Understanding and Mitigating Annotation Bias in Facial Expression Recognition

Yunliang Chen Jungseock Joo
University of California, Los Angeles
chenyunliang@ucla.edu, jjoo@comm.ucla.edu

Abstract

The performance of a computer vision model depends on the size and quality of its training data. Recent studies have unveiled previously-unknown composition biases in common image datasets which then lead to skewed model outputs, and have proposed methods to mitigate these biases. However, most existing works assume that human-generated annotations can be considered gold-standard and unbiased. In this paper, we reveal that this assumption can be problematic, and that special care should be taken to prevent models from learning such annotation biases. We focus on facial expression recognition and compare the label biases between lab-controlled and in-the-wild datasets. We demonstrate that many expression datasets contain significant annotation biases between genders, especially when it comes to the happy and angry expressions, and that traditional methods cannot fully mitigate such biases in trained models. To remove expression annotation bias, we propose an AU-Calibrated Facial Expression Recognition (AUC-FER) framework that utilizes facial action units (AUs) and incorporates the triplet loss into the objective function. Experimental results suggest that the proposed method is more effective in removing expression annotation bias than existing techniques.

1. Introduction

Computer vision models rely heavily on large sets of training images. Unfortunately, most datasets are "biased" in one way or another [58]. Traditional (*i.e.*, lab-controlled) datasets are often too small and not diverse enough to train a robust model. Recently, many large-scale image datasets have been created through web-scraping and crowdsourced annotations [16, 95]. While this practice helps researchers collect millions of diverse "in-the-wild" images rapidly at low cost, it also introduces an undesired problem of dataset bias [79, 78, 64]. To mitigate the problem of biases effectively, we need to know (1) what causes biases (source), (2) which specific problems, datasets, or models suffer from biases, and (3) which methods are effective in each situation. Machine learning models, unless explicitly modified, have been shown to be capable of learning bias from data [23] and, consequently, to produce biased outcomes against certain groups of people, undermining fairness and social trust of AI systems [83, 30, 94, 8, 9, 18, 26, 48].

We consider the scenario of supervised learning. Let $X = \{X_i\}_{i=1}^N$ denote the collection of input images, and $Y = \{Y_i\}_{i=1}^N$ be the set of labels. A dataset is unbiased if the joint distribution $P(X, Y)$ matches reality. In particular, this requires the annotated labels, $Y|X$, to be unbiased.

For tabulated data, label bias is a classical focus in the fairness literature, where machine learning models are applied to some historically discriminatory data whose labels are unfair to certain racial or gender groups [65], such as recidivism prediction [12], loan approval, and employment decisions [89, 63]. For crowdsourced image annotations, however, it is often assumed that the annotations are not *systematically* biased. Each annotator may have their personal biases, and there may be labeling mistakes, but given the diversity and large size of the data, they are generally assumed to be just another component of random noises [6, 96].

In reality, however, it is unlikely that people's biases are all idiosyncratic. In fact, annotators may possess *systematic* cultural or societal biases, and if not specifically trained, they may incorporate such biases into their annotations. As a result, models trained on such data will become unfair. In this paper, we investigate the presence of systematic annotation bias in large in-the-wild datasets. We focus on the task of facial expression recognition. Fairness in expression recognition has not received wide attention [87, 70], yet it has a profound impact: more and more companies nowadays conduct video job interviews in which algorithms are used to evaluate applicants' facial expressions, voice, and word selection to predict their skills, behaviors, and personality traits [90, 67, 28]; in addition, automated emotion analysis is already ubiquitous and used in consumer analysis, content recommendation, clinical psychology, lie detection, pain assessment, and many other human computer interfaces (*e.g*. "smile" shutters) [76, 68].

In the context of facial expression recognition, studies in psychology have shown that human observers are more likely to perceive women's faces as happier than men's faces even when their smiles have the same intensity [75], and it is believed that raters hold cultural stereotypes and that these stereotypes influence the judgment of emotions [31, 46]. We hypothesize that such bias is present in many in-the-wild expression datasets whose labels are annotated by non-experts. In particular, we seek to answer the question: "Are annotators equally likely to assign different expression labels between males and females?" As we will show, for subjective tasks such as facial expression recognition, image annotations can be systematically biased, and special efforts need to be taken to address such bias.

We note that, currently, most debiasing techniques in the deep learning literature focus on biases that come from the images themselves (*i.e*., the bias in the distribution $P(X)$). This is often known as "dataset bias" [79, 78] or "sample selection bias" [64]. It happens when the dataset is biased in its *composition* of images. As a result, models trained on one dataset do not generalize well to the real world due to the domain shift between the source and target. The trained model can also have undesirable accuracy differences across different groups or classes [9]. Additionally, the data may contain spurious or undesirable correlations. When such undesirable correlation involves protected attributes (*e.g*., gender, race, or age), the model is considered "unfair." Numerous methods have been proposed to decorrelate these attributes and ensure that models trained on such data do not discriminate people based on their protected attributes [71, 2, 34, 53, 62, 11, 60, 38].

However, debiasing $P(X)$ does not solve all problems since the joint distribution $P(X, Y)$ will still be biased if the annotated labels, $P(Y|X)$, are biased. As we will demonstrate in this paper, existing techniques that are designed to mitigate data *composition* bias fail to fully mitigate the bias that comes from *annotations*. On the other hand, classical methods that address label bias are intrusive in that they often involve changing the labels prior to training [55, 39]. In this paper, we address annotation bias that arises in facial expression recognition tasks. We propose an AU-Calibrated Facial Expression Recognition (AUC-FER) framework that uses the facial action units (AUs) and incorporates the triplet loss into the objective function to learn an embedding space in which the expressions are classified similarly for people with similar AUs. We demonstrate that the presented method more effectively mitigates annotation biases than existing methods. We note that although our framework is designed for facial expression recognition, it can be adapted to many other applications that require subjective human labeling such as activity recognition or image captioning but some fair or objective measures are available (such as the AUs and body keypoints).

The contribution of this paper is threefold:

- We compare the existence of annotation bias between lab-controlled datasets and in-the-wild datasets for facial expression recognition and observe that in-the-wild datasets often contain significant systematic bias in their annotations. To the best of our knowledge, this is the first work to demonstrate the effect of systematic annotation bias associated with image data.
- We further demonstrate that such systematic annotation bias will be learned by trained models and thus cannot be ignored as is often assumed in the literature.
- We propose a novel AU-Calibrated Facial Expression Recognition (AUC-FER) framework that utilizes facial action units to remove expression annotation bias. Experiments suggest that it outperforms existing debiasing techniques for removing annotation bias.

2. Related Work

As discussed in the previous section, the focus of this paper is the bias of $P(Y|X)$. We briefly review the literature on fairness and bias specific to this type, as well as research on facial expression recognition.

Fairness. Fairness generally means that the model is not discriminatory with respect to some protected attribute, such as race, color, religion, sex, or national origin [33]. Many formal definitions of fairness exist, and they generally can be divided into two types: *group fairness*, which requires different demographic groups to receive the same treatment on average [27], and *individual fairness*, which requires individuals who are similar to have similar probability distributions on classification outcomes [20]. As is common notation, we will denote the *protected variable* by Z and the model prediction by $\hat{Y}$. A major barrier to achieving individual fairness is the selection of a similarity measure between individuals. A closely related concept is *counterfactual fairness*, which requires the decision to be unchanged had the person belonged to a different demographic group while keeping everything else the same [49]. Denton *et al*. [17] build on this idea and use a generative model that can manipulate specific attributes of faces (*e.g*., from *young* to *old*) to reveal the biases of a smile classifier.

Debiasing techniques. Common techniques to address dataset bias include transfer learning [64], domain adaptation [80, 81, 25, 82], and adversarial mitigation [91, 83]. Many methods have also been proposed to remove or prevent models from learning spurious or undesirable correlations. Hardt *et al*. [27] propose a post-hoc correction technique that enforces equality of odds on a learned predictor. Other group fairness definitions have also been transformed into constrained optimization problems [63, 94, 88, 89].

Robinson *et al*. [71] propose learning subgroup-specific thresholds. In the realm of deep learning, modifying the loss functions to penalize unfairness [1] and adversarial learning [69, 91, 35, 60] are two common techniques, with the goal of learning a "fair" representation that does not contain information of the protected attribute Z.

In the case where the data labels are historically biased, data massaging is the most commonly used technique. This includes directly correcting the labels by changing them prior to training [55, 39], or use some weights or sampling techniques during training [40, 41].

Annotation bias. For tabulated data, historical label bias is a well-known issue [65]. Jiang and Nachum [36] propose a re-weighting scheme that can correct label bias under certain assumptions about the relationship between the biased labels and the true labels. In the case of large-scale in-the-wild datasets prepared for deep learning, however, annotation bias has received little attention compared to the more salient data composition bias. Regardless of the exact methods through which the images are labeled (manual, semi-automatic, or automatic), the general assumption is that they add random noise to the labels but are unbiased on average [6, 96]. In the case where each image is annotated by multiple workers, the focus has been on improving the compilation step of the dataset creation process to increase the accuracy of the labels. Methods have also been developed to fix errors in the case of multi-label supervised learning [15]. Zhuang and Young [96] note that presenting data items in batches to annotators can lead to in-batch annotation bias. In general, crowd annotators have lower accuracy when labeling difficult cases, but researchers have found that this is relatively unproblematic under certain conditions [6]. In this paper, we examine the bias of labels in the case of facial expression recognition, and we will show that, unlike what previous studies assumed, *systematic* bias exists and needs to be actively managed.

Facial expression recognition and facial action units. Facial expression recognition, which analyzes people's *expressed* emotions from visual data [76], is one of the central tasks in facial analysis and widely used in many domains such as media analytics [37, 86, 66], HCI [4, 14], education [45], and psychology [50, 13]. A seminal study conducted by Ekman and Friesen [22] identified six prototypical emotions: anger, disgust, fear, happiness, sadness, and surprise. They noticed that the association between certain facial muscular patterns and discrete emotions is universal and independent of gender and race, and adopted a Facial Action Coding System (FACS) consisting of facial action units (AUs) [21] that objectively code the fundamental muscle actions typically seen for various facial expressions of emotion [19]. Early works in facial expression recognition are often rule-based methods using FACS [77].

With deep learning, the average performance of expression recognition has significantly improved, and many works recently started to focus on model bias and dataset (composition) bias. A common observation is that disgust, anger, fear, and surprise are minority classes in datasets and harder to learn compared to happiness and sadness [51], and classical methods for addressing data composition bias such as weighting, re-sampling, data augmentation, hierarchical modeling [32], and confusion loss [87] have been proposed.

In another line of research, studies have shown that women look happier than men [75] and that people are faster and more accurate at detecting angry expressions on male faces and happy expressions on female faces [5]. As a result, correction of the annotations is necessary [75]. On a similar note, Denton *et al*. [17] find that a smiling classifier trained on CelebA is more likely to predict "smiling" when they remove the person's beard or apply makeup or lipstick to the image but keep everything else the same. Based on these psychological studies as well as the observed model bias, we hypothesize that systematic annotation bias exists in many large in-the-wild expression datasets and it (in addition to the data composition bias) contributes to the gender bias in trained models.

3. Annotation Bias in Expression Datasets

In this section, we illustrate the existence of systematic annotation bias in facial expression datasets. As previously noted, psychological studies have shown that raters tend to hold stereotypical biases that women are happier than men [31] and that they detect angry expressions on male faces more quickly [5], we hypothesize that these biases will manifest themselves in annotated datasets. In particular, we examine the "happiness" and "anger" annotations and ask the question: Are annotators equally likely to assign happiness/anger labels to male and female images – if they indeed show the same expression? In order to quantify the "same" expression, we use the AUs since they were specifically designed to measure facial expressions objectively, and past studies have used them to assess the accuracy in the imitation of facial expressions [46]. We mainly focus on gender due to its well-studied psychological connection to expression perception. We also conduct analysis on age and race. However, unlike gender, most public datasets used in our experiments are not well-balanced between different age and racial groups but are instead heavily dominated by younger and white people. The full analysis on age and race is included in the Supplementary Material.

3.1. Facial Action Units (AUs) Recognition

In the framework of FACS, happiness is defined as the combination of AU6 (cheeks raised and eyes narrowed) and AU12 (lip corners pulled up and laterally), and anger is defined as the combination of AU4 (brow lowerer), AU5 (upper lid raiser), AU7 (lid tightener), and AU23 (lip tightener)

	AU6		AU12	
	Accuracy	F1	Accuracy	F1
Male	0.859	0.613	0.887	0.830
Female	0.860	0.598	0.885	0.866
p-value	0.835		0.715	

Table 1. Accuracies and F1-scores of OpenFace AU Recognition, evaluated on 24,600 EmotioNet images with expert-coded AUs.

[57, 43]. Therefore, we will use them as objective benchmarks to evaluate the classification of emotions. Due to limited space, we include in the paper the numerical results for the happiness expression only; detailed analysis for anger is presented in the Supplementary Material.

We use OpenFace, a state-of-the-art facial behavior analysis toolkit [3], for our facial action unit recognition purpose. In order for it to serve as a benchmark for evaluating the bias of emotion annotations, we first check that its AU recognition is not biased between males and females itself.

We use EmotioNet [24], which includes 24,600 images with AUs manually annotated by experienced coders, to evaluate the performance of AU presence and intensity recognition by OpenFace. Since OpenFace and EmotioNet use different thresholds when binarizing the AU variables, we use OpenFace's AU intensity output to re-classify AU presence by choosing the threshold that optimizes the overall classification accuracy for each AU based on EmotioNet annotations.

We use the FairFace dataset [44] to train a simple gender classifier that achieves a test accuracy of 94.5%. We then use it to classify the 24,600 EmotioNet images and this enables us to test whether the performance of OpenFace differs by gender. Table 1 summarizes the accuracies and F1-scores for the calibrated OpenFace AU6 and AU12 output between males and females. We can see from the p-values of the t-tests for the accuracies that the differences are insignificant for both AU6 and AU12. Therefore, we conclude that even though OpenFace's AU6 and AU12 recognition is imperfect, it is unbiased between males and females and thus can be used as a proxy for the true AUs and as an objective benchmark for evaluating happiness annotations. A similar evaluation is conducted for AUs associated with the angry expression; see the Supplementary Material for details).

3.2. Expression Annotation Bias

As we mentioned previously, there are two potential sources of bias that in-the-wild datasets may contain: *data composition bias* (*e.g.*, the data contains significantly more happy women and unhappy men) and *annotation bias* (*e.g.*, even when two images are the same otherwise, a woman is more likely to be annotated as "happy" than a man). Since expressions are objectively defined as combinations of AUs, those respective AUs can help make the important distinction between these two biases.

Definition 1 ANNOTATION BIAS. *Let $Y \in \{0,1\}$ denote the emotion label. Let $Z \in \{M,F\}$ denote the gender (or some other protected attributes) of the person. We say that the expression annotations are unbiased if*

$$Y \perp\!\!\!\perp Z | AU \tag{1}$$

For happiness annotations, this means

$$\begin{aligned} &P(Y=1|AU6, AU12, Z=\mathrm{M}) \\ &= P(Y=1|AU6, AU12, Z=\mathrm{F}) \end{aligned} \tag{2}$$

where the AUs can be discrete (*i.e.*, $(AU6, AU12) \in \{(0,0),(0,1),(1,0),(1,1)\}$) or continuous (*i.e.*, intensity scores). The case for anger annotations is similar.

Remark. This definition is similar to equality of odds ($\hat{Y} \perp\!\!\!\perp Z|Y$) except that each image is conditioned on the AUs and the requirement is to the labels Y instead of model predictions $\hat{Y}$. Note that the conditioning on AU is crucial because otherwise, we would not have been able to separate annotation bias from data composition bias (that is, it is possible that the female images in the dataset are less happy than males on average, but they are annotated with a larger probability to be "happy" and so it looks as if the data does not contain any bias).

3.3. Evaluation on Various Datasets

We evaluate the expression annotations on various popular expression datasets. They can be categorized into two types: those whose images were collected in a laboratory-controlled condition and those whose images were scraped from the web (*i.e.*, "in-the-wild"). For the first type, we select the Karolinska Directed Emotional Faces database (KDEF) [54] and the Chicago Face Database (CFD) [56]. For the second type, we select the Expression in-the-Wild Database (ExpW) [92, 93], the Real-world Affective Face Database (RAF-DB) [52], and AffectNet [59].

KDEF [54]: KDEF contains 70 individuals displaying the 6 basic expressions plus neutral. Each expression is viewed from 5 angles and shot twice. However, for comparability with other databases, we will only use the 980 front-view photos among the 4,900 images.

CFD [56]: CFD contains photos of 597 individuals with a neutral expression. For a subset of 158 targets, it also includes happy, angry, and fearful expressions.

ExpW [92, 93]: ExpW is an in-the-wild dataset consisting of 91,793 faces. Each face is manually annotated as one of the 6 basic expressions plus neutral.

RAF-DB [52]: RAF-DB contains 29,672 facial images downloaded from the web. Using crowdsourcing, each image is independently labeled by about 40 annotators. In particular, 15,339 of them are classified into one of the 6 basic

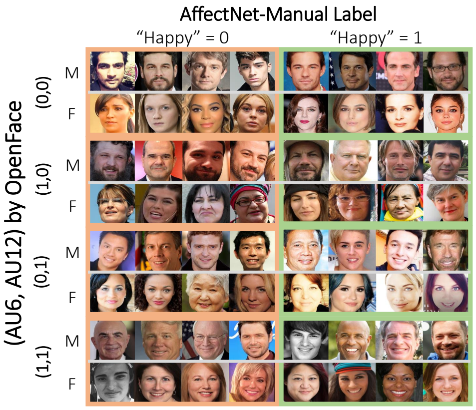

Figure 1. Examples of “happy” and “not happy” faces from AffectNet-Manual for each (AU6, AU12) combination and for each gender. The emotion labels in AffectNet-Manual come from manual annotation but possibly contain errors, and the AUs are detected using OpenFace.

expressions plus neutral. Gender, age, and race annotations are also provided.

AffectNet [59]: AffectNet contains about 1M facial images collected from the web. About half (420K) of the images (denoted as AffectNet-Manual) are manually annotated as one of the 6 basic expressions plus contempt and neutral. The rest (550K) (denoted as AffectNet-Automatic) are automatically annotated using ResNext Neural Network trained on all manually annotated training set samples with average accuracy of 65%. For the purpose of our evaluation, we will use random samples of size 38,889 and 45,369 for AffectNet-Manual and AffectNet-Automatic respectively instead of the entire datasets.

For each of the above datasets, we apply the OpenFace AU detector and obtain the AU6 and AU12 intensities for each image. They are then binarized into AU presence variables using the optimal thresholds found in Section 3.1. We also apply our gender classifier when gender information is not available (*i.e.*, for ExpW and AffectNet). Note that even though “happiness” is formally defined to be the presence of AU6 and AU12, the fact that both the expression and the AUs are inherently continuous-valued means that discretization may result in a few cases that violate the rule. In practice, AU detection and expression annotation are imperfect and will introduce additional noises. Nevertheless, the pattern should be similar between males and females if the errors are random. See Figure 1 for some examples of “happy” and “not happy” faces from AffectNet-Manual for each (AU6, AU12) combination.

Table 2 shows the proportion of “happy” labels among males and females conditioned on different values of AU6 and AU12. For each conditional distribution of “happy,” a chi-square test of independence is used to determine whether there is a significant relationship between the labels and gender after controlling for the AUs. Due to the limited

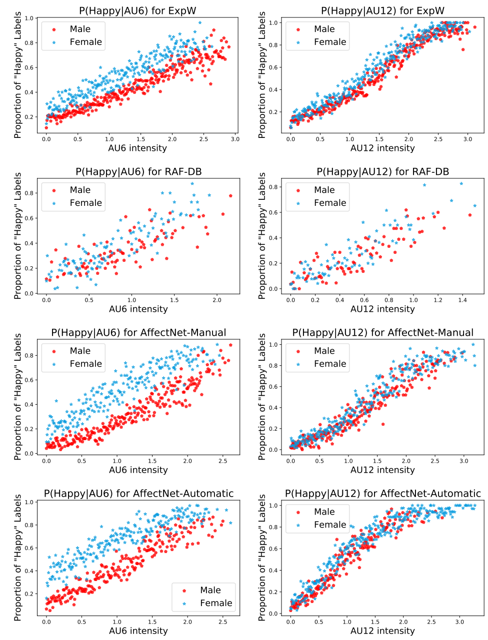

Figure 2. Proportion of “happy” labels among males and females conditioned on AU6 and AU12 intensity for each “in-the-wild” expression dataset. Significant differences can be seen between males and females, suggesting the presence of annotation bias.

sizes of KDEF and CFD, some (AU6, AU12) combinations do not contain enough data for the chi-square tests and thus a single AU (*i.e.*, AU6 only or AU12 only) is used as a condition. It is important to note that even though OpenFace is not perfectly accurate, we have demonstrated that it does not contain systematic bias with respect to gender (*i.e.*, its errors are random), and thus any systematic bias in emotion annotations conditioned on the AUs must be due to the bias in emotion annotations, not the AUs.

From Table 2, we can see significant differences between lab-controlled datasets and in-the-wild datasets. For both KDEF and CFD, the distribution of “happy” labels is independent of gender when AU6 and AU12 are controlled. On the other hand, for ExpW, RAF-DB, and AffectNet, the proportions of “happy” labels are significantly higher for females than males even when the AUs have been controlled. We believe that the significantly less annotation bias in lab-controlled datasets can be explained by the fact that those images are often carefully vetted by experts before being released while “in-the-wild” datasets are often annotated by laymen who are not specifically trained to overcome their cognitive bias or unconscious stereotyping. Comparing AffectNet-Manual and AffectNet-Automatic, we see that the levels of annotation bias are similar, indicating that the model used to automatically label the 540K images inherits the label bias in the manually-labeled dataset.

Figure 2 shows the proportion of “happy” labels as a function of AU6 and AU12 intensity for each in-the-wild

Data (Collecting Condition, Size)	Conditioned on Joint AU					Conditioned on Marginal AU				
	(AU6, AU12)	P(Happy |AU, M)	P(Happy |AU, F)	Δ	p-value of χ^2 test for $Y \perp\!\!\!\perp Z$	AU	P(Happy |AU, M)	P(Happy |AU, F)	Δ	p-value of χ^2 test for $Y \perp\!\!\!\perp Z$
KDEF (Lab, 980) [54]						AU6=0	0.036	0.016	-0.019	0.095 . [1]
						AU6=1	0.475	0.547	0.072	0.268
						AU12=0	0.000	0.005	0.005	-
	(1,1)	0.838	0.771	-0.067	0.304	AU12=1	0.769	0.673	-0.096	0.140
CFD (Lab, 1,207) [56]						AU6=0	0.059	0.079	0.021	0.222
						AU6=1	0.838	0.854	0.016	0.706
	(0,1)	0.383	0.487	0.104	0.228	AU12=0	0.005	0.005	-0.001	-
	(1,1)	0.884	0.890	0.006	0.877	AU12=1	0.725	0.751	0.026	0.546
ExpW (Web, 91,793) [92, 93]	(0,0)	0.176	0.215	0.039	0.000 ***	AU6=0	0.255	0.336	0.081	0.000 ***
	(1,0)	0.246	0.285	0.040	0.0488 *	AU6=1	0.646	0.770	0.124	0.000 ***
	(0,1)	0.663	0.770	0.107	0.000 ***	AU12=0	0.179	0.217	0.039	0.000 ***
	(1,1)	0.801	0.870	0.069	0.000 ***	AU12=1	0.716	0.806	0.091	0.000 ***
RAF-DB (Web, 15,339) [52]	(0,0)	0.197	0.192	-0.005	0.570	AU6=0	0.289	0.305	0.016	0.089 .
	(1,0)	0.232	0.254	0.022	0.537	AU6=1	0.632	0.785	0.153	0.000 ***
	(0,1)	0.822	0.868	0.047	0.013 *	AU12=0	0.200	0.195	-0.005	0.572
	(1,1)	0.808	0.905	0.097	0.000 ***	AU12=1	0.814	0.888	0.074	0.000 ***
AffectNet-Manual (Web, 420,299) [59]	(0,0)	0.086	0.125	0.039	0.000 ***	AU6=0	0.165	0.292	0.127	0.000 ***
	(1,0)	0.251	0.254	0.004	0.920	AU6=1	0.676	0.821	0.145	0.000 ***
	(0,1)	0.608	0.725	0.117	0.000 ***	AU12=0	0.093	0.127	0.034	0.000 ***
	(1,1)	0.778	0.860	0.082	0.000 ***	AU12=1	0.699	0.781	0.082	0.000 ***
AffectNet-Automatic (Web, 539,607) [59]	(0,0)	0.151	0.236	0.085	0.000 ***	AU6=0	0.246	0.410	0.164	0.000 ***
	(1,0)	0.431	0.518	0.087	0.024 *	AU6=1	0.822	0.908	0.086	0.000 ***
	(0,1)	0.811	0.873	0.062	0.000 ***	AU12=0	0.162	0.241	0.079	0.000 ***
	(1,1)	0.907	0.934	0.027	0.000 ***	AU12=1	0.861	0.899	0.038	0.000 ***

[1] Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2. Proportion of "happy" labels among males and females conditioned on AU6 and AU12 for each of the popular expression datasets. Here $Y \in \{0, 1\}$ is the "happy" label, $Z \in \{M, F\}$ is the gender attribute. Blanks and omitted p-values indicate that the (AU6, AU12) combinations do not contain enough data for the chi-square tests.

dataset (the size of lab-controlled data is too small to calculate average proportions). As expected, the proportion of "happy" labels is higher when AU6 and AU12 intensities are higher, but the effect is different between males and females. ExpW, AffectNet-Manual, and AffectNet-Automatic all show large discrepancies in the conditional distributions of "happy" labels between males and females while the difference for RAF-DB is smaller. In fact, a logistic regression would show that gender is a significant predictor even when AU6 and AU12 are controlled for all four datasets. This is consistent with the result in Table 2.

For the anger annotations, we also observe a consistent pattern of systematic annotation bias among all in-the-wild datasets, whereas lab-controlled datasets show no signs of annotation bias. For all in-the-wild datasets, males are more likely than females to be labeled as "angry" after the AUs are controlled; see the Supplementary Material for the results. We also check other expression annotations but do not find significant annotation biases between males and females as those observed with "happy" and "angry" annotations. This is partially because many expression classes' occurrence rates are too low in these datasets. For example, surprise, fear, and disgust account for about only 4%, 1%, and 1% of all images in AffectNet-Manual respectively and thus the differences between males and females are minor.

We also conduct analysis on the "happy" annotations across different age and racial groups following a similar procedure. We find that younger people are more likely to be annotated as "happy" compared to older people in general, although the saliency of such annotation bias varies across datasets. We do not find evidence of systematic annotation bias across different racial groups. The full results can be found in the Supplementary Material. For both age and race analyses, further analysis is needed on more balanced datasets (*i.e.*, datasets that have more older people and minority races).

To explain the seemingly contradictory observations that "happy" and "angry" expression labels suffer from significant annotation bias while many AU labels do not, we believe this is because facial action units are local attributes and so the gender information has little impact on the annotators' annotation, whereas when the annotators conduct expression annotation, they tend to look at the faces holistically, and so the gender of the face influences their annotation in a non-negligible way.

4. Bias Correction

4.1. Learned Bias in Trained Models

Having observed the existence of annotation bias in in-the-wild expression datasets, we hypothesize that a naive model trained on these data will learn such bias and that

Training Data	P($\hat{Y}$=1 |F)	P($\hat{Y}$=1 |M)	$Disc$
Raw ExpW	0.3916	0.3342	0.0574
Relabeled ExpW	0.3655	0.3603	0.0052

Table 3. Proportions of "happy" classification among males and females on the CFD test set by a ResNet-50 model trained on ExpW and relabeled ExpW data. For ground truth labels, the proportion of "happy" in the test set is 0.3629 for both males and females.

removing the annotation bias will reduce the bias of the model. To test our hypothesis, we use ExpW as our training data and CFD as our test data. We select CFD because the images are lab-controlled and thus contain fewer confounding factors (such as differences in the backgrounds) when we evaluate the model predictions between males and females. Following convention [39, 42, 97], we use the Calders-Verwer (CV) discrimination score [10] as our metric for the bias of the trained model:

$$Disc = P(\hat{Y} = Happy|F) - P(\hat{Y} = Happy|M) \quad (3)$$

Since the probabilities are no longer conditioned on the AUs, we will need to balance the test data (CFD) so that the proportions of true happy faces are the same between males and females.

We first train a naive happiness classifier using the raw ExpW dataset. We use ResNet-50 [29] pre-trained on ImageNet and fine-tuned by Adam optimization [47] with a learning rate of 0.0001 in PyTorch. To evaluate the effect of annotation bias, we relabel the ExpW data as follows: For each (AU6, AU12) presence combination, we first calculate the average proportion of images that are labeled as "happy." As females are more likely to be labeled as "happy" than males, we randomly sample some "unhappy" male images and relabel them as "happy" and sample some "happy" female images and relabel them as "unhappy" so that they both have the same proportion of images labeled as "happy" conditioned on the AUs. Even though this potentially introduces label errors, these modified labels are statistically fair, or, in other words, systematically unbiased. We then train a happiness classifier on a balanced subset (where each AU and gender combination has 3,000 samples) of the modified data using the same procedure.

For the test set, we remove a few easy happy and unhappy faces from CFD (whose predicted scores from the naive classifier >0.99999 or <0.00001) and then balance the proportions of happiness between males and females by removing some happy female images. As ExpW and CFD use different labeling criteria, the thresholds for binarizing the output of the trained classifier are adjusted to maximize the accuracy on the test set. Table 3 shows the model bias observed on the test set. We see significant bias in the prediction for the model trained on raw ExpW while there is little bias for the model trained on the relabeled ExpW data. This shows that annotation bias can have a significant impact on model fairness and thus should be actively managed.

4.2. Bias Correction

Since data massaging techniques such as changing the labels are intrusive and undesirable (it may have legal implications because it is a form of training on falsified data [36]), in this section, we propose an AU-Calibrated Facial Expression Recognition (AUC-FER) framework that can effectively achieve similar results without the need to modify the labels.

Our goal is to ensure that the model classifies expressions based on the AUs and not the gender, so we want to encourage the model to treat two samples in a similar way if their AUs are similar, even if their genders are different and the labels are different. We note that this is related to the concept of individual fairness (as opposed to group fairness). Our method is motivated by techniques in metric learning, which aims to learn an embedding space where the embedded vectors of similar samples are encouraged to be closer, while dissimilar ones are kept far from each other [74, 84]. In particular, we use the triplet loss function [73] as a regularizer to penalize unfairness.

From the training data, we construct triplets $\{X_i, X_j, X_k\}$ within each batch where X_i and X_j are images with the same AU presence (*e.g.*, (AU6, AU12) for happiness), and X_k is an image with a different AU presence from X_i. The triplet loss is then defined as:

$$\mathcal{L}_{trp} = \sum_{i,j,k}^{N_{trp}} [||f(X_i) - f(X_j)||_2^2 - ||f(X_i) - f(X_k)||_2^2 + \alpha]_+, \quad (4)$$

where $[z]_+ = \max(z, 0)$, and $f(.)$ is the feature representation of the images. The goal of the triplet loss function is to make the distance between X_i and X_j in the embedding space larger than the distance between X_i and X_k by at least a minimum margin α.

As usual, we have cross-entropy loss for classification:

$$\mathcal{L}_{softmax} = -\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i] log(p(y_i)). \quad (5)$$

The total loss function is then defined as the weighted sum of $L_{softmax}$ and L_{trp}:

$$\mathcal{L} = \mathcal{L}_{softmax} + \lambda \mathcal{L}_{trp}, \quad (6)$$

where λ measures how willing we are to deviate from the given biased labels and enforce fairness.

4.3. Experiments

We evaluate the proposed AUC-FER method by comparing it with other debiasing methods in the fairness literature. Popular methods include uniform confusion [1], gradient projection [91], domain discriminative training [85], and domain independent training [85]. Many are motivated by

Methods (ResNet-50 [29])	$Disc$	Compared to Baseline (%)
Baseline	0.059 ± 0.035	-
Uniform Confusion [1]	0.046 ± 0.008	77.6
Gradient Projection [91]	0.036 ± 0.014	60.0
Domain Discriminative [85]	0.076 ± 0.024	128.8
Domain Independent [85]	0.029 ± 0.015	49.4
AUC-FER (Ours)	$\mathbf{0.006 \pm 0.020}$	**10.6**

Table 4. Discrimination scores for various debiasing methods using the ResNet-50 architecture trained on random subsets of ExpW of size 20,000 and tested on CFD for the "happy" expression. The average discrimination scores are compared against the baseline model and shown as a percentage.

Methods (MobileNetV2 [72])	$Disc$	Compared to Baseline (%)
Baseline	0.079 ± 0.009	-
Uniform Confusion [1]	0.085 ± 0.021	107.8
Gradient Projection [91]	0.070 ± 0.036	88.2
Domain Discriminative [85]	0.064 ± 0.029	80.4
Domain Independent [85]	0.062 ± 0.035	78.4
AUC-FER (Ours)	$\mathbf{0.028 \pm 0.029}$	**35.9**

Table 5. Discrimination scores for debiasing methods using the MobileNetV2 architecture trained on random subsets of ExpW of size 8,000 and tested on CFD for the "happy" expression.

techniques in domain adaptation and are designed to reduce data composition bias. We compare them with AUC-FER to evaluate their effectiveness in mitigating annotation bias.

For the first set of experiments, we use the ResNet-50 architecture [29] pre-trained on ImageNet in PyTorch. For the four benchmark models, we follow Wang *et al*. [85] and replace the FC layer of the ResNet model with two consecutive FC layers both of size 2,048 with Dropout and ReLU in between. For AUC-FER, we use the PyTorch Metric Learning library [61] for the triplet loss implementation. All models are trained on random subsets of ExpW of size 20,000 and tested on the previously constructed CFD test set. The thresholds for binarizing the output scores are again chosen to maximize the accuracy on the test set, and the experiment is repeated 5 times for each model. To test the robustness of AUC-FER with respect to the model architecture and the size of training data, we repeat this experiment using MobileNetV2 [72] and a training set of size 8,000.

Tables 4 and 5 show the discrimination scores for the models and compare them with baseline ResNet-50 and MobileNetV2 models. AUC-FER obtains the lowest discrimination score, which is a 64-89% reduction in bias compared to the baseline models and is very close to the result we get by relabeling the training data. This shows that the proposed AUC-FER framework is effective in removing annotation bias. We also perform experiments for the angry expression using AffectNet-Automatic as training data, and AUC-FER again outperforms other debiasing techniques. The experiment details and analysis for the anger expression are included in the Supplementary Material.

5. Discussion

In this paper, we study systematic biases in human annotations in public datasets on facial expressions. To our knowledge, this is the first work in computer vision to demonstrate the systematic effect of annotators' perceptual bias as a potential source of bias that can be injected into computer vision models. We show that, contrary to the common assumption that annotation errors are just random noises, systematic biases exist in many facial expression datasets. The problem is more severe for in-the-wild datasets than lab-controlled datasets. We illustrate that if these biases are not addressed, trained models will also be biased. We further develop an AUC-FER framework to address annotation bias for expression recognition tasks and demonstrate that it is more effective in reducing annotation bias than existing debiasing methods.

The presented framework for facial expression recognition utilizes AUs as an auxiliary variable to enforce fairness since they are specifically designed to resolve subjectivity in facial analysis. This framework can be extended beyond expression recognition. In general, one can use any objective measures (*e.g*. body keypoints) for tasks requiring subjective human labeling (*e.g*. activity recognition or image captioning) within the proposed framework. Although such objective measures may not always be accurate in practice (*e.g*., applying OpenFace introduces additional noises), the belief is that because these measures (AUs, body keypoints) are often local attributes and less affected by other attributes of the subjects (*e.g*., gender, race, or age), they are fairer than the subjective labels in the training data and can thus be used as calibration for fairness.

For future work, we believe that combining our method with other debiasing techniques may potentially be effective when the training data suffers from multiple sources of biases (both composition bias and annotation bias).

This paper focuses on the identification and mitigation of systematic annotation bias. It would be interesting for dataset curators to study if such annotation bias varies across annotator subgroups. Recent work has also pointed out that the prototypical framework of six expressions does not capture the full facial expressions of humans [19], and compound emotions have been proposed to address the genuine ambivalence on some displayed facial expressions [19, 24, 52, 7]. Future work can study the role of these definitions and their interaction with bias.

Acknowledgement This work was supported by NSF SBE/SMA #1831848 "RIDIR: Integrated Communication Database and Computational Tools."

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[2] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE, 2015.

[4] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: development and applications to human computer interaction. In *2003 Conference on computer vision and pattern recognition workshop*, volume 5, pages 53–53. IEEE, 2003.

[5] D Vaughn Becker, Douglas T Kenrick, Steven L Neuberg, KC Blackwell, and Dylan M Smith. The confounded nature of angry men and happy women. *Journal of personality and social psychology*, 92(2):179, 2007.

[6] Eyal Beigman and Beata Beigman Klebanov. Learning with annotation noise. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 280–287, 2009.

[7] Christopher Blank, Shaila Zaman, Amanveer Wesley, Panagiotis Tsiamyrtzis, Dennis R Da Cunha Silva, Ricardo Gutierrez-Osuna, Gloria Mark, and Ioannis Pavlidis. Emotional footprints of email interruptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[10] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[11] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[13] Jeffrey F Cohn, Adena J Zlochower, James Lien, and Takeo Kanade. Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, 36(1):35–43, 1999.

[14] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[15] Zijun Cui, Yong Zhang, and Qiang Ji. Label error correction and generation through label relationships. In *AAAI*, pages 3693–3700, 2020.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019.

[18] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 2020.

[19] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[20] Cynthia Dwork and Christina Ilvento. Fairness under composition. *arXiv preprint arXiv:1806.06122*, 2018.

[21] Paul Ekman and Wallace Friesen. Action coding system: a technique for the measurement of facial movement. *Consulting Psychologists Press, Palo Alto*, 1978.

[22] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[23] Hachim El Khiyari and Harry Wechsler. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. *Journal of Biometrics and Biostatistics*, 7(323):11, 2016.

[24] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016.

[25] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[26] Patrick J Grother, Mei L Ngan, and Kayee K Hanaoka. Face recognition vendor test part 3: Demographic effects. 2019.

[27] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[28] Drew Harwell. Rights group files federal complaint against ai-hiring firm hirevue, citing 'unfair and deceptive' practices, Nov 2019.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018.

[31] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior*, 21(4):241–257, 1997.

[32] Ayanna Howard, Cha Zhang, and Eric Horvitz. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–7. IEEE, 2017.

[33] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

[34] Vasileios Iosifidis and Eirini Ntoutsi. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24, 2018.

[35] Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Right for the right reason: Training agnostic networks. In *International Symposium on Intelligent Data Analysis*, pages 164–174. Springer, 2018.

[36] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[37] Jungseock Joo, Erik P Bucy, and Claudia Seidel. Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, 2019.

[38] Jungseock Joo and Kimmo Kärkkäinen. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. *arXiv preprint arXiv:2005.10430*, 2020.

[39] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

[40] Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pages 1–6. Citeseer, 2010.

[41] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[42] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[43] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 46–53. IEEE, 2000.

[44] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021.

[45] Jihen Khalfallah and Jaleleddine Ben Hadj Slama. Facial expression recognition for intelligent tutoring systems in remote laboratories platform. *Procedia Computer Science*, 73:274–281, 2015.

[46] Janet E Kilbride and Matthew Yarczower. Ethnic bias in the recognition of facial expressions. *Journal of Nonverbal Behavior*, 8(1):27–41, 1983.

[47] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[48] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[49] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076, 2017.

[50] Peter Lewinski, Tim M den Uyl, and Crystal Butler. Automated facial coding: validation of basic emotions and facs aus in facereader. *Journal of Neuroscience, Psychology, and Economics*, 7(4):227, 2014.

[51] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[52] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 2584–2593. IEEE, 2017.

[53] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer, 2020.

[54] Daniel Lundqvist, Anders Flykt, and Arne Öhman. The karolinska directed emotional faces (kdef). *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, 91(630):2–2, 1998.

[55] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[56] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

[57] David Matsumoto and Paul Ekman. Facial expression analysis. *Scholarpedia*, 3(5):4237, 2008.

[58] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[59] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2017.

[60] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[61] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. Pytorch metric learning, 2020.

[62] Mkhuseli Ngxande, Jules-Raymond Tapamo, and Michael Burke. Bias remediation in driver drowsiness detection systems using generative adversarial networks. *IEEE Access*, 8:55592–55601, 2020.

[63] Mahbod Olfat and Anil Aswani. Spectral algorithms for computing fair support vector machines. In *International Conference on Artificial Intelligence and Statistics*, pages 1933–1942, 2018.

[64] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[65] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568, 2008.

[66] Yilang Peng. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5):920–941, 2018.

[67] Douglas Perry. Emotion-recognition technology doesn't work, but hiring professionals, others are using it anyway: report, Dec 2019.

[68] Diah Anggraeni Pitaloka, Ajeng Wulandari, T Basaruddin, and Dewi Yanti Liliana. Enhancing cnn with preprocessing stage in automatic emotion recognition. *Procedia computer science*, 116:523–529, 2017.

[69] Edward Raff and Jared Sylvester. Gradient reversal against discrimination. *arXiv preprint arXiv:1807.00392*, 2018.

[70] Lauren Rhue. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*, 2018.

[71] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

[72] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[73] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[74] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in neural information processing systems*, pages 1857–1865, 2016.

[75] John Eric Steephen, Samyak Raj Mehta, and Raju Surampudi Bapi. Do we expect women to look happier than they are? a test of gender-dependent perceptual correction. *Perception*, 47(2):232–235, 2018.

[76] Yingli Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression recognition. In *Handbook of face recognition*, pages 487–519. Springer, 2011.

[77] Ying-li Tian, T Kanada, and Jeffrey F Cohn. Recognizing upper face action units for facial expression analysis. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 294–301. IEEE, 2000.

[78] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[79] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[80] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[81] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[82] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019.

[83] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019.

[84] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[85] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8919–8928, 2020.

[86] Nan Xi, Di Ma, Marcus Liou, Zachary C Steinert-Threlkeld, Jason Anastasopoulos, and Jungseock Joo. Understanding the political ideology of legislators from social media im-

ages. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 726–737, 2020.

[87] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. *arXiv preprint arXiv:2007.10075*, 2020.

[88] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[89] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[90] Minda Zetlin. Ai is now analyzing candidates' facial expressions during video job interviews, Feb 2018.

[91] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[92] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015.

[93] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[94] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[95] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27:487–495, 2014.

[96] Honglei Zhuang and Joel Young. Leveraging in-batch annotation bias for crowdsourced active learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 243–252, 2015.

[97] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.