-v2 [23]). All of them are usually used as a training set for generative models to generate fake face images. This constitutes a total of 400,885 real face images in our dataset. In general, our dataset contains 28 subsets and 37 generation methods (*i.e.*, 5 in FF++, 5 in DFD, 8 in DFDC, 1 in Celeb-DF-v2, 10 GANs, and 8 DMs). For all images, we use RetinaFace [66] for detecting and cropping faces.

3.2. Annotation Generation

3.2.1. Skin Tone Annotation Generation

Skin tone is typically measured using an intuitive approach [67, 68], without requiring a predictive model. Inspired by [67], we developed a method to estimate skin tone using the Monk Skin Tone (MST) Scale [48] (including 10-shade scales: Tone 1 to 10) by combining facial landmark detection with color analysis. Specifically, utilizing Mediapipe's FaceMesh [69] for precise facial landmark localization, we isolate skin regions while excluding non-skin areas such as the eyes and mouth. Based on the detected landmarks, we generate a mask to extract skin pixels from the facial area. These pixels are then subjected to K-Means clustering [70] (we use K= 3 in practice) to identify the dominant skin color within the region of interest. The top-1 largest color cluster is mapped to the closest tone in the MST Scale by calculating the Euclidean distance between the cluster centroid and the MST reference colors in RGB space.

3.2.2. Gender and Age Annotation Generation

For generating gender and age annotations, the existing online software (*e.g.*, Face++ [71]) and open-source tools (*e.g.*, InsightFace [72]) can be used for the prediction. However, they fall short in our task due to two reasons: *1)* They are mostly designed for face recognition and trained on datasets of real face images but lack generalization capability for annotating AI-generated face images. *2)* Their use may introduce bias into our dataset, as they are typically designed and trained without careful consideration of bias and imbalance in the training set. See Appendix B.3 for our experimental study on these tools. To this end, we have to develop our specific annotators to predict gender and age annotations for each image in our dataset.

Problem Definition. Given a training dataset $\mathbb{D} = \{(X_i, A_i)\}_{i=1}^n$ with size n, where X_i represents the i-th face image and A_i signifies a demographic attribute associated with X_i. Here, $A_i \in \mathcal{A}$, where $\mathcal{A}$ represents user-defined groups (*e.g.*, for gender, $\mathcal{A} = \{\text{Female, Male}\}$. For age, $\mathcal{A} = \{\text{Child, Youth, Adult, Middle-age Adult, Senior}\}$). Our goal is to design a lightweight, generalizable annotator based on $\mathbb{D}$ that reduces bias while predicting facial demographic attributes for each image in our dataset. In practice, we use IMDB-WIKI [20] as training dataset, which contains

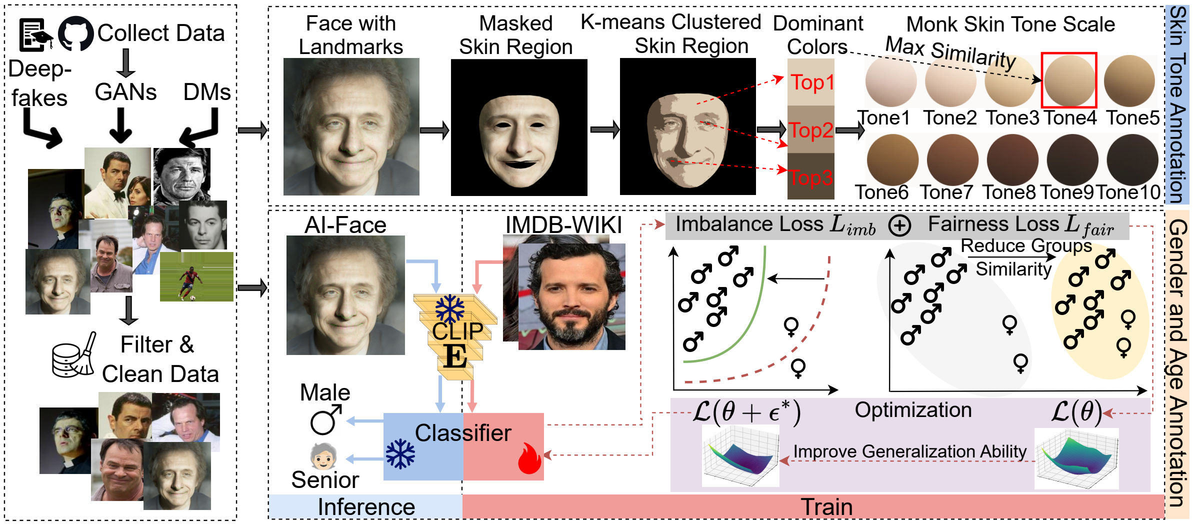


Figure 2. *Generation pipeline of our Demographically Annotated AI-Face Dataset. First, we collect and filter face images from Deepfake Videos, GAN-generated faces, and DM-generated faces found in public datasets. Second, we perform skin tone, gender, and age annotation generation. Skin tone is estimated by combining facial landmark detection with color analysis to generate the corresponding annotation. For gender and age, we develop annotators trained on the IMDB-WIKI dataset [20], then use them to predict attributes for each image.*

images along with profile metadata sourced from IMDb and Wikipedia, *ensuring that the demographic annotations are as accurate as possible*. We trained two annotators with identical architecture and training procedures for gender and age annotations, respectively.

Annotator Architecture. We build a lightweight annotator based on the CLIP [37] foundation model by leveraging its strong zero-shot and few-shot learning capabilities. Specifically, our annotator employs a frozen pre-trained CLIP ViT L/14 [73] as a feature extractor $\mathbf{E}$ followed by a trainable classifier parameterized by θ, which contains 3-layer Multi-layer Perceptron (MLP) $\mathbf{M}$ and a classification head h.

Learning Objective. Aware that neural networks can perform poorly when the training dataset suffers from class-imbalance [74] and CLIP is not free from demographic bias [75–77], we introduce an imbalance loss and fairness loss to address these challenges in the annotator training. Specifically, for image X_i, its feature f_i is obtained through f_i=$\mathbf{M}(\mathbf{E}(X_i))$. Next, two losses are detailed below.

Imbalance Loss: To mitigate the impact of imbalance data, we use Vector Scaling [78] loss, which is a re-weighting method for training models on the imbalanced data with distribution shifts and can be expressed as

$$L_{imb} = \frac{1}{n}\sum_{i=1}^{n} -u_{A_i} \log \frac{e^{\zeta_{A_i} h(f_i)_{A_i} + \Delta_{A_i}}}{\sum_{A\in\mathcal{A}} e^{\zeta_A h(f_i)_A + \Delta_A}},$$

where u_{A_i} is the weighting factor for attribute A_i. $h(f_i)_{A_i}$ is the predict logit on A_i. ζ_{A_i} is the multiplicative logit scaling factor, calculated as the inverse of A_i's frequency. Δ_{A_i} is the additive logit scaling factor, calculated as the log of A_i probabilities. More details about them are in appendix B.4.

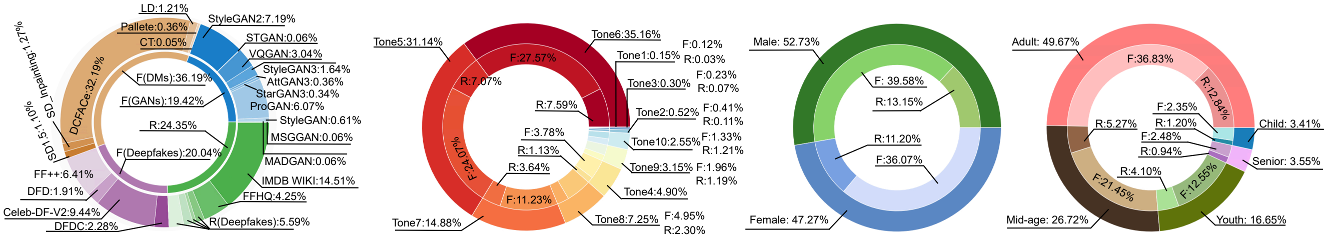
Fairness Loss: We introduce a fairness loss to minimize the disparity between the distribution $\mathcal{D}^f$ of f and the conditional distribution $\mathcal{D}^{f_A}$ of f on attribute $A \in \mathcal{A}$. Specifically, we follow [79, 80] to minimize the summation of the following Sinkhorn distance between these two distributions:

$$L_{fair} = \sum_{A\in\mathcal{A}} \inf_{\gamma\in\Gamma(\mathcal{D}^f,\mathcal{D}^{f_A})} \{\mathbb{E}_{X\sim\gamma}[c(p,q)]+\alpha H(\gamma|\mu\otimes\nu)\},$$

where $\Gamma(\mathcal{D}^f, \mathcal{D}^{f_A})$ is the set of joint distributions based on $\mathcal{D}^f$ and $\mathcal{D}^{f_A}$. Let p and q be the points from $\mathcal{D}^f$ and $\mathcal{D}^{f_A}$, respectively. Then, $c(p, q)$ represents the transport cost [80]. Let μ and ν be the reference measures from the set of measures on f. Then, $H(\gamma|\mu \otimes \nu)$ represents the relative entropy of γ with respect to the product measure $\mu \otimes \nu$. $\alpha \geq 0$ is a regularization hyperparameter. In practice, we use the empirical form of L_{fair}.

Total Loss: Therefore, the final learning objective becomes $\mathcal{L}(\theta) = L_{imb} + \lambda L_{fair}$, where λ is a hyperparameter.

Train. Traditional optimization methods like stochastic gradient descent can lead to poor model generalization due to sharp loss landscapes with multiple local and global minima. To address this, we use Sharpness-Aware Minimization (SAM) [81] to enhance our annotator's generalization by flattening the loss landscape. Specifically, flattening is attained by determining the optimal ϵ^* for perturbing model parameters θ to maximize the loss, formulated as: $\epsilon^* = \arg\max_{\|\epsilon\|_2\leq\beta} \mathcal{L}(\theta+\epsilon) \approx \arg\max_{\|\epsilon\|_2\leq\beta} \epsilon^\top \nabla_\theta \mathcal{L} = \beta \mathtt{sign}(\nabla_\theta \mathcal{L})$, where β controls the perturbation magnitude. The approximation is based on the first-order Taylor expansion with assuming ϵ is small. The final equation is obtained by solving a dual norm problem, where $\mathtt{sign}$ represents a sign function and $\nabla_\theta \mathcal{L}$ being the gradient of $\mathcal{L}$ with respect to θ. As a result, the model parameters are updated by solving: $\min_\theta \mathcal{L}(\theta + \epsilon^*)$.



(a) Subset (b) Skin Tone (c) Gender (d) Age

Figure 3. *Distribution of face images of the AI-Face dataset. The figure shows the (a) subset distribution and the demographic distribution for (b) skin tone, (c) gender, and (d) gender. The outer rings in (b), (c), and (d) represent the proportion of groups within each attribute category, while the inner rings indicate the distribution of fake (F) and real (R) images within those groups.*

Inference. We use the trained annotators to predict demographic labels for each image in AI-Face dataset, except for those from IMDB-WIKI, which already contain true labels.

3.3. Dataset Statistics

Fig. 3 illustrates the subset distribution and demographic attributes of the AI-Face dataset. The dataset contains approximately three times more generated images than real images, with diffusion model-generated images constituting the majority. In terms of demographic attributes, the majorities in skin tone are Tone 5 (31.14%) and Tone 6 (35.16%). The lightest skin tones (Tones 1-3) are underrepresented, comprising only 0.97% of the dataset. The dataset is relatively balanced across gender. Adult (25-44) (49.67%) is the predominant representation in age groups.

3.4. Annotation Quality Assessment

To assess the quality of demographic annotations in our AI-Face dataset, we conducted a user study. Three participants label the demographic attributes for the given images (the details of labeling activities are in appendix B.5), with the final ground truth determined by majority vote. We then compare our annotations with those in A-FF++, A-DFDC, A-CelebDF-V2, and A-DFD datasets. Specifically, we perform two assessments: *1) Strategic comparison*: We select 1,000 images from A-FF++ and A-DFDC that have different annotations from AI-Face. These images likely represent challenging cases. *2) Random comparison*: We randomly sampled 1,000 images from A-Celeb-DF-V2 and A-DFD. Due to the limited age classes in these datasets, only gender was evaluated. The results, presented in Table 3, demonstrate the high correctness of the AI-Face annotations and their superior quality compared to the annotations of other datasets. For example, our annotation quality (ACC) surpasses those in A-FF++ by 78.714% on gender and 48.000% on age.

4. Fairness Benchmark Settings

This section demonstrates the fairness benchmark settings for detection methods and evaluation metrics on AI-Face (80%/20%: Train/Test). More settings are in Appendix C.1.

Detection Methods. Our benchmark has implemented 12 detectors. The methodologies cover a spectrum that is specifically tailored to detect AI-generated faces from Deepfake Videos, GANs, and DMs. They can be classified into four types: *Naive detectors:* refer to backbone models that can be directly utilized as the detector for binary classification, including CNN-based (*i.e.*, Xception [82], EfficientB4 [83]) and transformer-based (*i.e.*, ViT-B/16 [84]). *Frequency-based:* explore the frequency domain for forgery detection (*i.e.*, F3Net [85], SPSL [86], SRM [87]). *Spatial-based:* focus on mining spatial characteristics (*e.g.*, texture) within images for detection (*i.e.*, UCF [26], UnivFD [88], CORE [89]). *Fairness-enhanced:* focus on improving fairness in AI-generated face detection by designing specific algorithms (*i.e.*, DAW-FDD [29], DAG-FDD [29], PG-FDD [30]).

Evaluation Type	Dataset	Gender ACC	Gender Precision	Gender Recall	Age ACC	Age Precision	Age Recall
Strategic	A-FF++	8.143	17.583	5.966	37.700	39.459	45.381
	AI-Face	**86.857**	**74.404**	**77.367**	**85.700**	**74.024**	**63.751**
	A-DFDC	21.600	28.604	23.082	33.400	38.011	40.165
	AI-Face	**91.700**	**92.129**	**83.448**	**77.000**	**76.184**	**62.646**
Random	A-Celeb-DF-V2	89.628	90.626	90.494		-	
	AI-Face	**91.206**	**91.474**	**91.767**		-	
	A-DFD	70.900	71.686	74.435		-	
	AI-Face	**92.300**	**91.060**	**91.727**		-	

Table 3. *Annotation quality assessment results (%) for A-FF++, A-DFDC, A-Celeb-DF-V2, A-DFD, and our AI-Face. ACC: Accuracy.*

Evaluation Metrics. To provide a comprehensive benchmarking, we consider 5 fairness metrics commonly used in fairness community [90–94] and 5 widely used utility metrics. For *fairness* metrics, we consider Demographic Parity (F_{DP}) [90, 91], Max Equalized Odds (F_{MEO}) [93], Equal Odds (F_{EO}) [92], and Overall Accuracy Equality (F_{OAE}) [93] for evaluating group (*e.g.*, gender) and intersectional (*e.g.*, individuals of a specific gender and simultaneously a specific skin tone) fairness. In experiments, the intersectional groups are Female-Light (F-L), Female-Medium (F-M), Female-Dark (Dark), Male-Light (M-L), Male-Medium (M-M), and Male-Dark (M-D), where we group 10 categories of skin tones into Light (Tone 1-3), Medium (Tone 4-6), and Dark (Tone 7-10) for simplicity according to [95]. We also use individual fairness (F_{IND}) [94, 96] (*i.e.*, similar individuals should have similar predicted outcomes) for estimation. For *utility* metrics, we employ the Area Under the ROC Curve (AUC), Accuracy (ACC), Average Precision (AP), Equal Error Rate (EER),

Measure	Attribute	Metric	Model Type											
			Naive			Frequency			Spatial			Fairness-enhanced		
			Xception [82]	EfficientB4 [83]	ViT-B/16 [84]	F3Net [85]	SPSL [86]	SRM [87]	UCF [26]	UnivFD [88]	CORE [89]	DAW-FDD [29]	DAG-FDD [29]	PG-FDD [30]
Fairness(%)↓	Skin Tone	F_{MEO}	8.836	8.300	6.264	19.938	8.055	10.002	17.325	**2.577**	10.779	14.118	6.551	6.465
		F_{DP}	9.751	**6.184**	7.728	12.876	9.379	10.897	12.581	8.556	10.317	10.706	8.617	9.746
		F_{OAE}	1.271	4.377	2.168	2.818	1.135	0.915	1.883	2.748	1.332	1.667	1.388	**0.882**
		F_{EO}	12.132	11.062	8.813	23.708	9.789	14.239	21.92	**5.536**	13.069	16.604	7.383	9.115
	Gender	F_{MEO}	3.975	5.385	5.104	4.717	4.411	6.271	5.074	4.503	5.795	5.510	5.910	**3.190**
		F_{DP}	1.691	1.725	1.344	1.864	1.827	1.957	1.736	**1.190**	2.154	2.015	2.151	1.252
		F_{OAE}	**0.975**	1.487	1.803	1.129	1.037	1.772	1.451	1.622	1.389	1.325	1.420	1.071
		F_{EO}	4.143	5.863	6.031	4.870	4.534	6.78	5.510	5.408	5.931	5.696	6.066	**3.702**
	Age	F_{MEO}	27.883	6.796	14.937	38.801	27.614	24.843	47.500	**5.436**	33.882	45.466	15.229	14.804
		F_{DP}	10.905	11.849	11.839	14.906	11.232	11.570	17.049	15.249	12.564	14.106	**9.633**	10.467
		F_{OAE}	7.265	**2.856**	6.838	10.116	7.270	6.524	11.652	3.793	8.760	11.878	5.533	5.009
		F_{EO}	42.216	**10.300**	30.795	55.032	40.943	38.528	67.545	14.148	48.729	64.384	30.182	29.585
	Intersection	F_{MEO}	10.505	17.586	9.384	21.369	10.379	15.142	20.134	**6.119**	15.34	16.565	12.178	9.578
		F_{DP}	14.511	**8.607**	11.535	17.175	13.259	15.186	17.03	14.026	14.301	14.088	11.705	14.697
		F_{OAE}	2.536	8.461	4.928	4.870	**2.464**	3.998	3.536	6.287	2.775	3.547	4.035	3.062
		F_{EO}	24.315	25.114	27.443	47.783	21.679	30.112	43.376	20.255	28.84	33.122	26.295	**18.348**
	Individual	F_{IND}	10.338	25.742	**0.022**	1.872	2.518	7.621	0.767	3.523	0.041	3.772	0.901	0.780
Utility(%)	-	AUC↑	98.583	98.611	98.69	98.714	98.747	97.936	98.082	98.192	98.579	97.811	98.771	**99.172**
		ACC↑	96.308	94.203	94.472	95.719	**96.346**	95.092	95.151	93.651	96.224	95.426	95.722	96.174
		AP↑	99.350	99.542	99.571	99.453	99.356	99.172	99.273	99.400	99.360	99.015	99.498	**99.694**
		EER↓	5.149	6.689	6.372	5.256	**4.371**	6.483	7.708	7.633	5.145	7.063	5.499	4.961
		FPR↓	12.961	20.066	16.426	14.679	13.661	15.746	13.646	18.550	13.410	16.670	14.844	**10.971**
Training Time / Epoch			**1h15min**	2h25min	2h40min	1h18min	1h20min	3h10min	5h05min	4h	1h16min	1h25min	1h17min	7h20min

Table 4. *Overall performance comparison of difference methods on the AI-Face dataset. The best performance is shown in* ***bold****.*

and False Positive Rate (FPR).

5. Results and Analysis

In this section, we estimate the existing AI-generated image detectors' fairness performance alongside their utility on our AI-Face Dataset. More results can be found in Appendix D.

5.1. General Fairness Comparison

Overall Performance. Table 4 reports the overall performance on our AI-Face test set. Our observations are: **1)** Fairness-Enhanced Models (specifically PG-FDD [30]) are the most effective in achieving both high fairness and utility, underscoring the effectiveness of specialized fairness-enhancement techniques in mitigating demographic biases. **2)** UnivFD [88], based on the CLIP backbone [73], also achieves commendable fairness, suggesting that foundation models equipped with fairness-focused enhancements could be a promising direction for developing fairer detectors. **3)** Naive detectors, such as EfficientB4 [83], trained on large, diverse datasets (*e.g.*, our AI-Face) can achieve competitive fairness and utility, highlighting the potential of fairness improvements by choosing specific architecture. **4)** 10 out of 12 detectors have an AUC higher than 98%, demonstrating our AI-Face dataset is significant for training AI-face detectors in resulting high utility. **5)** PG-FDD demonstrates superior performance but has a long training time, which can be explored and addressed in the future.

Performance on Different Subsets. **1)** Fig. 4 demonstrates the intersectional F_{EO} and AUC performance of detectors on each test subset. We observe that the fairness performance varies a lot among different generative methods for every detector. The largest bias on most detectors comes from detecting face images generated by diffusion models. **2)** DAG-FDD [29] and SRM [87] demonstrate the most consistent fairness across subsets, indicating a robust handling of bias introduced by different generative methods. **3)** Moreover, the stable utility demonstrates our dataset's expansiveness and diversity, enabling effective training to detect AI-generated faces from various generative techniques.

Performance on Different Subgroups. We conduct an analysis of all detectors on intersectional subgroups. **1)** As shown in Fig. 5, facial images with lighter skin tone are more often misclassified as fake, likely due to the underrepresentation of lighter tones (Tone 1-3) in our dataset (see Fig. 3 (b)). This suggests detectors tend to show higher error rates for minority groups. **2)** Although gender representation is relatively balanced (see Fig. 3 (c)) in our dataset, the detectors consistently exhibit higher false positive rates for female subgroups, indicating a persistent gender-based bias.

5.2. Fairness Reliability Assessment

Fairness Robustness Evaluation. We apply 6 post-processing methods: Random Crop (RC) [97], Rotation (RT) [34], Brightness Contrast (BC) [34], Hue Saturation Value (HSV) [34], Gaussian Blur (GB) [34], and JEPG Compression (JC) [98] to the test images. Fig. 6 shows each detector's intersectional F_{EO} and AUC performance changes after using post-processing. Our observations are: **1)** These impairments tend to wash out forensic traces, so that detectors have evident performance degradation. **2)** Post-processing does not always cause detectors more bias (*e.g.*, UCF, UnivFD, CORE, DAW-FDD have better fairness after rotation), though they hurt the utility. **3)** Fairness-enhanced detectors struggle to maintain fairness when images undergo post-processing. **4)** Spatial detectors have better fairness robustness compared with other model types.

Fairness Generalization Evaluation. To evaluate detectors' fairness generalization capability, we test them on Casual

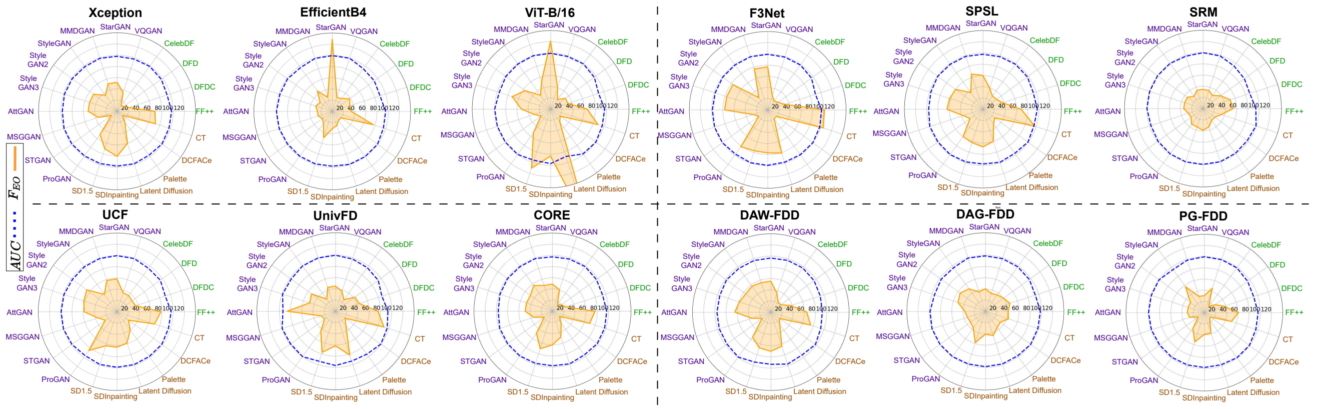


Figure 4. *Visualization of the intersectional F_{EO} (%) and AUC (%) of detectors on different subsets. The smaller F_{EO} polygon area represents better fairness. The larger AUC area means better utility.*

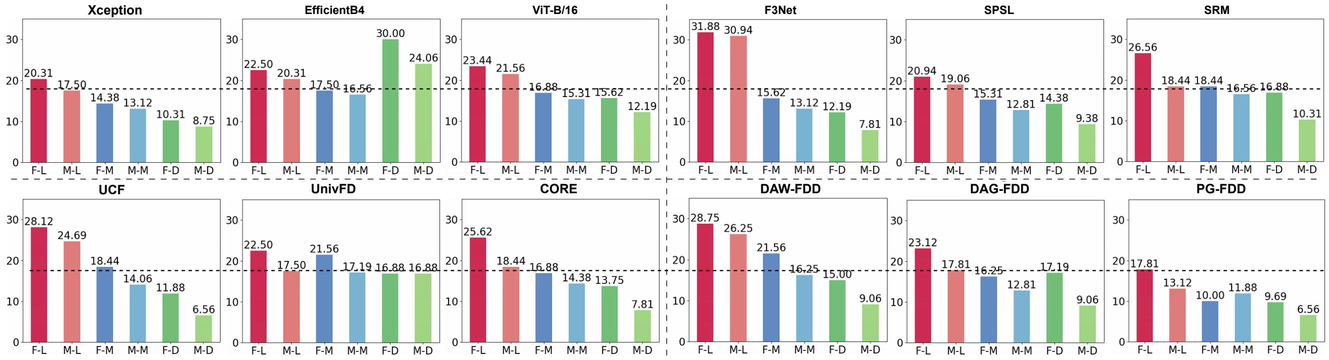


Figure 5. *FPR(%) of each intersectional subgroup The dashline represents the lowest FPR on Female-Light (F-L) subgroup.*

Conversations v2 (CCv2) [99], DF-Platter [16], and GenData [17], none of which are part of AI-Face. Notably, CCv2 is a dataset that contains only real face images with demographic annotations (*e.g.*, gender) self-reported by the participants. Results on gender attribute in Table 5 show that: **1)** Even well-designed detectors that focus on improving utility or fairness generalization (*e.g.*, UCF, PG-FDD) struggle to achieve consistently superior performance across different dataset domains. This highlights the remaining fairness generalization issue. **2)** DAW-FDD and PG-PDD are two fairness-enhanced detectors that require accessing demographic information during training, but their fairness does not encounter a drastic drop when evaluating on CCv2. This reflects the high accuracy of the annotations in our AI-face.

Effect of Training Set Size. We randomly sample 20%, 40%, 60%, and 80% of each training subset from AI-Face to assess the impact of training size on performance. Key observations from Fig. 7 (Left): **1)** Among all detectors, UnivFD demonstrates the most stable fairness and utility performance as the training dataset size changes, likely due to its fixed CLIP backbone. **2)** Increasing the training dataset size generally improves model utility, but this pattern does not extend to fairness metrics. In fact, certain detectors such as F3Net and UCF exhibit worsening fairness as the training size reaches its maximum. This suggests that more training data does not necessarily lead to fairer detectors.

Effect of the Ratio of Real and Fake. To examine how training real-to-fake sample ratios affect detector performance, we set the ratios at 1:10, 1:1, and 10:1 while keeping the total sample count constant. Experimental results in Fig. 7 (Right) show: **1)** Most detectors' fairness improves as real sample representation increases. Probably because increasing real and reducing fake samples helps detectors reduce overfitting to artifacts specific to fake samples. This makes it easier for detectors to distinguish real from fake, even for underrepresented groups, thereby enhancing fairness. **2)** Most detectors achieve the highest AUC with balanced data.

5.3. Discussion

According to the above experiments, we summarize the unsolved fairness problems in recent detectors: **1)** Detectors' fairness is unstable when detecting face images generated by different generative methods, indicating a future direction for enhancing fairness stability since new generative models continue to emerge. **2)** Even though fairness-enhanced detectors exhibit small overall fairness metrics, they still show biased detection towards minority groups. Future studies should be more cautious when designing fair detectors to ensure balanced performance across all demographic groups. **3)** There is currently no reliable detector, as all detectors experience severe large performance degradation under image post-processing and cross-domain evaluation. Future

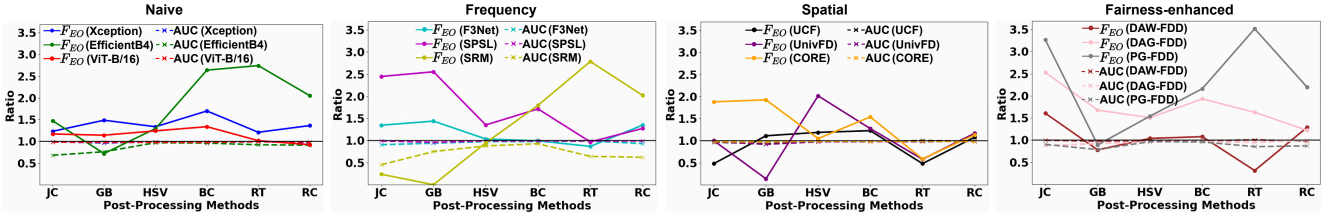


Figure 6. *Performance ratio after vs. before post-processing. Points closer to 1.0* (i.e., *no post-processing) indicate better robustness.*

Model Type	Detector	Dataset							
		CCv2 [99]		DF-Platter [16]			GenData [17]		
		Fairness(%)↓	Utility(%)↑	Fairness(%)↓		Utility(%)↑	Fairness(%)↓		Utility(%)↑
		F_{OAE}	ACC	F_{OAE}	F_{EO}	AUC	F_{OAE}	F_{EO}	AUC
Naive	Xception	1.006(+0.031)	86.465(-9.843)	6.836(+5.861)	9.789(+5.646)	81.273(-17.310)	2.539(+1.564)	13.487(+9.344)	96.971(-1.612)
	EfficientB4	4.077(+0.259)	82.980(-11.223)	8.786(+7.299)	12.370(+6.507)	67.694(-30.917)	3.304(+1.817)	**1.995** (-3.686)	93.213(-5.398)
	ViT-B/16	2.167(+0.364)	81.489(-12.983)	**0.015** (-1.788)	12.373(+6.342)	76.050(-22.640)	3.164(+1.361)	9.610(+3.579)	88.253(-10.437)
Frequency	F3Net	5.743(+4.614)	87.867(-7.852)	3.521(+2.392)	**6.445** (+1.575)	85.112(-13.602)	1.188(+0.059)	16.306(+11.436)	91.603(-7.111)
	SPSL	**0.601** (-0.436)	80.006(-16.340)	5.109(+4.072)	7.842(+3.308)	82.175(-16.572)	1.385(+0.348)	9.261(+4.272)	**98.838** (+0.091)
	SRM	7.000(+5.228)	79.768(-15.324)	3.823(+2.051)	6.567(-0.213)	66.401(-31.535)	3.281(+1.509)	7.907(+1.127)	90.049(-7.887)
Spatial	UCF	2.169(+0.718)	**93.009** (-2.142)	8.687(+7.236)	17.068(+11.558)	80.821(-17.261)	3.513(+2.062)	10.529(+5.019)	87.778(-10.304)
	UnivFD	7.625(+6.003)	67.983(-25.668)	4.540(+2.918)	9.950(+4.542)	76.443(-21.749)	1.645(+0.023)	3.848(-1.560)	94.418(-3.774)
	CORE	4.410(+3.021)	83.328(-12.896)	7.741(+6.352)	17.348(+11.417)	77.226(-21.353)	3.759(+2.370)	23.289(+17.358)	98.408(-0.171)
Fairness-enhanced	DAW-FDD	4.726(+3.401)	84.685(-10.741)	5.536(+4.211)	13.667(+7.791)	81.807(-16.004)	1.443(+0.118)	10.228(+4.532)	97.854(+0.043)
	DAG-FDD	2.364(+0.944)	83.918(-11.804)	3.064(+1.644)	22.203(+16.137)	75.206(-23.565)	**0.714** (-0.706)	10.332(+4.266)	92.108(-6.663)
	PG-FDD	1.513(+0.442)	92.852(-3.322)	4.565(+3.494)	9.717(+6.015)	**85.271** (-13.901)	3.063(+1.992)	9.479(+5.777)	93.329(-5.843)

Table 5. *Fairness generalization results based on the gender attribute. The smallest performance changes (in parentheses) and the best performance are in red and in* ***bold****, respectively. Only F_{OAE} fairness metric and ACC metric are used in CCv2 due to all samples are real.*

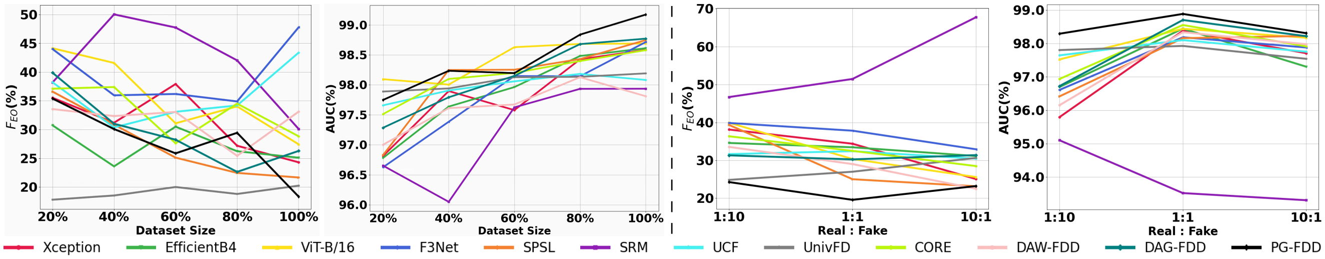


Figure 7. *Impact of the training set size (****Left****) and the ratio of real and fake (****Right****) on detectors' intersectional F_{EO}(%) and AUC (%).*

studies should aim to develop a unified framework that ensures fairness, robustness, and generalization, as these three characteristics are essential for creating a reliable detector. Moreover, integrating foundation models (*e.g.*, CLIP) into detector design may help mitigate bias.

6. Conclusion

This work presents the *first* demographically annotated million-scale AI-Face dataset, serving as a pivotal foundation for addressing the urgent need for developing fair AI face detectors. Based on this dataset, we conduct the *first* comprehensive fairness benchmark, shedding light on the fairness performance and challenges of current representative AI face detectors. Our findings can inspire and guide researchers in refining current models and exploring new methods to mitigate bias. **Limitation and Future Work:** One limitation is that our dataset's annotations are algorithmically generated, so they may lack 100% accuracy. This challenge is difficult to resolve, as demographic attributes for most AI-generated faces are often too ambiguous to predict and do not map to real-world individuals. We plan to enhance annotation quality through human labeling in the future. We also plan to extend our fairness benchmark to evaluate large language models like LLaMA2 [100] and GPT4 [101] for detecting AI faces. **Social Impact:** Malicious users could misuse AI-generated face images from our dataset to create fake social media profiles and spread misinformation. To mitigate this risk, only users who submit a signed end-user license agreement will be granted access to our dataset.

Ethics Statement

Our dataset collection and annotation generation are approved by Purdue's Institutional Review Board. The dataset is only for research purposes. All data included in this work are sourced from publicly available datasets, and we strictly comply with each dataset's license agreement to ensure lawful inclusion and permissible secondary use for training and testing. All collected data and their associated licenses are mentioned in the Datasheet of AI-Face in Appendix E. Our annotation processes prioritize ethical considerations: *1)* 76% images we annotated are generated facial images, ensuring no potential for harm to any individual. *2)* For real images, we only provide annotations for content either licensed by the original copyright holders or explicitly stated as freely shareable for research purposes.

Acknowledgments

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

References

[1] 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. 1

[2] 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. 1, 2, 3, 17, 18

[3] "Deepfakes github." https://github.com/deepfakes/faceswap. Accessed: 2024-04-17. 1, 2

[4] "Fakeapp." https://www.fakeapp.com/. Accessed: 2024-04-17. 1, 2

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *7th International Conference on Learning Representations, ICLR 2019*, 2019. 1

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019. 2, 3, 17

[7] 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.

[8] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in neural information processing systems*, vol. 34, pp. 852–863, 2021. 1, 2

[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1

[10] D. J. Tojin T. Eapen, "How generative ai can augment human creativity." https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity, 2023. Accessed: 2024-04-21. 1, 2

[11] B. News, "Trump supporters target black voters with faked ai images." https://www.bbc.com/news/world-us-canada-68440150, 2024. Accessed: 2023-05-09. 1

[12] H. S. Sætra, "Generative ai: Here to stay, but for good?," *Technology in Society*, vol. 75, p. 102372, 2023. 1

[13] M. Westerlund, "The emergence of deepfake technology: A review," *Technology innovation management review*, vol. 9, no. 11, 2019. 1, 2

[14] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898, 2020. 2

[15] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, "Deephy: On deepfake phylogeny," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, IEEE, 2022. 2

[16] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, and R. Singh, "Df-platter: multi-face heterogeneous deepfake dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9739–9748, 2023. 2, 7, 8

[17] C. Teo, M. Abdollahzadeh, and N.-M. M. Cheung, "On measuring fairness in generative models," *Advances in Neural Information Processing Systems*, vol. 36, 2023. 1, 2, 7, 8

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2

[19] Y. Xu, P. Terhöst, M. Pedersen, and K. Raja, "Analyzing fairness in deepfake detection with massively annotated databases," *IEEE Transactions on Technology and Society*, 2024. 1, 2, 13

[20] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 10–15, 2015. 2, 3, 4, 17

[21] 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, 3, 17, 18

[22] G. Research, "Contributing data to deepfake detection research," 2019. Accessed: 2024-04-12. 2, 3, 17, 18

[23] 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, 3, 17, 18

[24] 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. 1, 2

[25] 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.

[26] 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*, pp. 22412–22423, 2023. 5, 6, 18, 20, 22, 23, 24, 25, 26

[27] L. Papa, L. Faiella, L. Corvitto, L. Maiano, and I. Amerini, "On the use of stable diffusion for creating realistic faces: from generation to detection," in *2023 11th International*

Workshop on Biometrics and Forensics (IWBF), pp. 1–6, IEEE, 2023. 1, 2

[28] L. Trinh and Y. Liu, "An examination of fairness of ai models for deepfake detection," *IJCAI*, 2021. 1, 2, 3

[29] 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, 5, 6, 20, 22, 23, 24, 25, 26

[30] L. Lin, X. He, Y. Ju, X. Wang, F. Ding, and S. Hu, "Preserving fairness generalization in deepfake detection," *CVPR*, 2024. 1, 2, 5, 6, 18, 20, 22, 23, 24, 25, 26

[31] Z. Yan, T. Yao, S. Chen, Y. Zhao, X. Fu, J. Zhu, D. Luo, L. Yuan, C. Wang, S. Ding, *et al.*, "Df40: Toward next-generation deepfake detection," *NeurIPS*, 2024. 1, 2, 3

[32] C. Li *et al.*, "A continual deepfake detection benchmark: Dataset, methods, and essentials," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1349, 2023. 2, 3

[33] J. Deng, C. Lin, P. Hu, C. Shen, Q. Wang, Q. Li, and Q. Li, "Towards benchmarking and evaluating deepfake detection," *IEEE Transactions on Dependable and Secure Computing*, 2024. 3

[34] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: A comprehensive benchmark of deepfake detection," in *NeurIPS*, 2023. 3, 6

[35] B. M. Le, J. Kim, S. Tariq, K. Moore, A. Abuadbba, and S. S. Woo, "Sok: Facial deepfake detectors," *arXiv*, 2024. 2, 3

[36] 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. 2, 3

[37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021. 2, 4

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014. 2

[39] "Midjourney." https://mid-journey.ai/. Accessed: 2024-04-17. 2

[40] A. Ramesh *et al.*, "Hierarchical text-conditional image generation with clip latents," *arXiv*, vol. 1, no. 2, p. 3, 2022. 2

[41] D. O'Sullivan, "A high school student created a fake 2020 us candidate. twitter verified it." https://cnn.it/3HpHfzz, 2020. Accessed: 2024-04-21. 2

[42] S. Bond, "That smiling linkedin profile face might be a computer-generated fake." https://www.npr.org/2022/03/27/1088140809/fake-linkedin-profiles, 2022. Accessed: 2024-04-21. 2

[43] V. Albiero, K. Bowyer, K. Vangara, and M. King, "Does face recognition accuracy get better with age? deep face matchers say no," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 261–269, 2020. 3

[44] V. Albiero, K. Ks, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," in *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops*, pp. 81–89, 2020.

[45] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019. 3, 13

[46] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020. 13

[47] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The casual conversations v2 dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10–17, June 2023. 3

[48] Google, "The Monk Skin Tone Scale," 2024. [Accessed October 16, 2024]. 3, 13

[49] United States Department of State — Bureau of Consular Affairs, "Selecting your gender marker - travel," 2022. [Accessed October 16, 2024]. 3, 13

[50] Australian Bureau of Statistics, "Standard for Sex, Gender, Variations of Sex Characteristics and Sexual Orientation Variables," 2024. [Accessed October 16, 2024]. 3, 13

[51] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *2019 ieee 10th international conference on biometrics theory, applications and systems (btas)*, pp. 1–8, IEEE, 2019. 3, 13

[52] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019. 3, 13

[53] United Nations, "Provisional Guidelines on Standard International Age Classifications," 1982. [Accessed October 16, 2024]. 3, 13

[54] Statistics Canada, "Age Categories, Life Cycle Groupings," 2017. [Accessed October 16, 2024]. 3, 13

[55] O. Giudice, L. Guarnera, and S. Battiato, "Fighting deepfakes by detecting gan dct anomalies," *Journal of Imaging*, vol. 7, no. 8, p. 128, 2021. 3, 17

[56] V. Asnani, X. Yin, T. Hassner, and X. Liu, "Reverse engineering of generative models: Inferring model hyperparameters from generated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 17

[57] D. Beniaguev, "Synthetic faces high quality (sfhq) dataset," 2022. 3, 17

[58] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang, "Seeing is not always believing: Benchmarking

human and model perception of ai-generated images," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 3, 17

[59] L. M. Dang, S. I. Hassan, S. Im, J. Lee, S. Lee, and H. Moon, "Deep learning based computer generated face identification using convolutional neural network," *Applied Sciences*, vol. 8, no. 12, p. 2610, 2018. 3, 17

[60] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021. 3, 17

[61] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," *arXiv preprint arXiv:2303.09295*, 2023. 3, 17

[62] M. Kim, F. Liu, A. Jain, and X. Liu, "Dcface: Synthetic face generation with dual condition diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12715–12725, 2023. 3, 17

[63] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023. 3, 17

[64] M. Awsafur Rahman, B. Paul, N. Haque Sarker, Z. I. A. Hakim, and S. Anowarul Fattah, "Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection," *arXiv e-prints*, pp. arXiv–2302, 2023. 3, 17

[65] H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint arXiv:2309.02218*, 2023. 3, 17

[66] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020. 3, 30, 31

[67] K. S. Krishnapriya, G. Pangelinan, M. C. King, and K. W. Bowyer, "Analysis of manual and automated skin tone assignments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 429–438, January 2022. 3

[68] W. Thong, P. Joniak, and A. Xiang, "Beyond skin tone: A multidimensional measure of apparent skin color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4903–4913, October 2023. 3

[69] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "Mediapipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019. 3

[70] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979. 3

[71] Megvii Technology Limited, "Face++ Face Detection." https://www.faceplusplus.com/face-detection/. Accessed: 2024-03. 3, 13, 17, 18

[72] InsightFace Project Contributors, "InsightFace: State-of-the-Art Face Analysis Toolbox." https://insightface.ai/. Accessed: 2024-03. 3, 13, 17, 18

[73] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Open clip." https://github.com/mlfoundations/open_clip, 2021. 4, 6, 20

[74] 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

[75] S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, and M. Brundage, "Evaluating clip: towards characterization of broader capabilities and downstream implications," *arXiv preprint arXiv:2108.02818*, 2021. 4

[76] M. M. Tanjim, K. K. Singh, K. Kafle, R. Sinha, and G. W. Cottrell, "Discovering and mitigating biases in clip-based image editing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2984–2993, 2024.

[77] J. Wang and G. Kang, "Learn to rectify the bias of clip for unsupervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4102–4112, 2024. 4

[78] G. R. Kini, O. Paraskevas, S. Oymak, and C. Thrampoulidis, "Label-imbalanced and group-sensitive classification under overparameterization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18970–18983, 2021. 4

[79] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019. 4

[80] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013. 4

[81] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2020. 4

[82] 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. 5, 6, 18, 20, 22, 23, 24, 25, 26

[83] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019. 5, 6, 18, 20, 22, 23, 24, 25, 26

[84] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations*, 2021. 5, 6, 19, 20, 22, 23, 24, 25, 26

[85] 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. 5, 6, 20, 22, 23, 24, 25, 26

[86] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021. 5, 6, 20, 22, 23, 24, 25, 26

[87] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021. 5, 6, 18, 20, 22, 23, 24, 25, 26

[88] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. 5, 6, 20, 22, 23, 24, 25, 26

[89] Y. Ni, D. Meng, C. Yu, C. Quan, D. Ren, and Y. Zhao, "Core: Consistent representation learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12–21, 2022. 5, 6, 20, 22, 23, 24, 25, 26

[90] X. Han, J. Chi, Y. Chen, Q. Wang, H. Zhao, N. Zou, and X. Hu, "Ffb: A fair fairness benchmark for in-processing group fairness methods," in *ICLR*, 2024. 5

[91] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021. 5

[92] 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. 5

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

[94] 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

[95] "Monk skin tone scale," in *https://en.wikipedia.org/wiki/Monk_Skin_Tone_Scale*, Wikipedia, The Free Encyclopedia. 5

[96] S. Hu and G. H. Chen, "Fairness in survival analysis with distributionally robust optimization," *arXiv*, 2023. 5

[97] F. Cocchi, L. Baraldi, S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, "Unveiling the impact of image transformations on deepfake detection: An experimental analysis," in *International Conference on Image Analysis and Processing*, pp. 345–356, Springer, 2023. 6

[98] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of ai-generated image detection with clip," *arXiv preprint arXiv:2312.00195*, 2023. 6

[99] B. Porgali, V. Albiero, J. Ryda, C. C. Ferrer, and C. Hazirbas, "The casual conversations v2 dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10–17, 2023. 7, 8

[100] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023. 8

[101] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. 8

[102] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR, 2018. 13

[103] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 42–55, 2019. 13

[104] Z. Khan and Y. Fu, "One label, one billion faces: Usage and consistency of racial categories in computer vision," in *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, pp. 587–597, 2021. 13

[105] S. Sachdeva, "Fitzpatrick skin typing: Applications in dermatology," *Indian journal of dermatology, venereology and leprology*, vol. 75, p. 93, 2009. 13

[106] J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Reliability and validity of image-based and self-reported skin phenotype metrics," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 550–560, 2021. 13

[107] U. Okoji, S. Taylor, and J. Lipoff, "Equity in skin typing: why it is time to replace the fitzpatrick scale," *British Journal of Dermatology*, vol. 185, no. 1, pp. 198–199, 2021. 13

[108] M. Groh, C. Harris, R. Daneshjou, O. Badri, and A. Koochek, "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–26, 2022. 13

[109] R. Williamson and A. Menon, "Fairness risk measures," in *International conference on machine learning*, pp. 6786–6797, PMLR, 2019. 20

[110] 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. 20

[111] R. T. Rockafellar, S. Uryasev, *et al.*, "Optimization of conditional value-at-risk," *Journal of risk*, vol. 2, pp. 21–42, 2000. 20

[112] 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. 20

[113] 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. 20

[114] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. 30

Appendix

A. The Definition of Demographic Categories

Skin Tone: Skin tone is an important attribute of human appearance, with significant variation from pale to dark. Recently, AI systems, especially computer vision models, have become controversial over concerns about the potential bias of performance varying based on skin tone [46, 102, 103]. Additionally, existing research has pointed out that skin tone annotations can be potentially less biased than building a racial classifier [104]. And the ethnicity attribute is subjective and can conceptually cause confusion in many aspects; for example, there may be no difference in facial appearance of African-American and African people, although, they may be referred to with two distinct racial categories. We, therefore, have opted to annotate the apparent skin tone of each face image. The Monk Skin Tone Scale [48] is developed specifically for the computer vision use case. We intentionally use the Monk Skin Tone scale over the Fitzpatrick skin type [105], which is developed as means for determining one's likelihood of getting sunburn and lacks variance in darker skin tones [106, 107]. Additionally, Fitzpatrick skin type has been shown to be unreliable for image annotation [108].

Gender: Many governments [49, 50] have adopted binary gender (*i.e.*, Man/Male (M) and Woman/Female(F), defined as sex at birth, as a common choice for legal and institutional systems and official documents. Most facial recognition research [45, 51, 52] also considers binary genders in their analyses. Our AI-Face dataset adopts binary gender as gender attributes.

Age: Follow United Nations [53] and Statistics Canada [54], we have five distinct perceived age groups- Child (0-14), Youth (15-24), Adults (25-44), Middle-age Adults (45-64), and Seniors (65+).

The demographic attribute and its corresponding example images are shown from Fig. A.1 to Fig. A.3.

B. The Details of Demographically Annotated AI-Face Dataset

B.1. Detailed Information of Datasets

We build our AI-Face dataset by collecting and integrating public real and AI-generated face images sourced from academic publications, GitHub repositories, and commercial tools. We strictly adhere to the license agreements of all datasets to ensure that they allow inclusion in our datasets and secondary use for training and testing. Table B.1 shows the detailed information of each dataset we used in our AI-Face, including the number of samples, the link for downloading the dataset, the accessibility, and their licenses.

B.2. Artifacts of Deepfake Forgeries in Frequency

Leveraging frequency domain information plays a pivotal role in detecting AI-generated images. Frequency-based methods analyze the frequency components of an image, capturing information that may not be readily apparent in the spatial domain. In Fig. B.2, we present the mean Fast Fourier Transform (FFT) spectrum of images sampled from various sources in our AI-Face dataset. The results indicate that generative models often concentrate their output energy in the low-frequency range, represented by the central area of the FFT spectrum, resulting in overly smooth images. Notably, some models, such as StarGAN and Midjourney, exhibit distinct frequency artifacts, suggesting that they continue to struggle with eliminating generative patterns in the frequency domain. These artifacts serve as critical cues for distinguishing synthetic images from real ones. While most prior work has focused on applying frequency information to enhance the utility performance of detectors, exploring how frequency features can be leveraged to improve the fairness of detectors presents a promising direction for future research.

B.3. Experimental Study of Existing Face Attribute Prediction Tools

We compare current state-of-the-art face attribute prediction tools Face++ [71] and InsightFace [72] with our annotator. We perform *intra-domain* (train and test on IMDB-WIKI) and *cross-domain* (train on IMDB-WIKI, test on four AI-generated face datasets) evaluations. FF++, DFDC, DFD, and Celeb-DF-v2 are selected for cross-domain evaluation because they contain AI-generated faces, which match our objective and are not used to train Face++ and InsightFace. Additionally, they have demographic attribute annotations from [19], which can be used as ground truth for annotator evaluation. Since those annotations provided by [19] have limited age annotations, our evaluation of these four datasets is confined to gender. The intra-domain results are shown in Table B.2 and the results of cross-domain are in Table B.3. Those results demonstrate our annotator's superiority in demographic attribute prediction and generalization capability against Face++ and InsightFace. For example, under intra-domain evaluation (Table B.2), its precision surpasses the second-best method, InsightFace, by 3.47% on

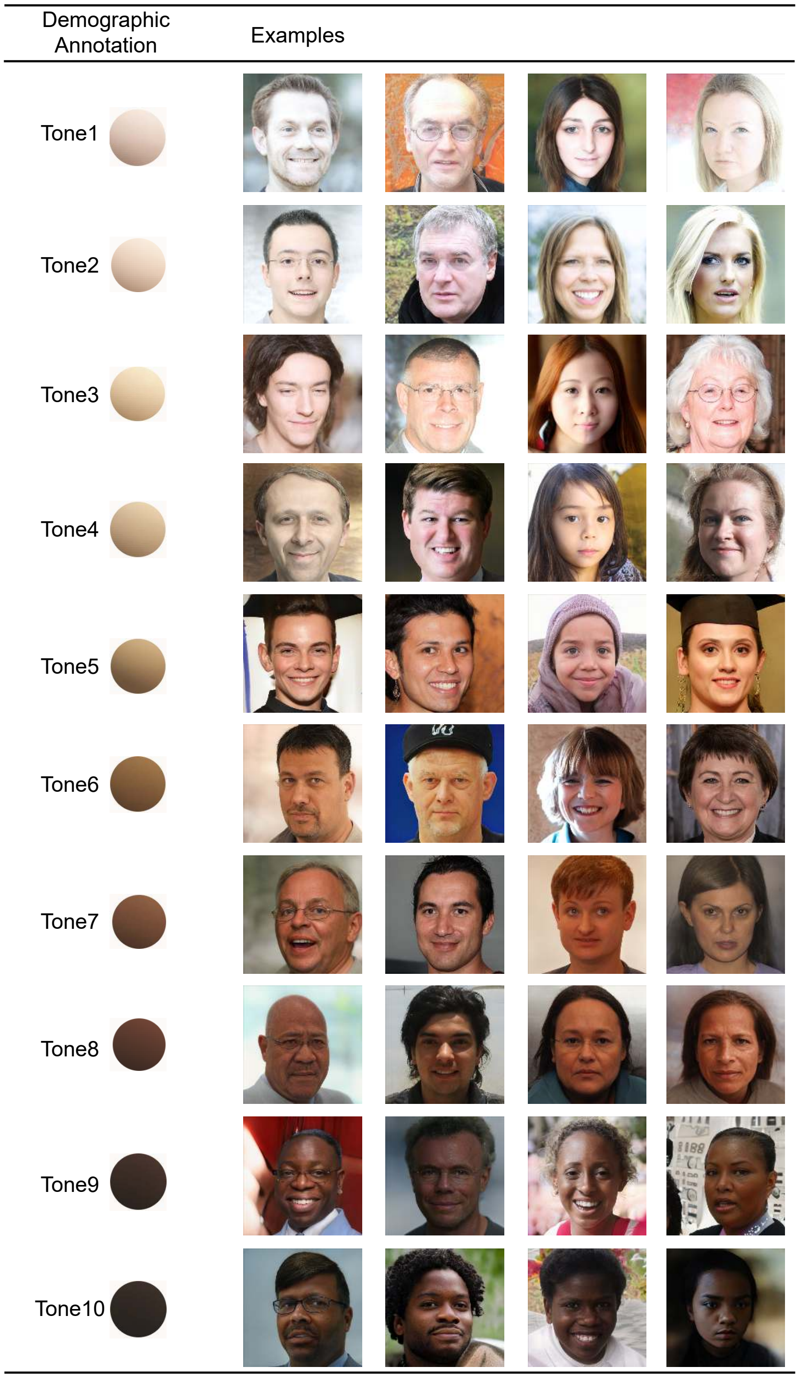


Figure A.1. *Demographic annotation definition and examples of **skin tone** attribute.*

Female and 24.81% on Senior. In cross-domain evaluation (Table B.2), our annotator maintains high accuracy on all datasets, reflecting good generalization. For instance, on the DFDC dataset, the precision our annotator outperforms Face++ by a margin of up to 1.07% and InsightFace by 3.32% on Female.

B.4. Annotator Implementation Detail

Our annotators are implemented by PyTorch and trained with a single NVIDIA RTX A6000 GPU. For training, we fix the batch size 64, epochs 32, and use Adam optimizer with an initial learning rate $\beta = 1e - 3$. Additionally, we employ a

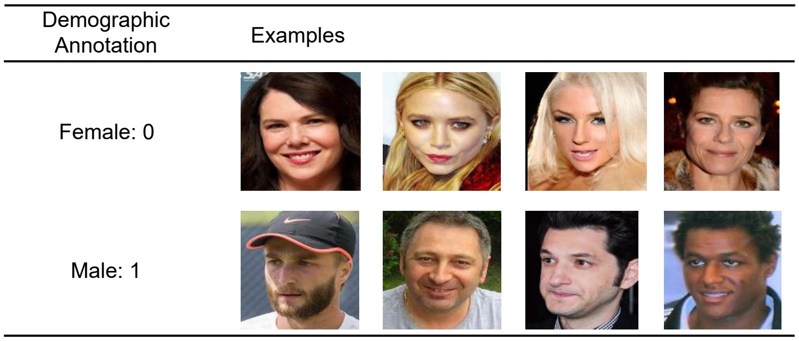


Figure A.2. *Demographic annotation definition and examples of* ***gender*** *attribute.*

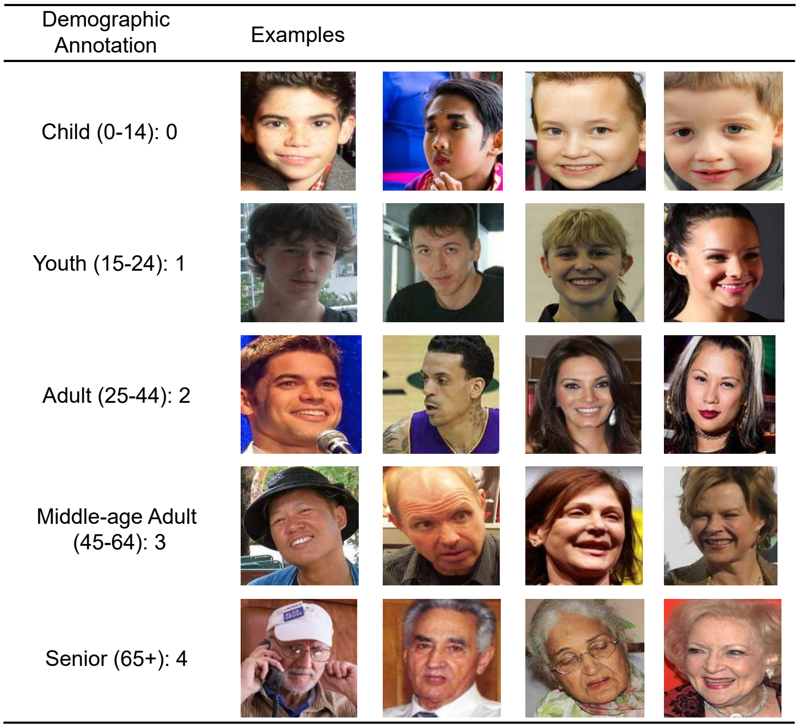


Figure A.3. *Demographic annotation definition and examples of* ***age*** *attribute.*

Cosine Annealing Learning Rate Scheduler to modulate the learning rate adaptively across the training duration. In terms of the imbalance loss, u_{A_i} is the weighting factor for attribute A_i. $h(f_i)_{A_i}$ is the predict logit on A_i. ζ_{A_i} is the multiplicative logit scaling factor, $\zeta_{A_i} = \left(\frac{N_{A_i}}{N_{\text{max}}}\right)^{\kappa}$, N_{max} is the number of samples in the most frequent class, κ is the hyperparameter controlling the sensitivity of scaling, it is set as 0.2 here. Δ_{A_i} is the additive logit scaling factor, calculated as the log of A_i probabilities $\Delta_{A_i} = \rho \cdot \log\left(\frac{N_{A_i}}{N_{\text{total}}}\right)$. The regularization hyperparameter α in fairness loss is 1e-4. The hyperparameter γ in SAM optimization is set as 0.05.

B.5. Details of Human Labeling Activities in Annotation Quality Assessment

The annotation process for assessing the quality of AI-generated face image annotations followed a structured and ethically grounded methodology. Prior to labeling, all human annotators signed an Annotator Agreement outlining the project objectives, confidentiality requirements, and detailed labeling guidelines for gender and age classification. This agreement emphasized impartiality, respect, and adherence to professional conduct throughout the annotation activities. Human annotators then underwent tutorial training using real example images to familiarize themselves with demographic attributes and labeling criteria, focusing on identifying gender-specific features, such as facial structure and presence of facial hair, and age-related

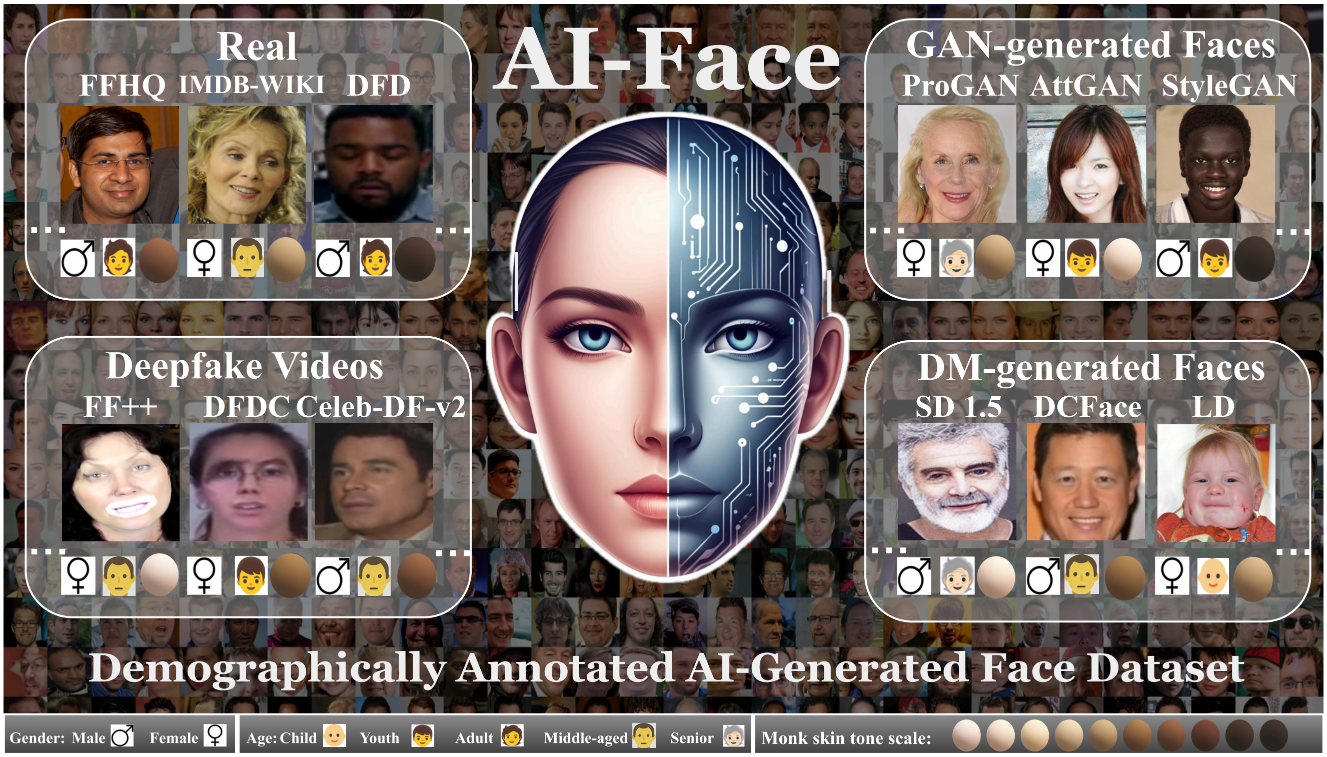


Figure B.1. *Overview of AI-Face dataset. Each face has three demographic annotations.*

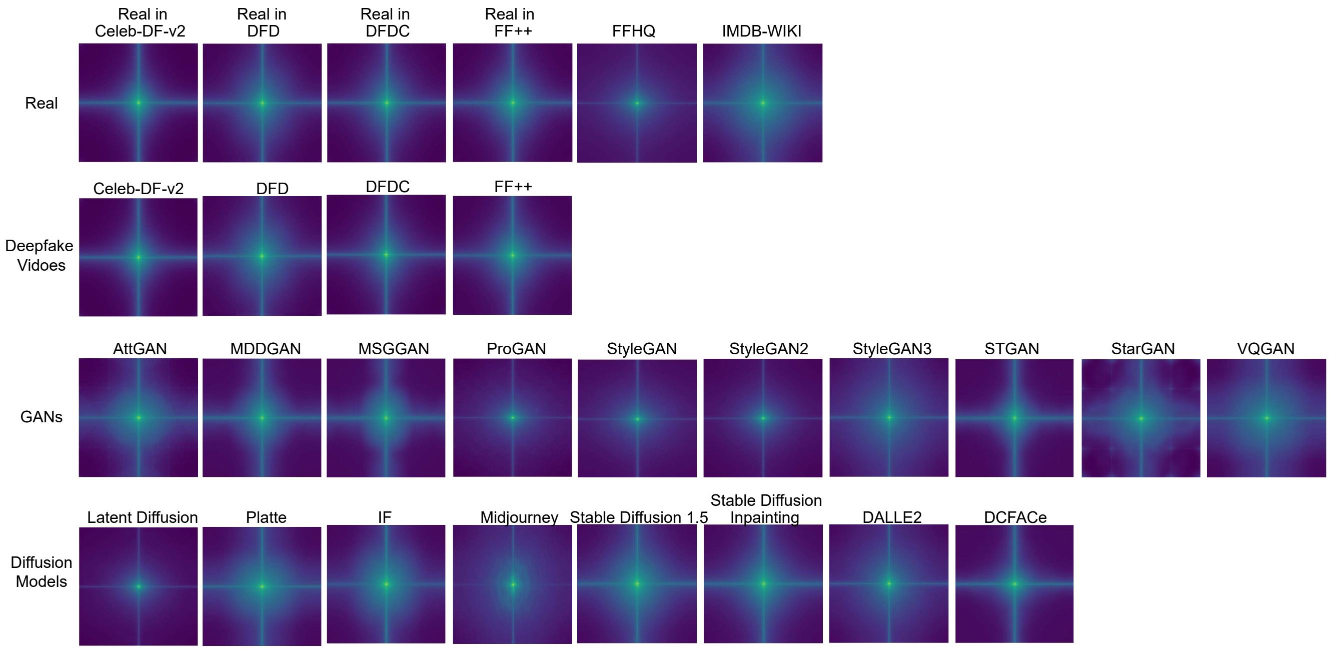


Figure B.2. *Frequency analysis on various sources. The mean FFT spectrum computation involves averaging over 2,000 images. DALLE2, IF, and Midjourney take average over 200, 500, and 100, respectively, due to their small number of samples.*

indicators, including wrinkles and skin elasticity.

Following the agreement and training, human annotators independently labeled the images based on the established criteria,

Dataset	#Samples	Link	Access	License
FF++ [2]	127K	https://github.com/ondyari/FaceForensics/tree/master/dataset	freely shared for a research purpose, submit aggreement	Non Commercial
DFDC [21]	75K	https://www.kaggle.com/c/deepfake-detection-challenge/data	the rights have been cleared for real videos, submit aggreement	Unknown
DFD [22]	40K	https://research.google/blog/contributing-data-to-deepfake-detection-research/		Non Commercial
Celeb-DF-v2 [23]	179K	https://cse.buffalo.edu/$\sim$siweilyu/celeb-deepfakeforensics.html	freely shared for a research purpose, submit aggreement	Non Commercial
AttGAN [55]	6K	https://iplab.dmi.unict.it/mfs/Deepfakes/PaperGANDCT-2021/	Online dataset, download directly, no license or agreement to sign	Unknown
StarGAN [55]	5.6K			
StyleGAN [55]	10K			
StyleGAN2 [57]	118K	https://github.com/SelfishGene/SFHQ-dataset	Since all images in this dataset are synthetically generated there are no privacy issues or license issues surrounding these images.	MIT License
StyleGAN3 [58]	26.7K	https://huggingface.co/datasets/InfImagine/FakeImageDataset	This dataset are fully open for academic research and can be used for commercial purposes with official written permission.	Apache-2.0
MMDGAN [56]	1K	https://github.com/vishal3477/Reverse_Engineering_GMs/blob/main/dataset/	The dataset can be used for research purposes only and can be used for commercial purposes with official written permission.	Non Commercial
MSGGAN [56]	1K			
STGAN [56]	1K			
ProGAN [59]	100K	https://drive.google.com/drive/folders/1jU-hzyvDZNn_M3ucuvs9xxtJNc9bPLGJ	Online dataset, download directly, no license or agreement to sign	Unknown
VQGAN [60]	50K	https://github.com/awsaf49/artifact	This dataset comes from ArtiFact dataset, which dataset takes leverage of data from multiple methods thus different parts of the dataset come with different licenses.	MIT License
DALLE2 [61]	204	https://github.com/ZhendongWang6/DIRE	freely shared for a research purpose	Unknown
IF [61]	505			
Midjourney [61]	100			
DCFACe [62]	529K	https://github.com/mk-minchul/dcface	freely shared for a research purpose	
Latent Diffusion [63]	20K	https://github.com/grip-unina/DMimageDetection	Copyright 2024 Image Processing Research Group of University Federico II of Naples ('GRIP-UNINA'). All rights reserved.Licensed under the Apache License, Version 2.0 (the "License")	Apache-2.0
Palette [64]	6K	https://github.com/awsaf49/artifact/?tab=readme-ov-file#data-generation	This dataset comes from ArtiFact dataset, which dataset takes leverage of data from multiple methods thus different parts of the dataset come with different licenses.	MIT License
SD v1.5 [65]	18K	https://huggingface.co/datasets/OpenRL/DeepFakeFace	freely shared for a research purpose	Apache-2.0
SD Inpainting [65]	20.9K			
FFHQ [6]	70K	https://github.com/NVlabs/ffhq-dataset	You can use, redistribute, and adapt it for non-commercial purposes, as long as you (a) give appropriate credit by citing our paper, (b) indicate any changes that you've made,and (c) distribute any derivative works under the same license.	Creative Commons BY-NC-SA 4.0 license
IMDB-WIKI [20]	239K	https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/	This dataset is made available for academic research purpose only. All the images are collected from the Internet, and the copyright belongs to the original owners.	Non Commericial

Table B.1. *A list of datasets used in AI-Face, including the number of samples, links, access details, and licenses.*

Method	Female			Male			Child			Young			Adult			Mid			Senior		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Face ++	0.9161 (0.0064)	0.9142 (0.0031)	0.9151 0.0045	0.9143 (0.0033)	0.9163 (0.0067)	0.9153 0.0048	0.8579 (0.0962)	0.0220 (0.0062)	0.0429 (0.0119)	0.3942 (0.0222)	0.2940 (0.0161)	0.3368 (0.0186)	0.3215 (0.0054)	0.6700 (0.0192)	0.4345 (0.0087)	0.5423 (0.0160)	0.6460 (0.0210)	0.5894 (0.0147)	0.8078 (0.0143)	0.7700 (0.0165)	0.7884 (0.0127)
Insightface	0.9648 (0.0055)	0.9405 (0.0058)	0.9525 0.0017	0.9420 (0.0050)	0.9656 (0.0057)	0.9537 0.0017	**1.0000** **(0.0000)**	0.0020 (0.0016)	0.0040 (0.0032)	0.3970 (0.0337)	0.0693 (0.0098)	0.1180 (0.0155)	0.2599 (0.0027)	0.6180 (0.0124)	0.3659 (0.0048)	0.4105 (0.0067)	0.5733 (0.0196)	0.4783 (0.0094)	0.7224 (0.0124)	0.7560 (0.0238)	0.7386 (0.0143)
Ours	**0.9995** **(0.0006)**	**0.9992** **(0.0006)**	**0.9993** **0.0006**	**0.9992** **(0.0006)**	**0.9995** **(0.0006)**	**0.9993** **0.0006**	0.9787 (0.0027)	**0.9780** **(0.0062)**	**0.9783** **(0.0039)**	**0.9560** **(0.0078)**	**0.9393** **(0.0080)**	**0.9476** **(0.0042)**	**0.9265** **(0.0122)**	**0.9547** **(0.0107)**	**0.9402** **(0.0058)**	**0.9498** **(0.0123)**	**0.9320** **(0.0123)**	**0.9408** **(0.0124)**	**0.9642** **(0.0085)**	**0.9700** **(0.0087)**	**0.9671** **(0.0082)**

Table B.2. *Detailed comparison of our annotator against Face++ [71] and InsightFace [72] on IMDB-WIKI [20] dataset. Prediction mean and standard deviation (in parentheses) of each method across 5 random samplings are reported. The best results are shown in Bold.*

categorizing gender and age into predefined classes, and recorded their annotations in a CSV file. A structured conflict resolution approach ensured accuracy and consistency in annotations. Labels agreed upon by a majority of annotators were finalized directly, while unanimous disagreements were resolved through collaborative discussions guided by the annotation guidelines. This process ensured that all annotations were objective, reliable, and aligned with ethical standards set forth in the signed agreement.

Dataset	Method	Female			Male		
		precision	recall	F1	precision	recall	F1
FF++	Face ++	**0.9816** **(0.3021)**	0.9795 (0.1360)	0.9805 (0.1459)	0.9795 (0.1312)	**0.9816** **(0.3084)**	0.9805 (0.1508)
	Insightface	0.9700 (0.4697)	0.9664 (0.6046)	0.9682 (0.3867)	0.9666 (0.5794)	0.9713 (0.4815)	0.9683 (0.3802)
	Ours	0.9799 (0.0022)	**0.9992** **(0.0006)**	**0.9894** **(0.0009)**	**0.9992** **(0.0007)**	0.9795 (0.0023)	**0.9892** **(0.0009)**
DFDC	Face ++	0.9412 (0.9771)	0.8992 (1.0095)	0.9197 (0.5639)	0.9035 (0.8353)	0.9437 (1.0246)	**0.9231** **(0.5353)**
	Insightface	0.9187 (0.9855)	0.7869 (1.7976)	0.8475 (0.7444)	0.8139 (1.1401)	0.9301 (1.0587)	0.8680 (0.3842)
	Ours	**0.9519** **(0.0106)**	**0.9741** **(0.0014)**	**0.9629** **(0.0059)**	**0.9735** **(0.0015)**	**0.9507** **(0.9619)**	0.0114 (0.0064)
DFD	Face ++	**0.9501** **(0.3773)**	0.8228 (1.8758)	0.8818 (1.0113)	0.8440 (1.3784)	**0.9568** **(0.3907)**	0.8968 (0.6856)
	Insightface	0.9441 (0.7765)	0.7557 (0.9555)	0.8394 (0.7937)	0.7964 (0.6834)	0.9552 (0.6344)	0.8686 (0.5961)
	Ours	0.9378 (0.0045)	**0.9365** **(0.0053)**	**0.9366** **(0.0033)**	**0.9366** **(0.0049)**	0.9379 (0.0048)	**0.9372** **(0.0033)**
Celeb-DF-v2	Face ++	0.9989 (0.0553)	0.9648 (0.4182)	0.9815 (0.2361)	0.9660 (0.3918)	0.9989 (0.0533)	0.9822 (0.2215)
	Insightface	0.9984 (0.0541)	0.9811 (0.3518)	0.9896 (0.1801)	0.9814 (0.3396)	0.9984 (0.0534)	0.9898 (0.1737)
	Ours	**1.0000** **(0.0000)**	**0.9997** **(0.0005)**	**0.9999** **(0.0003)**	**0.9997** **(0.0005)**	**1.0000** **(0.0000)**	**0.9999** **(0.0003)**

Table B.3. *Detailed comparison of our annotator against Face++ [71] and InsightFace [72] on FF++ [2], DFDC [21], DFD [22], and Celeb-DF-v2 [23] datasets. Prediction mean and standard deviation (in parentheses) of each method across 5 random samplings. The best results are shown in Bold.*

B.6. Visualization of Skin Tone Annotation Generation

The visualization shown in Fig. B.3 illustrates the skin tone estimation process using the Monk Skin Tone (MST) Scale. Each row represents a sample image, showing the progression from the original face with facial landmarks to the masked skin region that excludes non-skin areas like eyes and lips. Subsequently, the K-Means clustered skin region highlights the dominant skin tones extracted from the facial area. On the right, bar plots display the proportions of the top three dominant tones within the clustered region, with the top tone (largest cluster) mapped to the closest MST Scale shade. This mapping is achieved by calculating the maximum similarity, as indicated by the Euclidean distance in RGB space between the cluster centroid and MST reference colors. This process visually demonstrates how the methodology isolates, clusters, and estimates skin tones for accurate skin tone annotation generation.

C. Fairness Benchmark Settings

C.1. Implementation Detail

For fairness benchmark, all experiments are based on the PyTorch with a single NVIDIA RTX A6000 GPU. During training, we utilize SGD optimizer with a learning rate of 0.0005, with momentum of 0.9 and weight decay of 0.005. The batch size is set to 128 for most detectors. However, for the SRM [87], UCF [26], and PG-FDD [30], the batch size is adjusted to 32 due to GPU memory. For hyperparameters defined in these detectors, we use the default values set in their original papers. All detectors are initialized with their official pre-trained weights, and trained for 10 epochs.

C.2. Details of Detection Methods

We summarized the backbone architecture, GitHub repository link, and publication venue of the detectors implemented in our fairness benchmark in Table C.1. A brief introduction to each detector is provided below:

Xception [82]: is a deep convolutional neural network (CNN) architecture that relies on depthwise separable convolutions. This approach significantly reduces the number of parameters and computational cost while maintaining high performance. Xception serves as a classic backbone in deepfake detectors.

EfficientB4 [83]: is part of the EfficientNet family [83], which utilizes a novel model scaling method that uniformly scales all

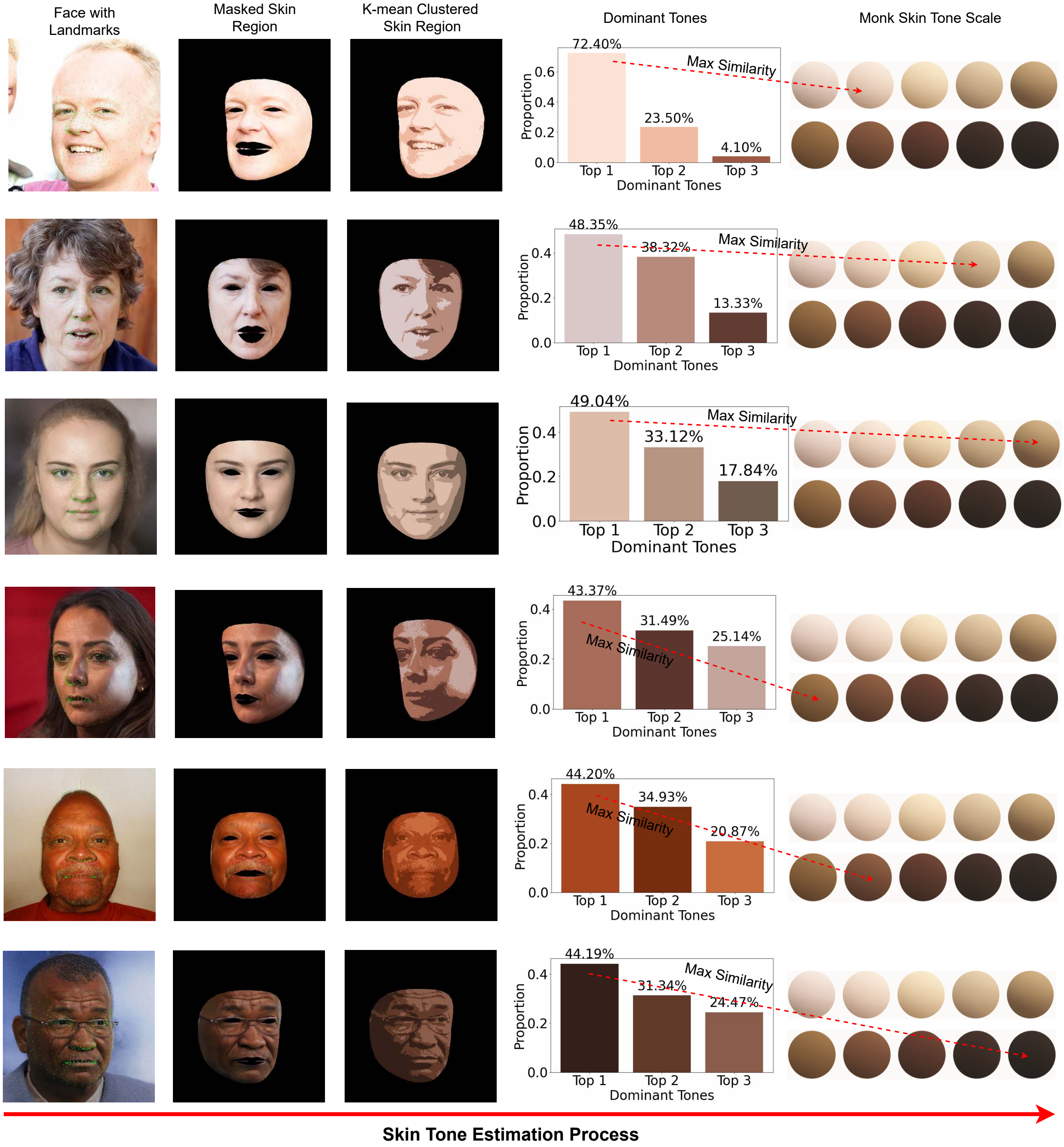


Figure B.3. *Visualization of the skin tone estimation process.*

dimensions of depth, width, and resolution using a compound coefficient. EfficientNet also serves as a classic backbone in deepfake detectors.

ViT-B/16 [84]: is a model that applies the transformer architecture, the 'B' denotes the base model size, and '16' indicates the patch size. ViT-B/16 splits images into 16 patches, linearly embeds each patch, adds positional embeddings, and feeds the resulting sequence of vectors into a standard transformer encoder.

Model Type	Detector	Backbone	GitHub Link	VENUE
Naive	Xception [82]	Xception	https://github.com/ondyari/FaceForensics/blob/master	ICCV-2019
	Efficient-B4 [83]	EfficientNet	https://github.com/lukemelas/EfficientNet-PyTorch	ICML-2019
	ViT-B/16 [84]	Transformer	https://github.com/lucidrains/vit-pytorch	ICLR-2021
Spatial	UCF [26]	Xception	https://github.com/SCLBD/DeepfakeBench/tree/main	ICCV-2023
	UnivFD [88]	CLIP VIT	https://github.com/Yuheng-Li/UniversalFakeDetect	CVPR-2023
	CORE [89]	Xception	https://github.com/niyunsheng/CORE	CVPRW-2022
Frequency	F3Net [85]	Xception	https://github.com/yyk-wew/F3Net	ECCV-2020
	SRM [87]	Xception	https://github.com/SCLBD/DeepfakeBench/tree/main	CVPR-2021
	SPSL [86]	Xception	https://github.com/SCLBD/DeepfakeBench/tree/main	CVPR-2021
Fairness-enhanced	DAW-FDD [29]	Xception	Unpublished code, reproduced by us	WACV-2024
	DAG-FDD [29]	Xception	Unpublished code, reproduced by us	WACV-2024
	PG-FDD [30]	Xception	https://github.com/Purdue-M2/Fairness-Generalization	CVPR-2024

Table C.1. *Summary of the implemented detectors in our fairness benchmark.*

F3Net [85]: utilizes a cross-attention two-stream network to effectively identify frequency-aware clues by integrating two branches: FAD and LFS. The FAD (Frequency-aware Decomposition) module divides the input image into various frequency bands using learnable partitions, representing the image with frequency-aware components to detect forgery patterns through this decomposition. Meanwhile, the LFS (Localized Frequency Statistics) module captures local frequency statistics to highlight statistical differences between authentic and counterfeit faces.

SPSL [86]: integrates spatial image data with the phase spectrum to detect up-sampling artifacts in face forgeries, enhancing the model's generalization ability for face forgery detection. The paper provides a theoretical analysis of the effectiveness of using the phase spectrum. Additionally, it highlights that local texture information is more important than high-level semantic information for accurately detecting face forgeries.

SRM [87]: extracts high-frequency noise features and combines two different representations from the RGB and frequency domains to enhance the model's generalization ability for face forgery detection.

UCF [26]: presents a multi-task disentanglement framework designed to tackle two key challenges in deepfake detection: overfitting to irrelevant features and overfitting to method-specific textures. By identifying and leveraging common features, this framework aims to improve the model's generalization ability.

UnivFD [88]: uses the frozen CLIP ViT-L/14 [73] as feature extractor and trains the last linear layer to classify fake and real images.

CORE [89]: explicitly enforces the consistency of different representations. It first captures various representations through different augmentations and then regularizes the cosine distance between these representations to enhance their consistency.

DAW-FDD [29]: a demographic-aware Fair Deepfake Detection (DAW-FDD) method leverages demographic information and employs an existing fairness risk measure [109]. At a high level, DAW-FDD aims to ensure that the losses achieved by different user-specified groups of interest (*e.g.*, different races or genders) are similar to each other (so that the AI face detector is not more accurate on one group vs another) and, moreover, that the losses across all groups are low. Specifically, DAW-FDD uses a CVaR [110, 111] loss function across groups (to address imbalance in demographic groups) and, per group, DAW-FDD uses another CVaR loss function (to address imbalance in real vs AI-generated training examples).

DAG-FDD [29]: a demographic-agnostic Fair Deepfake Detection (DAG-FDD) method, which is based on the distributionally robust optimization (DRO) [112, 113]. To use DAG-FDD, the user does not have to specify which attributes to treat as sensitive such as race and gender, only need to specify a probability threshold for a minority group without explicitly identifying all possible groups.

PG-FDD [30]: PG-FDD (Preserving Generalization Fair Deepfake Detection) employs disentanglement learning to extract demographic and domain-agnostic forgery features, promoting fair learning across a flattened loss landscape. Its framework combines disentanglement learning, fairness learning, and optimization modules. The disentanglement module introduces a loss to expose demographic and domain-agnostic features that enhance fairness generalization. The fairness learning module combines these features to promote fair learning, guided by generalization principles. The optimization module flattens the loss landscape, helping the model escape suboptimal solutions and strengthen fairness generalization.

C.3. Fairness Metrics

We assume a test set comprising indices $\{1, \dots, n\}$. Y_j and $\hat{Y}_j$ respectively represent the true and predicted labels of the sample X_j. Their values are binary, where 0 means real and 1 means fake. For all fairness metrics, a lower value means better performance. The formulations of fairness metrics are as follows,

$$
\begin{aligned}
F_{EO} &:= \sum_{\mathcal{J}_j \in \mathcal{J}} \sum_{q=0}^{1} \left| \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=1, D_j=\mathcal{J}_j, Y_j=q]}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=\mathcal{J}_j, Y_j=q]}} - \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=1, Y_j=q]}}{\sum_{j=1}^{n} \mathbb{I}_{[Y_j=q]}} \right|, \\
F_{OAE} &:= \max_{\mathcal{J}_j \in \mathcal{J}} \left\{ \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=Y_j, D_j=\mathcal{J}_j]}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=\mathcal{J}_j]}} \quad - \min_{\mathcal{J}_j{}' \in \mathcal{J}} \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=Y_j, D_j=\mathcal{J}_j{}']}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=\mathcal{J}_j{}']}} \right\}, \\
F_{DP} &:= \max_{q \in \{0,1\}} \left\{ \max_{J_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=q, D_j=J_j]}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=J_j]}} \quad - \min_{J_j' \in \mathcal{J}} \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=q, D_j=J_j']}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=J_j']}} \right\}, \\
F_{MEO} &:= \max_{q, q' \in \{0,1\}} \left\{ \max_{J_j \in \mathcal{J}} \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=q, Y_j=q', D_j=J_j]}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=J_j, Y_j=q]}} \quad - \min_{J_j' \in \mathcal{J}} \frac{\sum_{j=1}^{n} \mathbb{I}_{[\hat{Y}_j=q, Y_j=q', D_j=J_j']}}{\sum_{j=1}^{n} \mathbb{I}_{[D_j=J_j', Y_j=q]}} \right\}, \\
F_{IND} &:= \sum_{j=1}^{n-1} \sum_{l=j+1}^{n} [|f(X_j) - f(X_l)| - \delta \|X_j - X_l\|_2]_+,
\end{aligned}
$$

where D is the demographic variable, $\mathcal{J}$ is the set of subgroups with each subgroup $\mathcal{J}_j \in \mathcal{J}$. M is the set of detection models and F is the set of fairness metrics. F_{EO} measures the disparity in TPR and FPR between each subgroup and the overall population. F_{OAE} measures the maximum ACC gap across all demographic groups. F_{DP} measures the maximum difference in prediction rates across all demographic groups. And F_{MEO} captures the largest disparity in prediction outcomes (either positive or negative) when comparing different demographic groups. δ in F_{IND} is a predefined scale factor (0.08 in our experiments). $[\cdot]_+$ is the hinge function, $\| \cdot \|_2$ is the ℓ_2 norm. $f(X_j)$ represents the predicted logits of the model for input sample X_j. F_{IND} points that a model should be fair across individuals if similar individuals have similar predicted outcomes.

D. More Fairness Benchmark Results and Analysis

D.1. Detailed Results of Overall Performance Comparison

Detailed test results of each subgroup of each detector on AI-Face are presented in this section. Table D.1 provides comprehensive metrics of each subgroup on AI-Face. These results and findings align with the results reported in Table. 4 submitted main manuscript.

D.2. Performance on Different Age Subgroups

We conduct an analysis of all detectors on age subgroups. **1)** As shown in Fig. D.1, facial images with an age range of 0-14 (Child) are more often misclassified as fake, likely due to the underrepresentation of children in our dataset (see Fig. 3 (b)). This suggests detectors tend to show higher error rates for minority groups and show higher accuracy for the majority (Adult). **2)** Among those detectors, EfficientB4, UnivFD, and PG-FDD demonstrate a smaller FPR gap between age subgroups, indicating these models may be less susceptible to age bias.

D.3. Details of Post-Processing

In Section 4 we have applied 6 post-processing methods to evaluate detectors' robustness. Fig. D.2 visualizes the image after being applied different post-processing methods. We describe each post-processing method as follows:

JPEG Compression: Image compression introduces compression artifacts and reduces the image quality, simulating real-world scenarios where images may be of lower quality or have compression artifacts. In Fig. 6 we apply image compression with quality 80 to each image in the test set.

Gaussian Blur: This post-processing reduces image detail and noise by smoothing it through averaging pixel values with a Gaussian kernel. In Fig. 6 we apply gaussian blur with kernel size 7 to each image in the test set.

Hue Saturation Value: Alters the hue, saturation, and value of the image within specified limits. This post-processing technique is used to simulate variations in color and lighting conditions. Adjusting the hue changes the overall color tone, saturation controls the intensity of colors, and value adjusts the brightness. The results in Fig. 6 are after we adjust hue, saturation, and value with shifting limits 30.

Random Brightness and Contrast: This post-processing method adjusts the brightness and contrast of the image within specified limits. By applying random brightness and contrast variations, it introduces changes in the illumination and contrast

Model Type	Method	Metric	Gender		Skin Tone			Age					Intersection					
			M	F	L	M	D	Child	Young	Adult	Middle	Senior	M-L	M-M	M-D	F-L	F-M	F-D
Naive	Xception [82]	AUC	98.90	98.20	97.69	98.44	98.88	95.95	97.86	99.10	98.66	96.54	97.88	98.70	99.22	97.53	98.17	98.19
		FPR	11.12	15.10	18.37	14.72	9.54	36.83	18.26	8.95	12.78	20.42	17.70	12.82	8.42	18.93	16.58	11.32
		TPR	99.38	99.22	98.94	99.62	98.43	99.55	99.38	99.21	99.49	98.51	98.55	99.64	98.87	99.16	99.59	97.58
		ACC	96.77	95.80	95.16	96.43	96.03	89.83	94.98	97.10	97.06	92.06	94.26	96.79	96.78	95.70	96.10	94.67
	EfficientB4 [83]	AUC	98.86	98.31	99.23	98.94	97.59	99.63	98.61	98.44	98.82	98.69	99.01	99.10	98.36	99.39	98.78	95.94
		FPR	17.57	22.96	21.32	17.17	25.47	16.13	22.92	19.79	19.41	19.07	20.50	15.44	20.74	22.00	18.87	33.02
		TPR	99.04	98.56	99.49	99.02	98.19	99.56	98.89	98.55	99.14	98.66	99.22	99.25	98.60	99.64	98.81	97.40
		ACC	94.91	93.42	94.95	95.43	91.05	95.37	93.45	93.81	95.48	92.62	94.01	95.90	93.04	95.51	94.99	87.44
	ViT-B/16 [84]	AUC	99.02	98.26	97.50	98.77	98.49	95.55	98.23	98.98	98.80	97.28	97.21	99.03	99.03	97.80	98.50	97.21
		FPR	14.06	19.17	21.74	16.86	15.48	28.28	21.09	13.34	17.22	20.92	21.43	14.93	12.61	22.00	18.76	20.05
		TPR	98.43	97.51	96.75	98.23	97.36	94.58	97.88	98.14	98.40	96.16	96.77	98.56	98.21	96.74	97.93	95.72
		ACC	95.33	93.52	92.72	94.88	93.49	88.48	93.16	95.17	95.31	90.34	91.96	95.48	95.10	93.16	94.33	90.55
Frequency	F3Net [85]	AUC	99.09	98.24	98.51	98.68	98.79	96.05	98.17	99.15	98.97	97.40	98.66	99.00	99.27	98.30	98.36	97.80
		FPR	12.49	17.21	30.72	16.54	10.78	49.32	21.21	10.51	12.52	20.00	30.75	14.33	9.38	30.69	18.72	13.02
		TPR	99.15	99.00	99.69	99.41	98.11	99.77	99.40	99.00	99.06	97.96	99.55	99.34	98.76	99.76	99.48	96.86
		ACC	96.25	95.12	93.05	95.87	95.43	86.66	94.27	96.54	96.77	91.84	91.55	96.22	96.42	93.95	95.55	93.63
	SPSL [86]	AUC	98.88	98.58	98.99	98.70	98.81	97.41	98.05	99.17	98.75	97.32	99.05	98.78	99.02	98.92	98.61	98.36
		FPR	11.62	16.03	19.50	14.77	11.44	37.20	20.20	9.58	13.12	19.07	18.94	12.88	9.57	19.95	16.62	14.43
		TPR	99.64	99.52	99.53	99.73	99.15	99.78	99.58	99.53	99.69	99.12	99.44	99.77	99.37	99.58	99.70	98.73
		ACC	96.84	95.80	95.38	96.51	95.96	89.90	94.65	97.17	97.16	92.92	94.59	96.88	96.80	95.85	96.17	94.42
	SRM [87]	AUC	98.45	97.40	98.71	97.94	97.95	97.24	97.27	98.42	98.13	97.78	99.24	98.46	98.52	98.36	97.48	96.82
		FPR	12.84	19.11	22.30	17.50	12.30	36.76	22.42	11.92	15.33	19.64	18.94	14.70	9.92	25.06	20.25	16.10
		TPR	98.84	98.33	99.41	99.02	97.35	99.23	98.91	98.29	98.97	97.52	99.89	99.20	98.06	99.16	98.85	95.99
		ACC	95.93	94.16	94.67	95.35	94.44	89.62	93.59	95.65	96.14	91.67	94.91	96.03	95.76	94.53	94.72	92.03
Spatial	UCF [26]	AUC	98.62	97.45	97.20	97.92	98.49	95.59	97.26	98.74	98.67	97.04	97.67	98.44	99.00	96.88	97.41	97.47
		FPR	11.30	16.37	26.23	16.01	8.90	55.93	21.10	8.43	11.70	20.40	24.85	13.82	7.23	27.37	18.16	11.57
		TPR	98.20	97.77	98.75	98.37	96.89	99.53	98.50	97.58	98.35	96.91	99.00	98.47	97.63	98.61	98.28	95.45
		ACC	95.84	94.39	93.30	95.18	95.14	84.71	93.61	96.03	96.36	91.01	92.70	95.66	96.24	93.65	94.73	93.15
	UnivFD [88]	AUC	98.55	97.76	98.38	98.47	97.40	98.13	98.07	98.33	98.11	96.94	98.24	98.76	98.13	98.46	98.19	95.84
		FPR	16.46	20.97	20.06	19.04	17.60	19.90	19.91	16.70	22.13	17.04	17.70	16.84	15.88	22.00	21.20	20.36
		TPR	98.02	97.12	97.57	98.26	95.69	97.43	97.40	97.39	98.41	94.49	97.10	98.54	96.99	97.83	98.02	93.17
		ACC	94.42	92.80	93.73	94.43	91.68	92.80	93.09	93.74	94.35	90.56	93.19	95.03	93.29	94.04	93.87	88.74
	CORE [89]	AUC	99.04	98.01	97.80	98.47	98.79	95.88	97.82	99.09	98.77	97.11	98.12	98.91	99.26	97.57	98.02	97.84
		FPR	10.73	16.52	21.46	14.76	10.68	43.07	19.42	9.19	12.34	20.10	18.63	12.14	8.45	23.79	17.34	14.25
		TPR	99.40	99.27	99.61	99.51	98.84	99.83	99.27	99.27	99.46	99.00	99.78	99.53	99.13	99.52	99.49	98.28
		ACC	96.88	95.49	95.01	96.34	95.97	88.37	94.61	97.08	97.13	92.49	94.91	96.87	96.95	95.07	95.85	94.17
Fairness-enhanced	DAW-FDD [29]	AUC	98.36	97.15	96.84	97.45	98.51	94.97	96.41	98.62	98.00	94.95	96.83	98.05	98.85	96.77	96.87	97.91
		FPR	14.12	19.63	26.93	18.59	12.81	56.69	23.64	11.23	15.40	26.17	26.40	16.16	10.80	27.37	20.98	16.03
		TPR	99.42	99.24	99.73	99.38	99.20	99.77	99.18	99.28	99.51	98.98	99.78	99.47	99.32	99.70	99.29	98.97
		ACC	96.05	94.73	93.91	95.39	95.58	84.69	93.49	96.56	96.56	90.41	92.86	95.90	96.41	94.53	94.91	94.06
	DAG-FDD [29]	AUC	99.05	98.44	97.56	98.79	98.73	96.55	98.14	99.10	98.91	97.59	98.11	99.00	99.16	97.10	98.57	97.83
		FPR	12.11	18.02	20.76	15.10	14.21	26.95	20.01	11.72	14.61	22.40	18.32	13.04	10.58	22.76	17.14	20.00
		TPR	99.21	99.05	98.98	99.25	98.83	97.64	98.92	99.18	99.37	99.13	98.77	99.33	98.98	99.10	99.17	98.54
		ACC	96.39	94.97	94.67	96.06	94.90	91.07	94.20	96.36	96.61	91.79	94.26	96.51	96.23	94.92	95.65	92.47
	PG-FDD [30]	AUC	99.36	98.94	98.94	99.18	99.13	98.83	98.89	99.35	99.27	97.85	98.97	99.35	99.39	98.89	99.02	98.51
		FPR	9.49	12.68	15.29	12.06	8.82	22.77	14.31	7.97	11.64	19.74	13.35	10.90	7.30	16.88	13.20	11.26
		TPR	98.73	98.21	98.63	98.85	97.43	99.11	98.44	98.28	98.87	97.74	98.66	99.02	98.12	98.61	98.69	96.10
		ACC	96.68	95.61	95.59	96.43	95.55	93.27	95.26	96.66	96.79	91.78	95.49	96.75	96.56	95.65	96.12	93.69

Table D.1. *Detailed test results of each subgroup of each detector on the AI-Face. In the Skin Tone groups, 'L' represents Light (Tone 1-3), 'M' is Medium (4-6), 'D' is Dark (Tone 7-10).*

levels of the images. This evaluates detector's robustness to different illumination conditions. The results in Fig. 6 are after we adjust brightness and contrast with shifting limits 0.4.

Random Crop: Resizes the image to a specified size and then randomly crops a portion of it to the target dimensions. This post-processing method is used to evaluate the detector's robustness to variations in the spatial content of the image. The results in Fig. 6 are after we randomly crop the image with target dimension of 244 × 244.

Rotation: Rotates the image within a specified angle limit. This post-processing method is used to evaluate the detector's robustness to changes in the orientation of objects within the image. The results in Fig. 6 are after we randomly rotate the image within a range of -30 to 30 degrees.

D.4. Additional Fairness Robustness Evaluation Results

Fig. D.3 to Fig. D.7 demonstrate detectors' robustness analysis in more detail as a function of different degrees of post-processing. Overall, ViT-B/16 [84] and UnivFD [88] show stronger robustness to various post-processing methods compared to other detection methods. Fairness-enhanced detectors do not have robustness against post-processing; this would be a direction for future studies to work on. Figure D.3 presents a detailed robustness analysis in terms of utility and fairness under varying degrees of JPEG compression. The utility of all detectors decreases as image quality is reduced. Among the detectors,

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception	EfficientB4	ViT-B/16	F3Net	SPSL	SRM	UCF	UnivFD	CORE	DAW-FDD	DAG-FDD	PG-FDD
			[82]	[83]	[84]	[85]	[86]	[87]	[26]	[88]	[89]	[29]	[29]	[30]
Fairness(%)	Skin Tone	F_{MEO}	10.901	4.384	17.219	14.583	9.620	15.508	14.978	2.441	13.135	12.519	12.597	13.965
		F_{DP}	11.274	9.191	10.713	12.739	11.711	11.282	11.549	8.117	11.968	11.179	12.026	11.768
		F_{OAE}	2.434	3.609	2.276	2.232	2.814	1.780	1.950	2.940	1.658	0.878	1.753	1.439
		F_{EO}	0.160	0.093	0.205	0.209	0.156	0.191	0.186	0.034	0.176	0.165	0.176	0.176
	Gender	F_{MEO}	5.475	5.458	8.003	5.749	5.754	5.848	5.575	3.244	4.367	5.186	5.808	4.086
		F_{DP}	1.205	1.412	2.416	1.340	1.445	1.959	1.810	0.781	0.980	1.715	1.470	1.545
		F_{OAE}	2.043	1.800	1.896	2.054	1.969	1.471	1.569	1.413	1.848	1.430	2.012	1.133
		F_{EO}	0.066	0.063	0.083	0.068	0.067	0.062	0.060	0.043	0.056	0.056	0.068	0.044
	Age	F_{MEO}	28.244	7.460	38.521	27.860	24.768	40.542	44.342	8.584	34.156	36.450	35.031	36.197
		F_{DP}	11.228	12.245	12.140	11.395	11.466	14.564	15.856	16.134	13.525	12.256	13.478	12.082
		F_{OAE}	7.138	5.234	10.940	6.933	6.053	11.126	11.481	4.171	8.294	9.192	8.636	8.934
		F_{EO}	0.460	0.175	0.560	0.460	0.410	0.550	0.610	0.191	0.508	0.537	0.524	0.539
	Intersection	F_{MEO}	15.752	10.644	24.460	18.455	15.157	18.381	17.397	5.300	16.257	14.806	15.219	16.517
		F_{DP}	16.943	14.565	13.773	18.071	17.490	14.943	15.612	12.967	17.063	14.802	16.786	15.513
		F_{OAE}	6.805	8.029	5.025	6.658	7.200	3.614	3.532	6.226	5.079	3.314	5.757	2.989
		F_{EO}	0.355	0.307	0.441	0.440	0.366	0.382	0.382	0.178	0.371	0.336	0.399	0.354
Utility(%)	-	AUC	0.968	0.968	0.981	0.966	0.968	0.967	0.977	0.979	0.975	0.970	0.973	0.978
		ACC	0.931	0.922	0.924	0.930	0.929	0.946	0.951	0.933	0.947	0.941	0.941	0.952
		AP	0.987	0.988	0.994	0.986	0.988	0.985	0.991	0.993	0.989	0.987	0.989	0.991
		EER	0.093	0.101	0.082	0.096	0.095	0.076	0.074	0.083	0.076	0.085	0.084	0.072
		FPR	0.205	0.219	0.290	0.199	0.190	0.205	0.163	0.188	0.144	0.186	0.168	0.151

Table D.2. *Detailed fairness and utility evaluation results on 20% training subset.*

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception	EfficientB4	ViT-B/16	F3Net	SPSL	SRM	UCF	UnivFD	CORE	DAW-FDD	DAG-FDD	PG-FDD
			[82]	[83]	[84]	[85]	[86]	[87]	[26]	[88]	[89]	[29]	[29]	[30]
Fairness(%)	Skin Tone	F_{MEO}	9.815	4.414	12.194	10.801	9.275	24.037	12.299	2.568	15.734	12.628	10.463	10.982
		F_{DP}	10.080	10.413	9.475	11.137	10.644	12.550	10.632	8.322	11.251	10.157	10.456	10.928
		F_{OAE}	0.122	0.095	0.146	0.154	0.122	0.281	0.150	0.043	0.180	0.154	0.135	0.143
		F_{EO}	1.472	3.796	3.395	2.088	1.631	3.280	1.533	2.898	2.384	1.571	1.323	1.188
	Gender	F_{DP}	5.576	3.592	6.089	5.985	4.368	7.959	4.400	3.438	5.234	5.960	4.797	5.390
		F_{MEO}	1.817	0.822	1.853	1.566	1.227	2.658	1.481	0.797	1.866	2.053	1.458	2.052
		F_{OAE}	1.559	1.622	1.829	1.966	1.595	1.722	1.338	1.507	1.369	1.473	1.530	1.303
		F_{EO}	0.060	0.047	0.067	0.069	0.052	0.080	0.049	0.045	0.055	0.062	0.055	0.055
	Age	F_{MEO}	32.781	9.931	18.050	26.665	33.004	54.967	43.829	7.840	38.202	41.707	34.285	33.582
		F_{DP}	12.161	12.428	14.762	10.954	13.272	16.006	14.243	16.076	13.394	13.661	11.630	12.955
		F_{OAE}	8.535	4.312	6.949	6.831	8.210	14.795	11.248	4.185	9.906	10.841	8.738	8.520
		F_{EO}	0.474	0.235	0.348	0.432	0.470	0.738	0.611	0.179	0.541	0.585	0.509	0.500
	Intersection	F_{MEO}	13.451	8.558	15.615	13.559	12.356	30.133	15.278	5.585	19.342	17.006	12.784	13.197
		F_{DP}	13.795	16.478	14.409	16.096	15.138	15.893	14.797	13.263	15.123	14.586	14.345	15.166
		F_{OAE}	4.424	7.462	5.221	6.173	4.728	5.030	2.886	6.298	3.741	3.541	4.402	2.911
		F_{EO}	0.312	0.236	0.416	0.360	0.309	0.550	0.304	0.185	0.374	0.324	0.310	0.301
Utility(%)	-	AUC	0.979	0.976	0.980	0.974	0.982	0.957	0.979	0.979	0.981	0.976	0.978	0.982
		ACC	0.951	0.940	0.933	0.937	0.951	0.938	0.955	0.934	0.957	0.948	0.949	0.960
		AP	0.991	0.991	0.993	0.989	0.993	0.983	0.992	0.993	0.991	0.989	0.991	0.992
		EER	0.068	0.078	0.081	0.083	0.066	0.117	0.072	0.082	0.058	0.074	0.074	0.055
		FPR	0.165	0.147	0.136	0.186	0.147	0.245	0.157	0.180	0.151	0.184	0.169	0.130

Table D.3. *Detailed fairness and utility evaluation results on 40% training subset.*

ViT-B/16 [84] exhibits the highest utility robustness, ViT-B/16 [84] and UnivFD [88] both demonstrate the strongest fairness robustness. When considering Gaussian blur, ViT-B/16 again stands out as the most robust detector in terms of utility, whereas DAW-FDD [29] and UnivFD [88] show the great robustness in terms of fairness. Against Hue Saturation Value adjustments, SPSL [86] shows the strongest utility robustness, while the fairness of DAW-FDD [29] fluctuates less with different Hue Saturation Value adjustments. ViT-B/16 demonstrates superior robustness in both utility and fairness when facing rotations. For brightness contrast variations, SPSL [86] is the most robust detector in terms of utility, while UnivFD once again shows superior robustness in terms of fairness. Last, we can get the same conclusion from Fig. D.3 to Fig. D.7 as in the main manuscript, that post-processing clearly impairs detectors' utility but does not necessarily make detectors more biased.

D.5. Full Results of Effect of Increasing the Size of Train Set

In this section, we provide the full evaluation results tested under different sizes of train set, as shown from Table D.2 to Table D.5. Intersection F_{EO} and AUC align with the results in Fig. 7 of the submitted manuscript.

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception [82]	EfficientB4 [83]	ViT-B/16 [84]	F3Net [85]	SPSL [86]	SRM [87]	UCF [26]	UnivFD [88]	CORE [89]	DAW-FDD [29]	DAG-FDD [29]	PG-FDD [30]
Fairness(%)	Skin Tone	F_{MEO}	9.086	14.704	4.388	15.303	6.813	14.516	14.952	2.186	9.689	13.488	9.672	4.108
		F_{DP}	10.232	11.784	7.714	11.225	9.979	14.909	13.116	8.004	10.666	10.844	10.054	8.575
		F_{OAE}	1.531	2.017	2.572	2.320	1.247	1.733	1.562	2.777	1.208	1.664	1.259	1.383
		F_{EO}	0.124	0.194	0.084	0.177	0.100	0.234	0.208	0.043	0.131	0.163	0.125	0.055
	Gender	F_{MEO}	4.418	6.743	8.445	5.545	5.242	7.331	5.713	4.182	4.395	5.579	4.978	2.430
		F_{DP}	1.243	2.063	2.697	2.096	1.736	3.656	1.845	1.142	1.622	2.153	1.600	1.061
		F_{OAE}	1.567	1.846	2.052	1.318	1.499	0.630	1.651	1.489	1.217	1.267	1.470	0.858
		F_{EO}	0.053	0.072	0.086	0.057	0.057	0.087	0.062	0.050	0.048	0.056	0.055	0.029
	Age	F_{MEO}	35.231	27.998	17.573	42.366	35.428	37.043	34.243	5.520	28.666	40.326	38.409	24.355
		F_{DP}	12.874	12.663	11.691	15.070	12.924	16.570	13.579	15.256	11.264	13.929	13.112	10.449
		F_{OAE}	8.954	7.379	7.004	10.921	8.915	8.078	8.232	3.900	7.519	10.379	9.909	6.870
		F_{EO}	0.514	0.411	0.320	0.585	0.519	0.528	0.520	0.134	0.438	0.570	0.558	0.362
	Intersection	F_{MEO}	11.554	18.923	10.063	18.907	10.175	20.404	18.818	5.414	12.995	15.944	12.552	5.425
		F_{DP}	14.625	15.884	9.908	15.240	14.093	18.967	18.087	12.890	15.331	14.584	13.949	13.379
		F_{OAE}	4.755	4.997	5.459	3.617	4.210	2.965	4.832	6.152	3.747	3.106	4.102	3.306
		F_{EO}	0.279	0.405	0.311	0.362	0.251	0.477	0.431	0.200	0.276	0.331	0.282	0.159
Utility(%)	-	AUC	0.976	0.980	0.986	0.981	0.983	0.976	0.981	0.981	0.982	0.977	0.982	0.982
		ACC	0.948	0.945	0.943	0.961	0.952	0.927	0.951	0.935	0.956	0.960	0.950	0.960
		AP	0.989	0.992	0.996	0.991	0.993	0.991	0.992	0.994	0.992	0.987	0.993	0.991
		EER	0.074	0.071	0.065	0.058	0.067	0.092	0.068	0.078	0.062	0.060	0.069	0.058
		FPR	0.166	0.183	0.177	0.137	0.154	0.137	0.137	0.198	0.145	0.143	0.172	0.127

Table D.4. *Detailed fairness and utility evaluation results on 60% training subset.*

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception [82]	EfficientB4 [83]	ViT-B/16 [84]	F3Net [85]	SPSL [86]	SRM [87]	UCF [26]	UnivFD [88]	CORE [89]	DAW-FDD [29]	DAG-FDD [29]	PG-FDD [30]
Fairness(%)	Skin Tone	F_{MEO}	15.463	6.826	7.442	13.642	4.221	9.425	13.574	2.368	13.487	10.127	8.763	7.613
		F_{DP}	11.994	10.440	8.378	11.211	8.998	6.875	11.085	8.231	11.506	10.182	8.920	11.914
		F_{OAE}	1.998	2.390	1.540	1.713	2.141	7.511	1.661	2.777	1.538	1.375	1.484	1.822
		F_{EO}	0.192	0.107	0.095	0.171	0.064	0.187	0.168	0.036	0.170	0.119	0.114	0.129
	Gender	F_{MEO}	4.209	3.639	9.461	5.189	4.116	2.328	5.402	4.084	5.058	4.112	4.143	4.035
		F_{DP}	1.171	1.043	3.025	1.749	1.191	2.803	1.778	1.082	1.899	1.537	1.560	1.401
		F_{OAE}	1.579	1.395	2.163	1.499	1.507	1.248	1.560	1.519	1.277	1.218	1.158	1.353
		F_{EO}	0.051	0.045	0.095	0.057	0.050	0.037	0.059	0.050	0.053	0.045	0.045	0.046
	Age	F_{MEO}	33.930	16.272	10.222	45.076	20.048	11.857	45.508	5.788	37.055	30.409	29.707	20.058
		F_{DP}	15.167	11.254	11.643	15.938	11.379	8.360	15.357	15.620	13.565	11.925	9.880	10.487
		F_{OAE}	8.520	4.738	5.662	11.409	5.123	10.447	11.546	3.777	9.454	7.784	8.513	4.450
		F_{EO}	0.488	0.286	0.231	0.623	0.322	0.228	0.625	0.136	0.526	0.441	0.459	0.357
	Intersection	F_{MEO}	19.488	11.396	16.807	15.829	6.631	12.599	16.926	5.116	16.594	13.597	11.226	10.523
		F_{DP}	17.093	15.841	11.946	15.708	13.376	11.010	15.607	13.121	15.834	14.758	12.474	16.906
		F_{OAE}	3.583	5.700	5.469	3.139	5.452	10.020	2.878	6.209	2.852	2.313	2.755	4.348
		F_{EO}	0.392	0.262	0.340	0.349	0.225	0.420	0.342	0.188	0.345	0.254	0.226	0.294
Utility(%)	-	AUC	0.985	0.985	0.987	0.981	0.984	0.979	0.982	0.981	0.984	0.981	0.987	0.988
		ACC	0.950	0.949	0.940	0.958	0.950	0.816	0.956	0.936	0.959	0.963	0.960	0.953
		AP	0.994	0.994	0.996	0.992	0.994	0.992	0.993	0.994	0.993	0.989	0.995	0.996
		EER	0.065	0.064	0.062	0.069	0.064	0.066	0.068	0.078	0.060	0.053	0.057	0.058
		FPR	0.145	0.162	0.206	0.134	0.148	0.039	0.141	0.189	0.138	0.118	0.144	0.099

Table D.5. *Detailed fairness and utility evaluation results on 80% training subset.*

D.6. Full Results of Effect of the Ratio of Real and Fake

In this section, we provide the full evaluation results tested under the train set with different ratios of real and fake, as shown from Table D.6 to Table D.8. Intersection F_{EO} and AUC align with the results in Fig. 7 of the submitted manuscript.

D.7. Comparison Results with Foundation Model

In the Discussion (Section 5.3) of the main manuscript, we highlighted the potential of integrating foundation models (*e.g.*, CLIP) into detector design as a strategy for mitigating bias. To explore this, we conducted a preliminary experiment by designing a detector using a frozen CLIP model combined with a trainable 3-layer MLP. This model was trained and tested on the AI-Face dataset. For comparison, we selected one representative detector from each model type: EfficientB4 [83], SPSL [86], UnivFD [88], and PG-FDD [30]. These four detectors' results are consistent with those reported in Table 4. As shown in Table D.9, the CLIP+MLP detector demonstrates a clear advantage in both fairness and utility metrics, suggesting that foundation models hold significant promise for bias mitigation. For instance, its F_{EO} score is 3.11% lower than the

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception	EfficientB4	ViT-B/16	F3Net	SPSL	SRM	UCF	UnivFD	CORE	DAW-FDD	DAG-FDD	PG-FDD
			[82]	[83]	[84]	[85]	[86]	[87]	[26]	[88]	[89]	[29]	[29]	[30]
Fairness(%)	Skin Tone	F_{MEO}	9.750	11.908	5.095	12.896	5.859	18.083	10.363	5.926	10.955	9.570	8.980	3.889
		F_{DP}	16.824	14.637	12.547	17.621	14.479	19.676	15.803	8.744	16.109	16.063	15.700	12.899
		F_{OAE}	2.401	2.973	1.714	3.630	0.905	5.752	3.249	5.302	3.524	2.641	2.684	1.164
		F_{EO}	0.164	0.155	0.119	0.183	0.113	0.248	0.147	0.083	0.149	0.142	0.132	0.069
	Gender	F_{MEO}	3.535	0.952	4.275	4.788	4.439	4.201	4.655	6.898	3.981	3.848	3.674	3.792
		F_{DP}	2.241	0.715	3.129	3.184	3.060	2.245	3.426	4.328	2.850	2.562	2.514	2.790
		F_{OAE}	2.482	0.833	2.263	2.690	2.496	2.920	2.332	3.028	2.236	2.451	2.360	2.166
		F_{EO}	0.051	0.011	0.047	0.057	0.052	0.062	0.049	0.070	0.048	0.051	0.049	0.045
	Age	F_{MEO}	23.282	14.195	13.627	32.679	26.796	52.939	30.332	8.251	32.612	23.934	23.834	15.408
		F_{DP}	17.521	18.304	16.426	20.437	18.489	28.303	18.892	12.343	19.913	17.103	18.221	15.255
		F_{OAE}	11.063	3.474	9.185	16.411	13.102	26.805	15.234	6.602	16.794	11.909	11.523	7.893
		F_{EO}	0.381	0.203	0.303	0.490	0.436	0.769	0.472	0.162	0.469	0.378	0.370	0.273
	Intersection	F_{MEO}	11.921	17.564	7.727	17.622	10.201	21.679	15.031	12.980	12.775	13.010	11.104	7.173
		F_{DP}	23.484	21.311	17.501	23.519	21.115	24.119	22.167	13.434	21.787	22.542	22.120	18.247
		F_{OAE}	4.000	4.650	3.969	5.957	4.306	9.974	5.003	11.530	5.750	3.903	4.600	4.426
		F_{EO}	0.344	0.335	0.304	0.378	0.250	0.515	0.324	0.270	0.325	0.290	0.303	0.196
Utility(%)	-	AUC	0.984	0.984	0.984	0.982	0.982	0.935	0.981	0.979	0.986	0.983	0.987	0.989
		ACC	0.937	0.892	0.922	0.928	0.929	0.900	0.930	0.830	0.933	0.937	0.944	0.940
		AP	0.980	0.986	0.986	0.977	0.978	0.910	0.977	0.980	0.982	0.980	0.984	0.987
		EER	0.062	0.064	0.067	0.061	0.066	0.129	0.065	0.082	0.056	0.061	0.053	0.052
		FPR	0.087	0.005	0.111	0.122	0.111	0.164	0.116	0.337	0.115	0.095	0.083	0.095

Table D.6. *Detailed fairness and utility evaluation results on a training subset with the ratio of real vs fake is 1:1.*

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception	EfficientB4	ViT-B/16	F3Net	SPSL	SRM	UCF	UnivFD	CORE	DAW-FDD	DAG-FDD	PG-FDD
			[82]	[83]	[84]	[85]	[86]	[87]	[26]	[88]	[89]	[29]	[29]	[30]
Fairness(%)	Skin Tone	F_{MEO}	11.678	10.565	8.595	12.629	11.790	17.068	9.661	5.615	11.138	11.726	8.680	6.435
		F_{DP}	14.133	12.859	10.579	15.157	13.983	16.962	14.081	8.438	14.161	14.724	13.388	13.256
		F_{OAE}	4.539	3.671	4.407	4.104	4.894	5.036	3.931	5.461	4.389	3.379	3.006	2.232
		F_{EO}	0.128	0.151	0.102	0.148	0.127	0.200	0.107	0.081	0.121	0.141	0.103	0.081
	Gender	F_{MEO}	6.942	2.295	10.586	7.818	7.327	8.990	6.054	6.525	7.518	6.259	5.934	4.944
		F_{DP}	4.378	0.094	6.203	4.881	4.632	5.572	4.086	4.136	4.867	4.093	3.917	3.565
		F_{OAE}	3.225	1.508	4.799	3.669	3.303	4.060	2.906	2.842	3.445	2.890	2.795	2.407
		F_{EO}	0.072	0.023	0.106	0.080	0.074	0.091	0.063	0.067	0.075	0.065	0.062	0.052
	Age	F_{MEO}	36.384	13.175	25.574	35.508	30.942	36.806	32.134	6.629	34.717	32.860	29.474	28.923
		F_{DP}	18.815	19.006	16.522	18.393	17.007	18.947	17.086	12.524	18.454	16.798	15.004	19.634
		F_{OAE}	19.144	2.128	14.899	18.668	16.331	19.294	16.714	6.598	18.249	17.426	15.605	15.099
		F_{EO}	0.524	0.193	0.373	0.525	0.442	0.553	0.490	0.147	0.507	0.482	0.447	0.420
	Intersection	F_{MEO}	16.037	18.196	19.081	16.394	19.895	21.535	13.589	12.135	15.921	14.772	12.424	10.340
		F_{DP}	16.749	17.705	16.207	17.525	18.313	19.312	17.029	12.639	17.144	16.813	15.850	16.846
		F_{OAE}	7.914	4.565	12.301	6.936	8.704	9.372	5.877	11.469	7.025	5.523	6.909	5.109
		F_{EO}	0.381	0.346	0.400	0.399	0.394	0.467	0.317	0.249	0.364	0.336	0.313	0.243
Utility(%)	-	AUC	0.958	0.967	0.975	0.966	0.964	0.951	0.976	0.978	0.969	0.962	0.967	0.983
		ACC	0.864	0.864	0.823	0.876	0.855	0.862	0.909	0.829	0.888	0.874	0.882	0.925
		AP	0.934	0.972	0.976	0.948	0.946	0.938	0.964	0.979	0.952	0.942	0.951	0.977
		EER	0.087	0.098	0.091	0.082	0.081	0.116	0.064	0.085	0.073	0.084	0.079	0.056
		FPR	0.267	0.012	0.349	0.242	0.285	0.272	0.172	0.339	0.221	0.246	0.227	0.142

Table D.7. *Detailed fairness and utility evaluation results on a training subset with the ratio of real vs fake is 1:10.*

second-best method, PG-FDD, for the Skin Tone group, and 14.046% lower for the Intersection group.

Measure	Attribute	Metric	Model Type											
			Native			Frequency			Spatial			Fairness-enhanced		
			Xception [82]	EfficientB4 [83]	ViT-B/16 [84]	F3Net [85]	SPSL [86]	SRM [87]	UCF [26]	UnivFD [88]	CORE [89]	DAW-FDD [29]	DAG-FDD [29]	PG-FDD [30]
Fairness(%)	Skin Tone	F_{MEO}	6.557	11.357	7.743	7.364	5.850	18.290	5.814	12.493	6.397	5.083	6.515	5.122
		F_{DP}	13.087	13.112	12.348	15.751	11.728	22.766	14.881	13.771	14.364	12.265	15.536	12.718
		F_{OAE}	1.779	2.939	1.589	1.352	2.402	3.823	1.495	2.386	0.996	1.401	1.471	1.742
		F_{EO}	0.108	0.140	0.093	0.153	0.090	0.331	0.135	0.155	0.125	0.084	0.143	0.096
	Gender	F_{MEO}	1.808	1.803	2.106	1.659	2.316	3.035	3.035	1.997	3.499	1.920	1.812	3.016
		F_{DP}	1.769	0.136	0.369	1.321	0.470	2.871	2.503	2.008	2.119	1.097	1.405	2.633
		F_{OAE}	1.385	1.269	1.615	1.661	1.777	1.270	1.817	0.579	2.662	1.928	1.734	1.677
		F_{EO}	0.025	0.018	0.027	0.032	0.032	0.038	0.035	0.020	0.052	0.036	0.033	0.033
	Age	F_{MEO}	8.571	11.954	9.832	8.680	9.523	36.509	11.812	10.258	13.180	8.539	9.553	11.109
		F_{DP}	17.740	18.338	18.662	17.788	18.910	27.851	16.998	16.446	16.719	18.069	18.034	16.696
		F_{OAE}	3.191	0.656	1.405	4.291	2.010	12.139	5.174	2.737	5.723	3.924	4.283	4.859
		F_{EO}	0.196	0.157	0.141	0.228	0.174	0.685	0.281	0.173	0.288	0.209	0.240	0.246
	Intersection	F_{MEO}	11.785	18.146	14.840	14.116	12.940	23.975	11.313	16.240	12.729	10.948	13.399	9.441
		F_{DP}	20.192	17.674	19.309	23.135	18.594	29.625	22.394	19.583	21.935	18.674	23.584	17.906
		F_{OAE}	3.419	4.233	3.638	4.263	4.020	5.166	3.479	3.133	4.945	3.388	3.885	3.667
		F_{EO}	0.251	0.312	0.256	0.329	0.232	0.678	0.305	0.307	0.285	0.225	0.314	0.232
Utility(%)	-	AUC	0.978	0.973	0.982	0.979	0.982	0.933	0.978	0.975	0.979	0.980	0.982	0.983
		ACC	0.920	0.862	0.895	0.928	0.916	0.832	0.921	0.849	0.921	0.920	0.930	0.933
		AP	0.978	0.977	0.984	0.978	0.984	0.915	0.979	0.978	0.978	0.979	0.981	0.984
		EER	0.070	0.088	0.075	0.066	0.064	0.141	0.076	0.087	0.074	0.070	0.065	0.066
		FPR	0.034	0.008	0.009	0.042	0.018	0.116	0.054	0.004	0.055	0.037	0.040	0.051

Table D.8. *Detailed fairness and utility evaluation results on a training subset with the ratio of real vs fake is 10:1.*

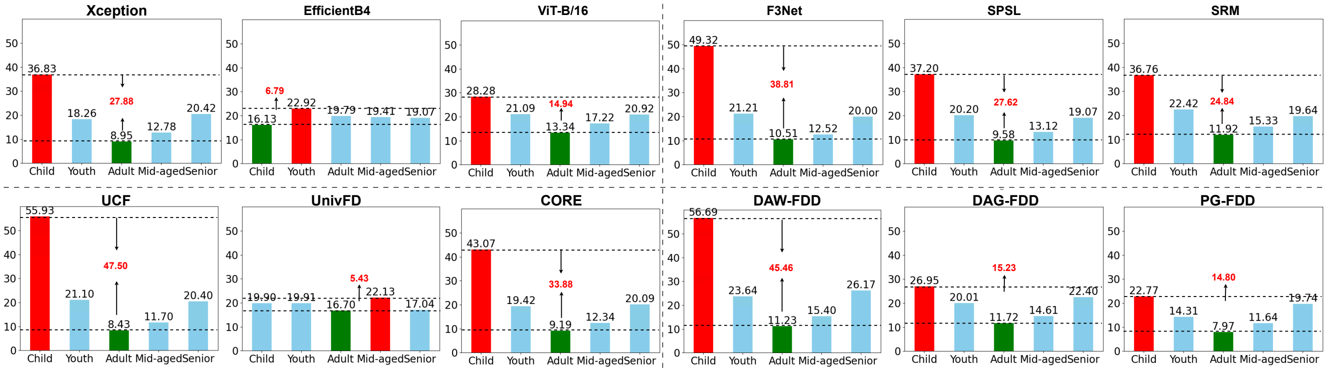


Figure D.1. *FPR(%) of each age subgroup. The subgroup with the highest FPR score is highlighted in red, while the subgroup with the lowest FPR score is shown in green.*

Method	Fairness																Utiliy				
	Skin Tone				Gender				Age				Intersection				-				
	F_{MEO}	F_{DP}	F_{OAE}	F_{EO}	F_{MEO}	F_{DP}	F_{OAE}	F_{EO}	F_{MEO}	F_{DP}	F_{OAE}	F_{EO}	F_{MEO}	F_{DP}	F_{OAE}	F_{EO}	AUC	ACC	AP	EER	FPR
EfficientB4	5.385	1.725	1.487	5.863	8.300	**6.184**	4.377	11.062	6.796	11.849	2.856	10.300	17.586	**8.607**	8.461	25.114	98.611	94.203	99.542	6.689	20.066
SPSL	4.411	1.827	1.037	4.534	8.055	9.379	1.135	9.789	27.614	11.232	7.270	40.943	10.379	13.259	2.464	21.679	98.747	96.346	99.356	4.371	13.661
UnivFD	4.503	1.19	1.622	5.408	2.577	8.556	2.748	5.536	5.436	15.249	3.793	14.148	6.119	14.026	6.287	20.255	98.192	93.651	99.400	7.633	18.550
PG-FDD	3.190	1.252	1.071	3.702	6.465	9.746	0.882	9.115	14.804	**10.467**	5.009	29.585	9.578	14.697	3.062	18.348	99.172	96.174	99.694	4.961	10.971
CLIP+MLP	**0.419**	**0.938**	**0.227**	**0.591**	**0.506**	8.658	**0.334**	**1.021**	**0.765**	14.473	**0.395**	**1.802**	**1.973**	13.992	**1.000**	**4.302**	**99.973**	**99.290**	**99.991**	**0.793**	**1.171**

Table D.9. *Fairness and utility performance of CLIP+MLP compared to representative detectors on the AI-Face dataset, highlighting the potential of foundation models for bias mitigation.*

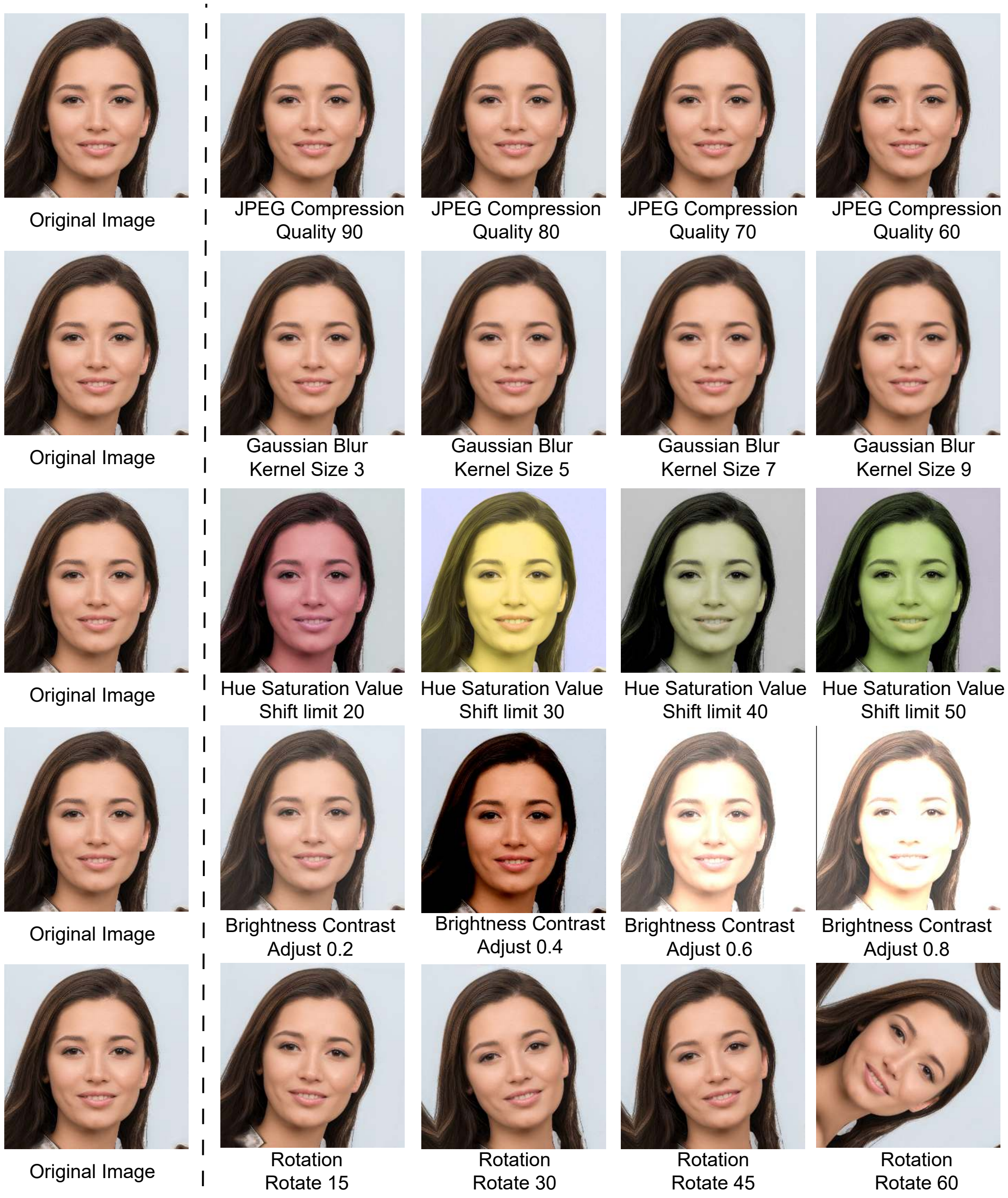

Original Image
JPEG Compression Quality 90
JPEG Compression Quality 80
JPEG Compression Quality 70
JPEG Compression Quality 60
Original Image
Gaussian Blur Kernel Size 3
Gaussian Blur Kernel Size 5
Gaussian Blur Kernel Size 7
Gaussian Blur Kernel Size 9
Original Image
Hue Saturation Value Shift limit 20
Hue Saturation Value Shift limit 30
Hue Saturation Value Shift limit 40
Hue Saturation Value Shift limit 50
Original Image
Brightness Contrast Adjust 0.2
Brightness Contrast Adjust 0.4
Brightness Contrast Adjust 0.6
Brightness Contrast Adjust 0.8
Original Image
Rotation Rotate 15
Rotation Rotate 30
Rotation Rotate 45
Rotation Rotate 60

Figure D.2. *Visualization of the image after different post-processing.*

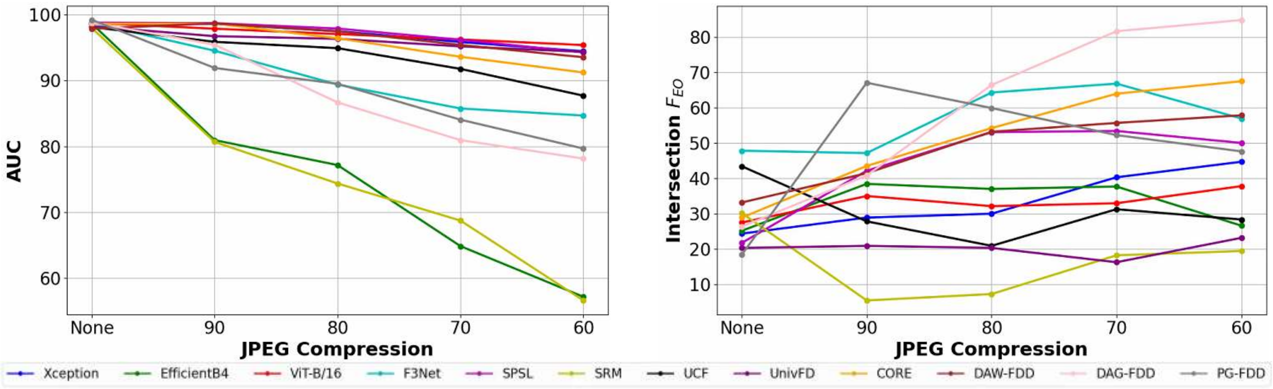


Figure D.3. *Robustness analysis in terms of utility and fairness under varying degrees of JPEG compression.*

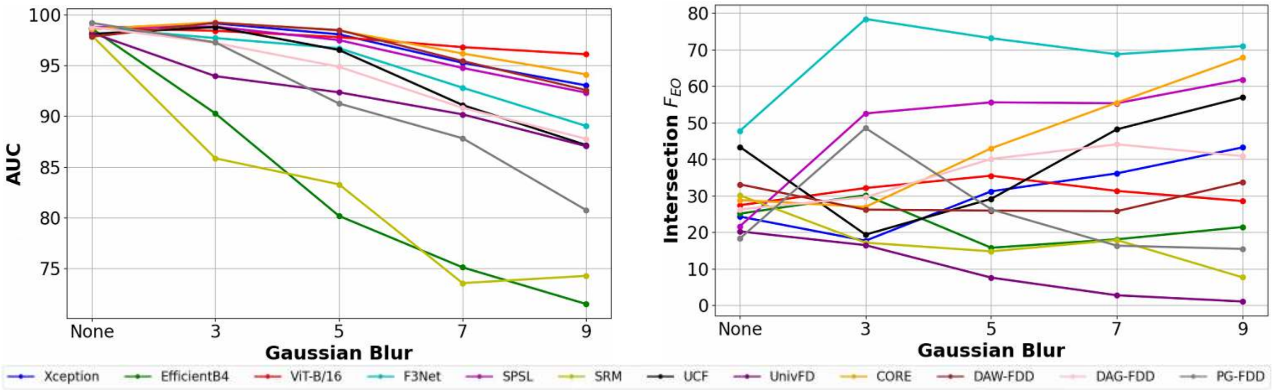


Figure D.4. *Robustness analysis in terms of utility and fairness under varying kernel sizes of Gaussian Blur.*

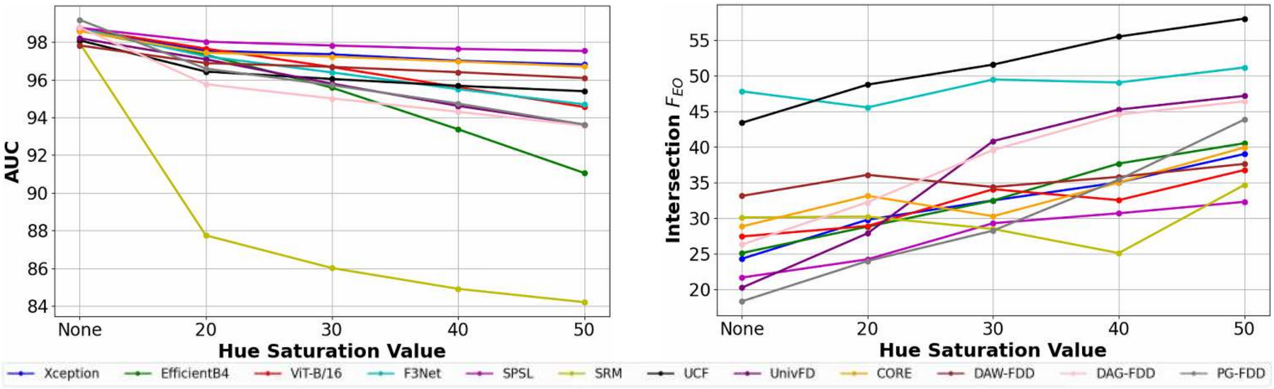


Figure D.5. *Robustness analysis in terms of utility and fairness under varying degrees of Hue Saturation Value.*

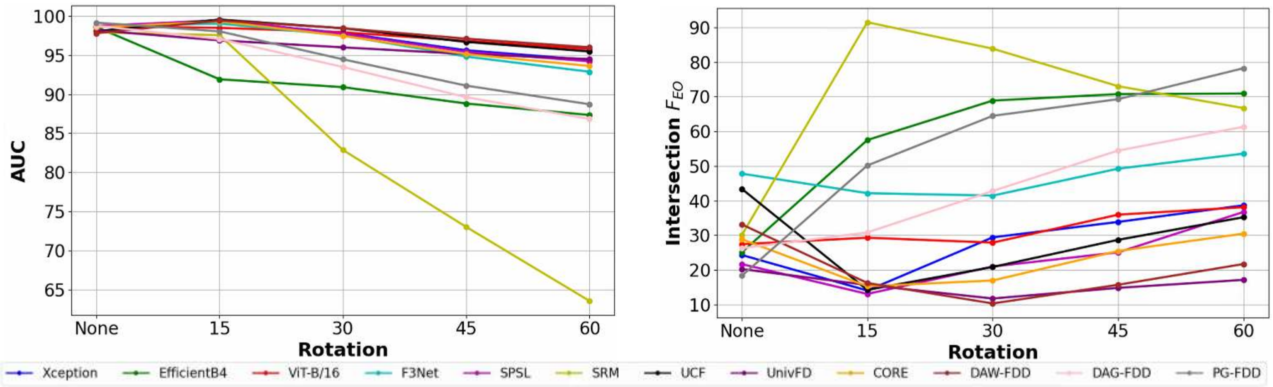


Figure D.6. *Robustness analysis in terms of utility and fairness under varying degrees of Rotations.*

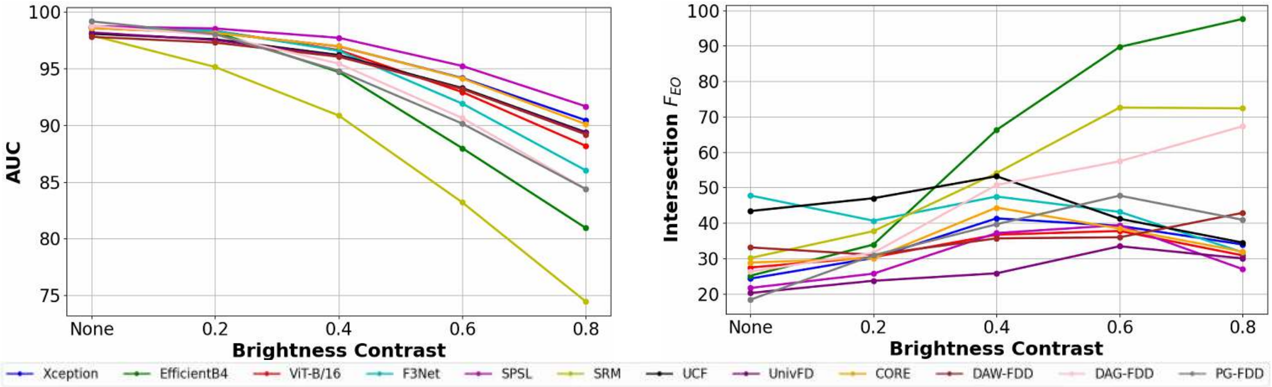


Figure D.7. *Robustness analysis in terms of utility and fairness under varying degrees of Brightness Contrast.*

E. Datasheet for AI-Face

In this section, we present a DataSheet [114] for AI-Face.

E.1. Motivation For Dataset Creation

- **Why is the dataset created?** For researchers to evaluate the fairness of AI face detection models or to train fairer models. Please see Section 2 'Background and Motivation' in the submitted manuscript.
- **Has the dataset been used already?** Yes. Our fairness benchmark is based on this dataset.
- **What (other) tasks could the dataset be used for?** Could be used as training data for generative methods attribution task.

E.2. Data Composition

- **What are the instances?** The instances that we consider in this work are real face images and AI-generated face images from public datasets.
- **How many instances are there?** We include 1,646,545 face images from public datasets. Please see Table B.1 for details.
- **What data does each instance consist of?** Each instance consists of an image.
- **Is there a label or target associated with each instance?** Each image is associated with gender annotation, age annotation, skin tone annotation, intersectional attribute (gender and skin tone) annotation, and target label (fake or real).
- **Is any information missing from individual instances?** No.
- **Are relationships between individual instances made explicit?** Not applicable – we do not study the relationship between each image.
- **Does the dataset contain all possible instances or is it a sample?** Contains all instances our curation pipeline collected. Since the current dataset does not cover all available images online, there is a high probability more instances can be collected in the future.
- **Are there recommended data splits (*e.g.*, training, development/validation, testing)?** For detector development and training, the dataset can be split as 6:2:2.
- **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.** Yes. Despite our extensive efforts to mitigate the bias that may introduced by the automated annotator and reduce demographic label noise, there may still be mislabeled instances. Given the dataset's size of over 1 million images and most are generated face images, it is impractical for humans to manually check and correct each image individually.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** The dataset is self-contained.

E.3. Collection Process

- **What mechanisms or procedures were used to collect the data?** We build our AI-Face dataset by collecting and integrating public AI-generated face images sourced from academic publications, GitHub repositories, and commercial tools. Please see 'Data Collection' in Section 3.1
- **How was the data associated with each instance acquired? Was the data directly observable (*e.g.*, raw text, movie ratings), reported by subjects (*e.g.*, survey responses), or indirectly inferred/derived from other data?** The data can be acquired after our verification of user submitted and signed EULA.
- **If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?** Not applicable. We did not sample data from a larger set. But we use RetinaFace [66] for detecting and cropping faces to ensure each image only contains one face.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (*e.g.*, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.** The data was collected from February 2024 to April 2024, even though the data were originally released before this time. Please refer to the cited papers in Table B.1 for specific original data released time.

E.4. Data Processing

- **Was any preprocessing/cleaning/labeling of the data done (*e.g.*, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** Yes. We discussed in 'Data Collection' in Section 3.1.

- **Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*, to support unanticipated future uses)? If so, please provide a link or other access point to the 'raw' data.** The 'raw' data can be acquired through the original data publisher. Please see the cited papers in Table B.1.
- **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.** Yes. We use RetinaFace [66] for detecting and cropping faces to ensure each image only contains one face. Demographic annotations are given by our annotator, see 'Annotation Generation' in Section 3.2. Our annotator code will not be released considering the ethical guidelines.
- **Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?** Yes. The dataset does allow for the study of our goal, as it covers comprehensive generation methods, demographic annotations for evaluating current detectors and training fairer detectors.

E.5. Dataset Distribution

- **How will the dataset be distributed?** We distribute all the data as well as CSV files that formatted all annotations of images under the CC BY-NC-ND 4.0 license and strictly for research purposes.
- **When will the dataset be released/first distributed? What license (if any) is it distributed under?** The dataset will be released following the paper's acceptance, and it will be under the permissible CC BY-NC-ND 4.0 license for research-based use only. Users can access our dataset by submitting an EULA.
- **Are there any copyrights on the data?** We believe our use is 'fair use' since all data in our dataset is collected from public datasets.
- **Are there any fees or access restrictions?** No.

E.6. Dataset Maintenance

- **Who is supporting/hosting/maintaining the dataset?** The first author of this paper.
- **Will the dataset be updated? If so, how often and by whom?** We do not plan to update it at this time.
- **Is there a repository to link to any/all papers/systems that use this dataset?** Our fairness benchmark uses this dataset, a brief instruction of how to use this dataset and the code of fairness benchmark is on `https://github.com/Purdue-M2/AI-Face-FairnessBench`.
- **If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?** Not at this time.

E.7. Legal and Ethical Considerations

- **Were any ethical review processes conducted (*e.g.*, by an institutional review board)?** No official processes were done since all data in our dataset were collected from the existing public datasets.
- **Does the dataset contain data that might be considered confidential?** No. We only use data from public datasets.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why** No. It is a face image dataset, we have not seen any instance of offensive or abusive content.
- **Does the dataset relate to people?** Yes. It is a face image dataset containing real face images and AI-generated face images.
- **Does the dataset identify any subpopulations (*e.g.*, by age, gender)?** Yes, through demographic annotations.
- **Is it possible to identify individuals (*i.e.*, one or more natural persons), either directly or indirectly (*i.e.*, in combination with other data) from the dataset?** Yes. It is a face image dataset. The age, gender, and skin tone can be identified through the face image, also through the demographic annotation we provide. All of the images that we use are from publicly available data.

E.8. Author Statement and Confirmation of Data License

The authors of this work declare that the dataset described and provided has been collected, processed, and made available with full adherence to all applicable ethical guidelines and regulations. We accept full responsibility for any violations of rights or ethical guidelines that may arise from the use of this dataset. We also confirm that the dataset is released under the CC BY-NC-ND 4.0 license, permitting sharing and downloading of the work in any medium, provided the original author is credited, and it is used non-commercially with no derivative works created.