Enhancing Fairness in Face Detection in Computer Vision Systems by Demographic Bias Mitigation

Yu Yang*
Amazon Alexa AI
yuyang@cs.ucla.edu

Aayush Gupta
Amazon Alexa AI
aayugupt@amazon.com

Jianwei Feng
Amazon Alexa AI
jianwef@amazon.com

Prateek Singhal
Amazon Alexa AI
prtksngh@amazon.com

Vivek Yadav
Amazon Alexa AI
ydvivek@amazon.com

Yue Wu
Amazon Alexa AI
wuayue@amazon.com

Pradeep Natarajan
Amazon Alexa AI
natarap@amazon.com

Varsha Hedau
Amazon Alexa AI
hedauv@amazon.com

Jungseock Joo
Amazon Alexa AI
jjoo@comm.ucla.edu

ABSTRACT

Fairness has become an important agenda in computer vision and artificial intelligence. Recent studies have shown that many computer vision models and datasets exhibit demographic biases and proposed mitigation strategies. These works attempt to address accuracy disparity, spurious correlations, or unbalanced representations in datasets in tasks such as face recognition, verification and expression and attribute classification. These tasks, however, all require face detection as the first preprocessing step, and surprisingly, there has been little effort in identifying or mitigating biases in face detection. Biased face detectors themselves pose a threat against fair and ethical AI systems, and their biases may be further passed on to subsequent downstream tasks such as face recognition in a computer vision pipeline. This paper therefore investigates the problem of biases in face detection, focusing on accuracy disparity of detectors between demographic groups including gender, age group, and skin tone. We collect perceived demographic attributes on a popular face detection benchmark dataset, WIDER FACE, report skewed demographic distributions, and compare detection performance between groups. In order to mitigate the biases, we apply three mitigation methods that have been introduced in the recent literature and also propose two novel methods. Experimental results show that these methods are effective in reducing demographic biases. We also discuss how the effectiveness varies by demographic attributes, detection easiness, and multiple detectors, which will shed light on this new topic of addressing face detection bias.

*Yu was a summer intern at Amazon while performing this work. She is currently a PhD student at UCLA.

AIES '22, August 1–3, 2022, Oxford, United Kingdom.

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534153

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; *Regularization*; *Cross-validation*; *Object detection.*

KEYWORDS

Fairness in computer vision, Face detection bias, Bias in large-scale facial image dataset

ACM Reference Format:
Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. 2022. Enhancing Fairness in Face Detection in Computer Vision Systems by Demographic Bias Mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3514094.3534153

1 INTRODUCTION

Fairness has become an important agenda in computer vision and artificial intelligence research. Recent studies have shown that many computer vision models and datasets exhibit demographic biases [6, 24] and also proposed mitigating strategies [10, 43]. Many real-world AI systems such as self-driving cars or personal assistance devices are getting equipped with visual sensors and inference modules based on computer vision models. Biases in underlying machine learning models can lead to unexpected negative outcomes such as discriminating certain groups of people in employment [35] or education [33].

Previous works on fairness in computer vision have attempted to address accuracy disparity, spurious correlations, or unbalanced representations in datasets in various automated tasks using facial images such as face recognition, verification and expression and attribute classification [1, 6, 9, 24, 39]. A face contains diverse cues from which both humans and models may infer various information about the person, and there have been numerous applications and datasets for automated facial analysis. The first major step in these automated systems is face detection [42] – to find and localize any faces in a given image. This is followed by subsequent processing for each detected face to generate final model outcomes such as identity, demographic attributes, or expression (Fig 1). Surprisingly, there has been little effort in identifying or mitigating biases in face detection. Existing studies have focused on biases in downstream

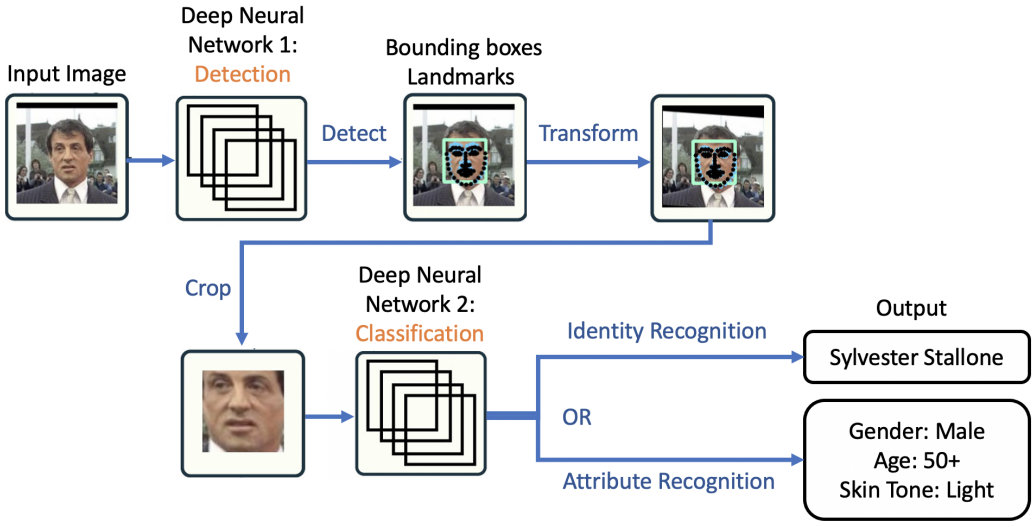

Figure 1: A common processing pipeline for automated face analysis contains multiple steps and models. The first step in these systems is face detection. Any hidden biases in the detection module can lead to biased sample selection in training datasets for subsequent modules (i.e. classification) as well as the accuracy disparity of the whole system between demographic groups. Previous studies on biases in computer vision facial models have only examined the biases of the second-stage classifier in isolation. Our paper is aimed at investigating the biases of the first module, face detection, which will enable, along with previous findings, a comprehensive approach in addressing pipeline biases.

tasks, assuming that there is no systematic biases in its preprocessing steps including the detection stage. Biased face detectors themselves pose a threat against fair and ethical AI systems, and their biases may be further passed on to subsequent downstream tasks in a computer vision system pipeline.

This paper therefore investigates the problem of biases in face detection, focusing on accuracy disparity of detectors between demopgrahic groups including gender, age group, and skin tone. We collect perceived demographic attributes on a popular face detection benchmark dataset, WIDER FACE, report skewed demographic distributions, and compare detection performance between groups. In order to mitigate the biases, we apply three different mitigation methods that have been introduced in the recent literature and also propose two novel methods. Experimental results show that these methods are effective in reducing demographic biases. We also discuss how the effectiveness varies by demographic attributes, detection easiness, and multiple detectors, which will shed light on this new topic of addressing face detection bias. Our main contributions are as follows:

- We construct a novel dataset of demographic labels on a public face detection benchmark dataset (WIDER FACE).
- We show that existing face detectors exhibit demographic biases measured by performance disparity between groups (equalized odds), which can further affect downstream tasks in a pipeline.
- By using existing bias mitigation methods, we show that we can significantly reduce the detection bias. We also propose two novel mitigation methods, which are computationally efficient and effective in bias mitigation. Previous studies on biases in facial models have only examined the biases of the second-stage classifier in isolation.

2 RELATED WORK

Fairness and Biases in Computer Vision: While fairness and demographic biases are relatively new topics in computer vision, earlier works examined the issue of dataset bias in large scale image datasets, i.e. how well a model trained from one dataset will perform on another dataset [16, 26, 41], which is also related to domain adaptation and transfer [40]. [29, 30] introduced and discussed the notion of representation bias, which refers to biases that occur in a hierarchy of representations in which a task that requires high level representation can be solved by only using low level representation, e.g. the action of "playing the piano" can be recognized by just seeing the piano in one static frame. Similar observations have also been made in visual question answering [7, 22] and visual reasoning [31]. All of these works address biases in computer vision, mostly concerning with limited or idiosyncratic ways in data collection and their impacts on model generalizability, but these biases are not specifically related to social or demographic biases.

Another line of work has focused on demographic biases, i.e. biased datasets or models in the context of social groups and discrimination, and investigated whether and how a model treats specific individuals or groups, especially socially less-privileged groups, in an undesirable or unfair way. Klare et al. [27] reported that various face recognition methods consistently show lower matching performance on females, Blacks, and age group 18–30 than other groups. Buolamwini and Gebru [6] also reported that the gender classification performance of commercial computer vision APIs

was consistently lower for dark-skinned females than light-skinned males. Similar demographic biases have been reported for a range of tasks such as image classification [15, 23, 47], face attribute classification [10, 13, 24], expression recognition [9, 46], face recognition and verification [18, 39], visual semantic role labeling [51], image captioning [20],

Visual Bias Measurement and Mitigation: There has been a great deal of effort to quantify biases and mitigate them in computer vision datasets and models. Traditional datasets are often biased as they disproportionately represent demographic groups, e.g. White-dominated, which can lead to the performance disparity of the trained models. A number of new datasets with balanced representation on gender, race, and age-groups have been proposed to allow accurate measurement and mitigation of such bias [19, 24, 39]. Collecting real images is effective but expensive and it is more difficult to collect data from underrepresented groups. Therefore, researchers have also proposed to use synthetic images to fill the gap between groups. These synthetic or manipulated images can also serve as counterfactual examples to measure counterfactual or causal fairness [3, 13, 23], i.e. the exact impact of demographic cues on the model bias.

Algorithmic approaches have also been proposed to mitigate model bias given a fixed biased dataset. A straightforward mitigation strategy is to balance samples from a dataset, e.g. adjusting sampling frequency or weight for each example based on the proportion of its group in the dataset [8]. Another approach is fairness through blindness [1] in which protected attributes are omitted in the input and feature space such that trained models can't use the protected attributes in predictions. The omission of an input variable doesn't always prevent the model from using it because the signal can still be inferred from other proxy variables. Adversarial learning has been used to remove cues related to protected attributes in the feature representation of a model [1, 43, 44]. Another line of work have used generated models to synthesize new images or modify existing ones which can be used to either augment unbalanced training dataset [38] or to measure model's sensitivity to protected attributes [3, 13, 23]. Recent methods have also considered the fairness issue in computer vision methods in connection to other data type (e.g. text) for multimodal analysis such as image captioning [4, 20, 50].

To the best of our knowledge, there has been no work about measuring or mitigating demographic biases in face detection.[1] The closest work to our paper are biases in pedestrian detection [5, 45] and facial landmark detection [28], but these domains are still very different from face detection in terms of methods, datasets, and applications. Two recent studies have measured demographic biases in face detection [14, 24], but they both used datasets in which the faces were automatically selected by an existing face detector, which is already biased, and thus these samples (which tend to be much easier to detect than challenging examples in our dataset – WiderFace) are not appropriate for the purpose of auditing a detector. In addition, these papers do not propose any algorithmic solution to mitigate biases in face detection. The lack of in-depth analysis on face detection bias poses a serious risk to a large number of computer vision systems that use face detection as the first preprocessing step. That is, while there have been **numerous previous studies on computer vision fairness**, most of these studies potentially bear the same critical limitation in that they overlook the fact that the very first computational module of these systems may produce and transmit biases to the entire systems. Our paper is aimed at investigating this critical issue for the first time by quantitatively measuring biases in face detection and applying popular mitigation techniques.

[1]The term of "face detection" is sometimes misused in the literature. For example, [2] studies biases in face classification without localization but uses the term of detection. Face detection refers to the task of finding and localizing face instances in an image. See Section 4.1 for the formal task definition.

3 A NEW BENCHMARK FOR FAIR FACE DETECTION: WIDERFACE-DEMO

There has been no dataset that can be used to measure demographic biases of face detectors. Face datasets with demographic labels have been developed for other tasks such as face recognition [34] but they are typically constructed semi-automatically by first running a face detector to obtain candidate images and can't be used for our purpose because we are interested in what would be missed by the detector. Instead of constructing a new dataset, we augment an existing popular benchmark dataset, WIDER FACE [48], with demographic labels such that current and future bias analysis can be aligned with state-of-the-art face detection methods. WIDER FACE contains 393,703 ground-truth face bounding boxes from 32,203 images. Since the bounding boxes in the test set (50% of the dataset) are not publicly released, we only use the train set (40%) and the validation set (10%).

We obtained perceived demographic attributes – gender, skin tone, and age group – by a commercial annotation service. We note that people's true demographic attributes may not be perfectly inferred from their facial images and therefore explicitly define these as "perceived" attributes as shown in Figure 3. We use a demographic categorization scheme commonly used in the literature on fairness [3, 24]. Each face image was given to three annotators and we took the majority responses as the assigned groups. Images that received more than one "not sure" response were excluded for the corresponding attribute. These demographic attributes are auxiliary variables in our main task of face detection, which will be used to compute specific bias-related loss functions and evaluate accuracy disparity between groups. Therefore, we annotated a subset of faces in the dataset (12,000 in the validation set and 25,000 in the training set) to reduce cost.

WIDER FACE is imbalanced in terms of the demographic attributes we annotated. To increase the diversity of examples in the annotated subset in the training set, we first used a publicly available face attribute model (FairFace [24]) to obtain demographic attribute predictions and incrementally increased the dataset size (the White-male category was quickly filled up and excluded in the process). We did not use the protocol to sample from the validation set because the set is small. Figure 2 shows the distribution of the demographic attributes in the entire WiderFace-Demo set.

4 BIAS MEASUREMENT AND MITIGATION IN FACE DETECTION

In this section, we formulate our problem by introducing the face detection task and the definition of fairness used for the task. We then

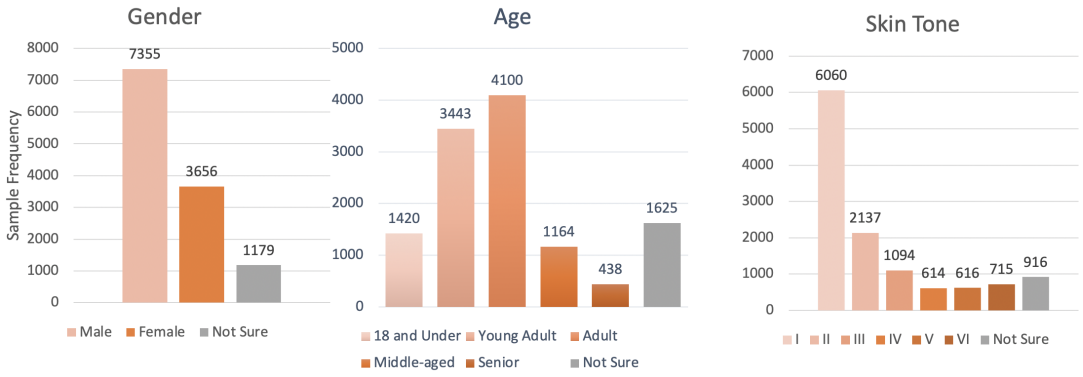

Figure 2: The demographic distribution of faces in Wider Face (random samples from the validation set).

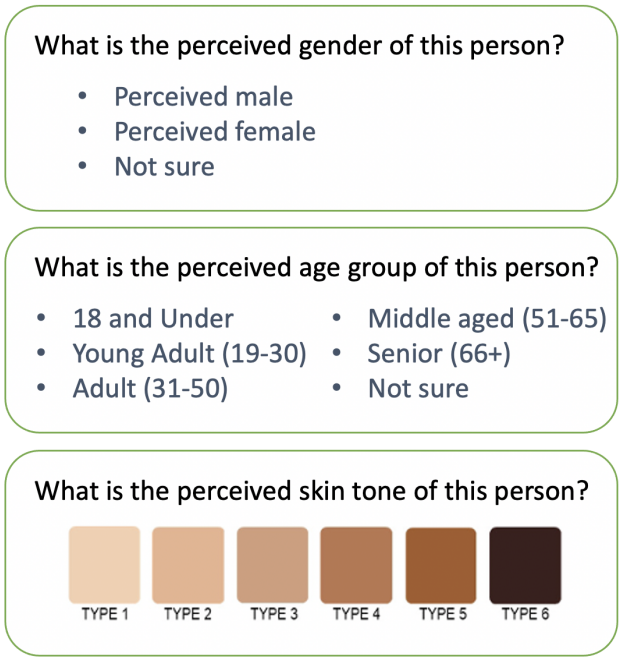

Figure 3: The annotation questionnaires and categories.

introduce existing bias mitigation approaches and make adaptation to them for the face detection task. Finally, we propose two novel bias mitigation methods **Attribute-Orthogonal Detection** and **Attribute-Attended Learning** which extend previous approaches.

4.1 Problem Formulation

4.1.1 Face Detection. The goal of face detection is to find every instance of face in an input image. In computer vision, the term of detection refers to localization, i.e. predicting the location and size of an object rather than just classifying its type. The location and size of an object instance is typically specified by a bounding box, and there can be multiple bounding boxes in one image. The input is an image I and the outputs are a set of face bounding boxes with confidence $O = \{t_i = (x_i, y_i, w_i, h_i, p_i)\}$. For the ith bounding box (t_i), x_i, y_i are its left top coordinates, and w_i, h_i are its width and height. p_i is the confidence of the ith bounding box being a face.

Currently, the most popular approach in face detection is to use convolutional neural networks [11, 21, 36]. These models typically contain a feature extractor network f and a light weighted detector d which detects face bounding boxes from feature maps, namely $O = d(f(I))$. During training, the objective for face detection is formed by a bounding box regression loss, L_{reg} and a classification loss, L_{cls}.

$$\min_{\theta_f, \theta_d} L = L_{reg} + L_{cls}, \tag{1}$$

where L_{cls} is typically a binary cross entropy (log loss) and L_{reg} is a smooth-L1 loss which penalizes the misalignment between the predicted bounding box and the ground truth bounding box [17].

4.1.2 Fairness Definition for Face Detection. The goal of our paper is to minimize the accuracy disparity of a face detector between demographic groups. The accuracy of a face detector is commonly measured by average precision (AP) by treating each bounding box as an example in a binary classification task. However, we use recall instead of AP as our main measure for the following reason. The face detection performance (i.e. AP) can be measured from precision and recall (AP is computed as the area-under-curve of a precision-recall curve).

$$pr = \frac{TP}{TP + FP}; rc = \frac{TP}{TP + FN} \tag{2}$$

where TP, FP, FN refers to true positive, false positive and false negative respectively. A true positive happens when a predicted bounding box matches with a ground truth bounding box (i.e. Intersection over Union, IoU above certain threshold). A false positive happens when a predicted bounding box doesn't match with any ground truth bounding box (i.e. wrong detection). A false negative happens when a ground truth bounding box is not detected (i.e. missed detection). Since our goal is to minimize the gap between demographic groups, every example needs to be associated with a demographic label. False positive bounding boxes do not have these labels, and thus do not affect the accuracy gap between groups in any meaningful way. We therefore select recall to measure face detection performance (i.e. true positive rate – **equal opportunity**)

and define the fairness goal by measuring standard deviation of recall between groups. As recall is also quite sensitive to threshold set for IoU, we apply the standard setting in face detection literature and calculate overall detection rate as the average recall in IoU range 0.5:0.05:0.95 (from 0.5 to 0.95 with step size 0.05).

Since WiderFace is imbalanced across different demographic groups (Fig. 2), we use the calibrated detection rate (recall) instead of the overall detection rate as our main metric. The calibrated detection rate is obtained by computing the average of the detection rates of all the demographic groups, i.e., $rc = \frac{1}{|D|}\sum_{k\in D} rc_k$, where D is the set of demographic groups and rc_k is the detection rate (recall) for each group. We then measure the standard deviation of the calibrated detection rate over different demographic groups as our bias metric, following the literature in fairness in computer vision [24].

4.2 Mitigating Face Detection Biases

In this section we describe existing methods commonly used for bias mitigation based on [44] and adapt them for face detection task.

4.2.1 Sample Weighting. The simplest approach in bias mitigation is to adjust weights for samples by the inverse of the probability of the inclusion of their corresponding demographic groups in the dataset. This will make each demographic group equally weighted in training [44].

For each ground truth face bounding box i, we assign a weight w_i:

$$w_i = \frac{min_{g,a,s} P_{tr}(g,a,s)}{P_{tr}(g_i,a_i,s_i)} \tag{3}$$

where g_i, a_i, s_i are the demographic labels (gender, age, skin tone) for bounding box i. $P_{tr}(g,a,s)$ is the distribution of demographic labels in the dataset and $\sum_{g\in G,a\in A,s\in S} P_{tr}(g,a,s) = 1$. The objective for face detection then becomes

$$\min_{\theta_f,\theta_d} L = \frac{1}{N}\sum_{i=1}^{N} w_i(L_{reg,i} + L_{cls,i}) \tag{4}$$

where N is the number of samples in the dataset, $L_{reg,i}$ and $L_{cls,i}$ are the regression and classification losses on sample i. For convenience, we refer to this method as "Weighted" in figures and tables.

4.2.2 Adversarial Training. We aim to learn a feature extractor f such that the feature embedding for each image is informative for the face detection task, d, whilst being uninformative for a demographic classification model c.

Uniform Confusion Loss. Following [44], we consider the uniform confusion loss [1].

$$L_{adv} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|D|}\sum_{k\in D} -\log c_k(f(t_i)) \tag{5}$$

where N is the number of samples in the dataset and $|D|$ is the number of classes of certain demographic attribute (e.g. 2 classes for gender). $f(t_i)$ is the feature representation for a bounding box prediction, t_i. $c_k()$ is the output of the demographic attribute classifier for the k-th group. This loss function is minimized when the classifier, c, produces even outputs across all the categories, i.e., the feature representation, f, does not contain any information about protected attributes.

The overall objective for face detection becomes

$$\min_{\theta_f,\theta_d,\theta_c} L = L_{cls} + \lambda \cdot L_{reg} + \alpha \cdot L_{adv} \tag{6}$$

where α is the weight for adversarial loss. The weight for the regression loss, λ is given from the original paper of RetinaFace [44]. We refer to this method as "Adv-Confusion" for adversarial uniform confusion loss.

Gradient Reversal. We also consider the gradient reversal technique proposed in [37]. The adversarial loss for gradient reversal is a standard classification loss:

$$L_{adv} = \frac{1}{N}\sum_{i=1}^{N}\log c_{k_i}(f(t_i)), \tag{7}$$

where k_i is the demographic category label for the face bounding box, t_i.

Gradient Reversal training is standard adversarial training where the feature extractor tries to generate features that confuse the demographic classifier while the demographic classifier tries to classify the features into different demographic groups. It encourages the feature extractor to generate features that contain minimum demographic information. The training procedure is as follows.

Algorithm 1 Gradient Reversal Training

```
for epochs do
    for mini-batches do
        Optimize demographic classifier. min_{θ_c} L_adv
        Optimize feature extractor and detector with reversed gradients from classifier.
        min_{θ_fe,θ_det} L_reg + L_cls − α ∗ L_adv
    end for
end for
```

4.3 Attribute-Orthogonal Detection

The existing methods for bias mitigation have a few limitations. First, Sample Reweighting usually struggles when the dataset does not contain sufficient samples for a specific demographic group. Second, adversarial training attempts to achieve fairness by enforcing a model to not learn features that are useful for demographic group classification. However, such features may still be useful for detection. For example, eye-glasses may be correlated with age group but can help face detection. Adversarial training will force to not learn these features. Recent studies have also shown that removing spurious features can decrease the model accuracy [25].

Therefore, we propose an alternative learning algorithm called Attribute-Orthogonal Detection that can still encourage to learn useful features but decorrelate features for detection and demographic attribute classification. This is done by simply adding a regularization term that penalizes the correlation between the parameters of detector θ_{det} and the parameters of the demographic attribute classifier θ_{cls} which share the same features from feature

extractor θ_{fe}.

$$L_{ortho} = \frac{||\theta_{det} \cdot \theta_{cls}||_1}{||\theta_{det}||_2 \cdot ||\theta_{cls}||_2} \tag{8}$$

This regularization term allows us to encourage the detector and the attribute classifier independent in the feature space while not completely prohibiting the overlap between them. The optimization is also simpler than adversarial training, which requires to solve a min-max optimization problem. The objective for face detection then becomes

$$\min_{\theta_{fe},\theta_{det},\theta_{cls}} L_{reg} + L_{cls} + \alpha * L_{ortho} \tag{9}$$

4.4 Attribute-Attended Learning

We propose another mitigation method that can learn effective features for diverse demographic groups using attention. This method is related to domain independent training proposed by [44], where a model is trained to treat each demographic category (domain) separately. For example, separate detectors can be trained to detect male faces and female faces in this case (or their feature networks may be shared). While such a method can effectively separate distinct classes, we also note that boundaries between demographic categories are not very clear and many face images may fall into areas between classes (e.g., skintone type 3 vs. 4).

To effectively model this continuous space, we use an attention mechanism as shown in Fig. 4. The core utility of the attention in this method is to let the model attend to different features for the detection task based on its beliefs about demographic attributes. Specifically, the model first computes the feature for a bounding box t_i, $f(t_i)$, and this is used to predict demographic attributes, $c(f(t_i))$. This prediction is transformed into the feature attention vector, a_i, of the same dimension as $f(t_i)$. Finally, $f(t_i)$ is **reweighted** by a_i by an element-wise multiplication before it is passed to the detector, d. Therefore, a_i controls how much each feature in $f(t_i)$ should contribute to the final detection output.

5 EXPERIMENTS

5.1 Measuring Face Detection Bias

We use three widely-used face detectors in the experiments. RetinaFace [12] is a popular single stage face detector using a feature pyramid and context modules. We adopt two variants of RetinaFace with different backbone networks: ResNet-50 and MobileNet-0.25. The former has a better accuracy and the latter is more efficient. MobileNet is a very popular light-weight model, frequently used in mobile devices, which may have more direct impacts on ordinary people. We also use MTCNN [49] to evaluate face detection bias effect on downstream task. In all cases, we use the model implementation provided by the authors and follow the same training protocols.[2]

We first measured the biases of these existing detectors for different demographic groups. As mentioned before, we define our fairness metric as the standard deviation of calibrated detection rate over the range of IoU thresholds in 0.5:0.05:0.95. We found a considerable amount of biases from the given range of IoU thresholds as shown in Figure 6. Models with lower detection performance tend to show larger biases than more accurate models. RetinaFace with ResNet-50 shows a lower bias between skin-tone groups than RetinaFace with the MobileNet backbone while having similar biases for gender and age group.

Fig. 7 also shows that the detector yields lower detection performance for the darker skintone groups than the lighter groups. This pattern is also consistent with the relative size of the groups in the training dataset. The same pattern was found for the case of age group with an exception that the 18 and Under group yields the best detection performance while they represent a smaller subset in the training dataset compared to other groups. We believe this is because the people in this age group tend to appear in the center of images, many of which are clean portraits targeting them.

5.2 Mitigating Face Detection Bias

We report the effectiveness of mitigation methods introduced in Section 4.2 using the same dataset. Ideally, an effective mitigation method should be able to minimize the bias in each demographic group while maintaining higher calibrated detection rates. Table. 1 and Table. 2 summarize the calibrated detection rates and bias for different mitigation methods. Fig. 5 also visualizes the trade-off between overall detection performance and inter-group bias. For mitigation methods "Adv-Reverse", "Adv-Confusion", "Ortho", "Attend", we train several models variants with different hyper-parameters and visualize several data points in Fig. 5 . Generally, different mitigation methods show similar calibrated detection rate compared to baseline. Among them, "weighted" (simply reweighting samples) and "Adv-Reverse" show limited bias reduction compared to baseline, while other methods generally shows certain amount of reduction in bias. On smaller model (MobileNet-0.25), Our proposed methods "Ortho" demonstrates the best bias reduction by reducing 0.15% in overall bias measured by standard deviation between groups (10.2% relatively compared to baseline). Our proposed methods "Attend" and "Adv-Confusion" (adversarial confusion loss) also show decent amount of reduction in bias by reducing 0.07% (4.8% relatively) and 0.08% (5.4% relatively). On larger model (Resnet-50), we observe a similar trend, where our proposed methods "Ortho" and "Attend" and "Adv-Confusion" show most improvement in bias reduction while maintaining similar calibrated detection rate. Note that although "Adv-Reversal" reduces significant bias 0.18%(15.1% relatively), it also has a significant drop in calibrated detection rate by 1.1%. From Fig. 5, we also observe similar trend that "Adv-Confusion", "Ortho" and "Attend" have more improvement on bias. These results (the difference between the baseline and other methods) are statistically significant and all the mitigation methods significantly improve the bias of the baseline model on at least one demographic category (p-val < 0.001) with an exception of the age group for the MobileNet model. This was tested by comparing the detection performance between methods for the group which shows the lowest performance via McNemar's Test.

5.3 Estimating Impact of Face Detector Bias in Downstream Tasks in CV Pipelines

We also discuss the potential impact of the bias of a face detector on downstream applications which take detected bounding boxes as an

[2]As part of our work, we did not collect any face landmarks or biometric identifiers.

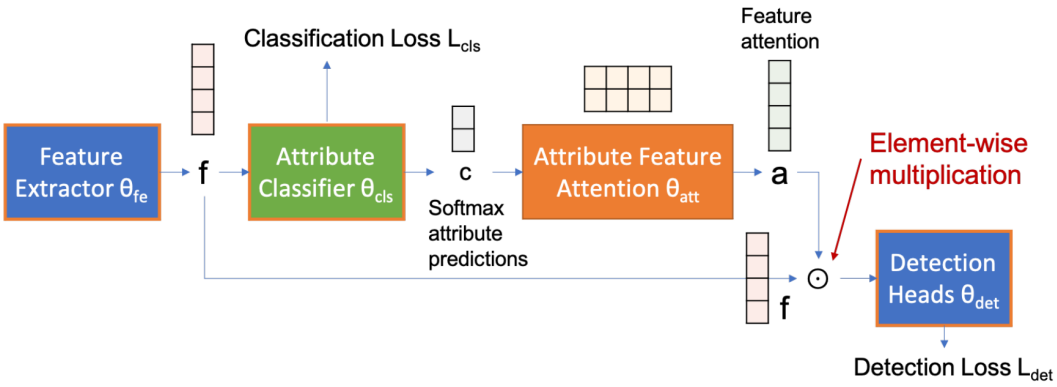

Figure 4: Our attribute feature attention pipeline.

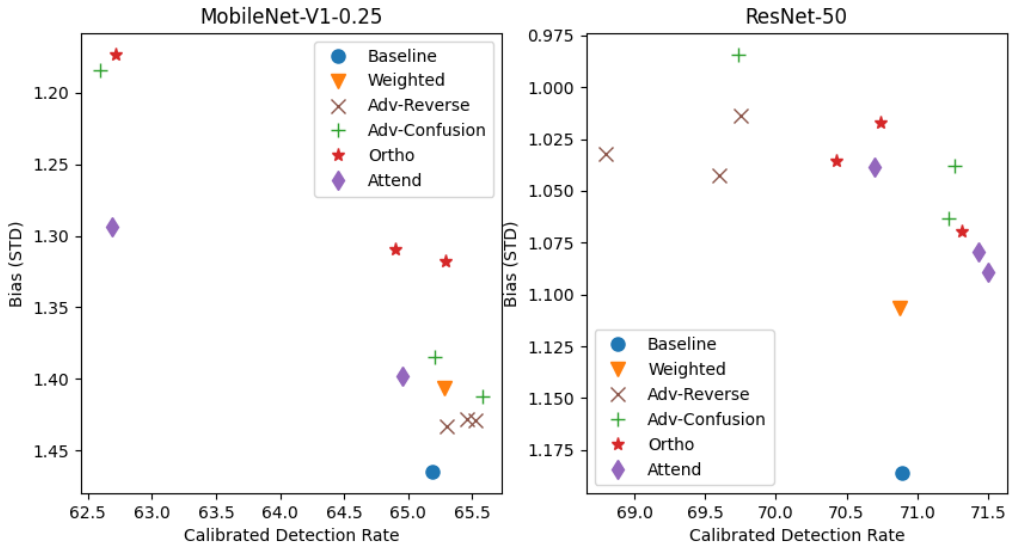

Figure 5: Trade-off between overall detection performance and inter-group bias. Multiple points were obtained by varying the hyperparameter $\alpha \in \{0.1, 1, 10.0\}$

input to the systems. For example, a security application using face verification will first detect and localize a face in an input image and then verify the identity of the face. If a face is not detected at all, it will be impossible for the system to further process the face. Thus, any biases in a face detector in a computer vision pipeline will be accumulated in the overall bias of the system.

To measure the exact degree of the impact, however, is challenging because most face image datasets for face analysis, except the ones developed for face detection (e.g. Wider-Face), have been created by using face detectors, which we have shown are biased. That is, most faces in these datasets are detectable even by a biased face detector.

To understand the impact of bias accumulation in a computer vision pipeline, we instead measure the changes in classification accuracy for a downstream classification task due to the bounding box alignment error of a detector. A face detector performs a binary classification task (whether there is a face or not in a given location) and a prediction task for precise alignment for a bounding box. Any biases in a face detector can therefore have two types (face and bounding box) of impacts on subsequent modules. We are interested in measuring the second impact, i.e. a potential bias that can be caused due to the alignment error.

To this end, we use CelebA [32], a public face attribute classification benchmark dataset, to measure how the classification accuracy is affected by the alignment error. The CelebA dataset provides facial images which have been pre-detected and aligned such that a classifier can be trained and evaluated with these pre-aligned images. In a real-world computer vision system, however, any faces should be first detected and aligned using a detector. This detector will produce imperfect detection bounding boxes, whose errors

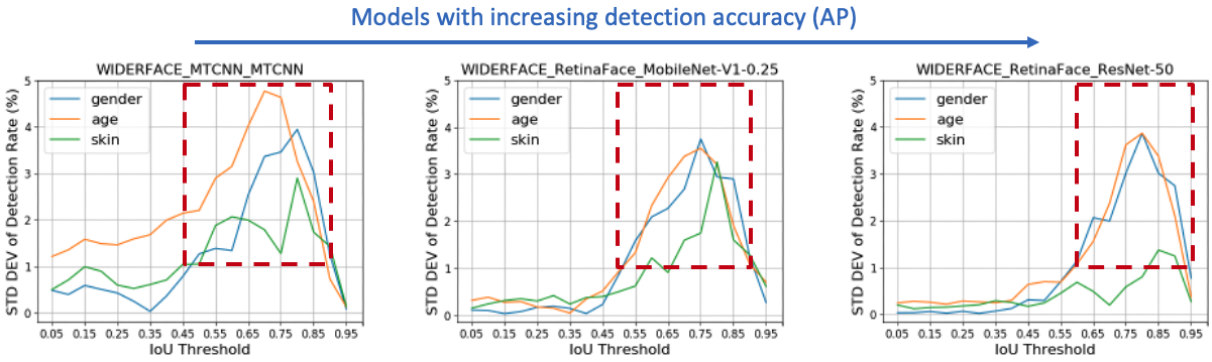

Figure 6: Demographic biases in face detectors measured at different IoU thresholds.

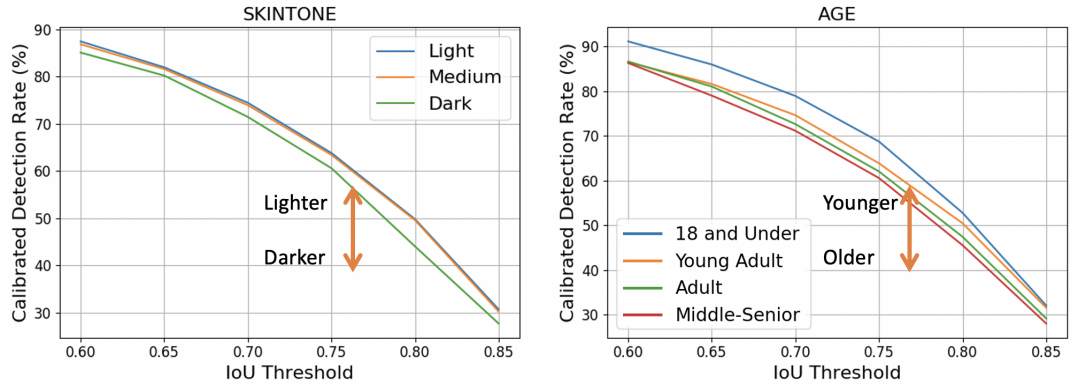

Figure 7: Performance disparity of a baseline detector (RetinaFace with Mobilenet-v1).

Table 1: Calibrated Detection Rates (Recall) averaged over IoU thresholds in the range of (0.5:0.05:0.95)

RetinaFace-MobileNet-0.25						
Attribute	Calibrated Detection Rates (%) ↑					
	Baseline	Weighted	Adv-Reverse	Adv-Confusion	Ortho	Attend
Overall	65.19	65.28	65.46	65.20	65.29	64.95
Gender	65.63	65.70	65.96	65.67	65.69	65.35
Age	65.19	65.18	65.46	65.21	65.35	64.88
SkinTone	64.75	64.98	64.95	64.73	64.83	64.62
RetinaFace-Resnet-50						
Attribute	Calibrated Detection Rates (%) ↑					
	Baseline	Weighted	Adv-Reversal	Adv-Confusion	Ortho	Attend
Overall	70.89	70.88	69.75	71.26	70.74	70.70
Gender	71.32	71.19	70.02	71.66	71.13	71.09
Age	70.87	70.86	69.73	71.26	70.77	70.74
SkinTone	70.48	70.58	69.51	70.87	70.33	70.26

are uneven across demographic groups as shown in the previous section. Therefore, we compare the performance gap between two cases – one using perfectly aligned bounding boxes, and the other using boxes produced by an automated detector (we used publicly available pre-trained detectors) – as a way to estimate the impact of detector errors on a subsequent classifier.

Figure 8 shows how much imperfect and biased detection can further increase the bias of face attribute classification. This result is also related to bias amplification where the bias of the dataset is amplified in the model [51]. In our case, we show that the bias of the detector can amplify the bias of subsequent downstream tasks. Note that we only consider the gender attribute in CelebA to

Table 2: Bias (standard deviation between groups) averaged over IoU thresholds in the range of (0.5:0.05:0.95). For overall bias comparison, we also show relative improvement over baseline. For example 1.32(10.2%) means relatively 10.2% improvement over baseline. Note that it is always possible to lower the model bias by compromising the detection performance (use higher α). To make meaningful comparisons, we chose the model whose detection performance is not lower than the baseline's performance by 0.5% (See Table 1). The only exception is Adv-Reversal for Resnet-50 whose best detection performance was 1.14% lower than the baseline.

RetinaFace-MobileNet-0.25						
Attribute	Biases (%) ↓					
	Baseline	Weighted	Adv-Reverse	Adv-Confusion	Ortho	Attend
Overall	1.47	1.41(4.1%)	1.43(2.7%)	1.39(5.4%)	**1.32(10.2%)**	1.40(4.8%)
Gender	1.55	1.54	1.49	1.39	**1.34**	1.44
Age	1.54	1.69	**1.48**	1.55	1.50	1.69
SkinTone	1.31	**0.98**	1.32	1.01	1.12	1.07
RetinaFace-Resnet-50						
Attribute	Biases (%) ↓					
	Baseline	Weighted	Adv-Reversal	Adv-Confusion	Ortho	Attend
Overall	1.19	1.11(6.7%)	**1.01(15.1%)**	1.04(12.6%)	1.02(14.3%)	1.04(12.6%)
Gender	1.08	1.01	**0.76**	0.97	1.00	0.94
Age	1.51	1.46	1.56	**1.23**	1.24	1.30
SkinTone	0.96	0.85	0.91	0.88	**0.81**	0.88

separate demographic groups and measure the classifier accuracy on the remaining 35 attributes because it does not have the same age and skintone categorization as our dataset. This inaccuracy originated from a detector will have an additional impact on top of the biases that already exist in the attribute classifier itself [6], i.e. **accumulated bias**. This suggests that it is critical to evaluate computer vision systems and pipelines holistically in addition to unit-level evaluations.

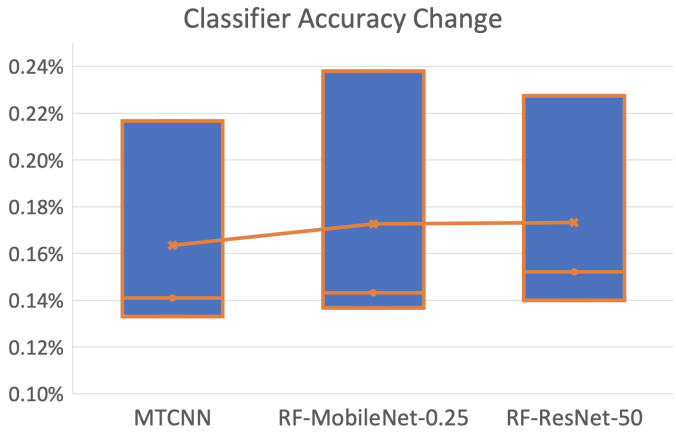

Figure 8: The impact of the precise bounding box alignment of a detector on the accuracy of a subsequent attribute classifier. The classifier becomes less accurate when using predicted bounding boxes than using annotated bounding boxes. The box indicates the minimum, median, and maximum performance change, The line plot indicates the mean of the changes. The differences are all significant.

6 CONCLUSION

Face detection is a widely used computer vision technique and a critical preprocessing step for other facial applications such as face recognition or attribute editing. While prior studies have attempted to address fairness and biases in various face-based computer vision models and datasets using an existing detector, little has been known about the existence and effects of biases in face detection itself. Since face detectors are commonly used as the required preprocessing step to find any faces in an image, any subsequent processing may be affected by the hidden biases of detectors.

In order to address this critical issue, we first created a novel dataset of demographic attributes of people using a popular face detection benchmark dataset, WiderFace. Using the labels, we showed that the dataset is significantly unbalanced and the existing face detectors trained from it also exhibit demographic biases. We then applied three different mitigation methods and proposed two novel mitigation methods and demonstrated that we can effectively reduce biases even with the biased training dataset.

REFERENCES

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2018. 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) Workshops*. 0–0.

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

[3] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. 2021. Towards Causal Benchmarking of Bias in Face Analysis Algorithms. In *Deep Learning-Based Face Analytics*. Springer, 327–359.

[4] Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578* (2019).

[5] Martim Brandao. 2019. Age and gender bias in pedestrian detection algorithms. *arXiv preprint arXiv:1906.10490* (2019).

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness,*

accountability and transparency. PMLR, 77–91.

[7] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems* 32 (2019), 841–852.

[8] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. 2020. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5671–5679.

[9] Yunliang Chen and Jungseock Joo. 2021. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14980–14991.

[10] Abhijit Das, Antitza Dantcheva, and Francois Bremond. 2018. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5203–5212.

[12] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. *CoRR* abs/1905.00641 (2019). arXiv:1905.00641 http://arxiv.org/abs/1905.00641

[13] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. 2019. Detecting bias with generative counterfactual face attribute augmentation. *arXiv e-prints* (2019), arXiv–1906.

[14] Samuel Dooley, George Z Wei, Tom Goldstein, and John P Dickerson. 2022. Are Commercial Face Detection Models as Biased as Academic Models? *arXiv preprint arXiv:2201.10047* (2022).

[15] Chris Dulhanty and Alexander Wong. 2019. Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. *arXiv preprint arXiv:1905.01347* (2019).

[16] Chen Fang, Ye Xu, and Daniel N Rockmore. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*. 1657–1664.

[17] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[18] Sixue Gong, Xiaoming Liu, and Anil K Jain. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision*. Springer, 330–347.

[19] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Casual Conversations: A Dataset for Measuring Fairness in AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2289–2293.

[20] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 771–787.

[21] Huaizu Jiang and Erik Learned-Miller. 2017. Face detection with the faster R-CNN. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 650–657.

[22] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.

[23] Jungseock Joo and Kimmo Kärkkäinen. 2020. Gender slopes: Counterfactual fairness for computer vision models by attribute manipulation. In *Proceedings of the 2nd International Workshop on Fairness, Accountability, Transparency and Ethics in Multimedia*. 1–5.

[24] Kimmo Karkkainen and Jungseock Joo. 2021. 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*. 1548–1558.

[25] Fereshte Khani and Percy Liang. 2021. Removing Spurious Features Can Hurt Accuracy and Affect Groups Disproportionately. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 196–205. https://doi.org/10.1145/3442188.3445883

[26] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*. Springer, 158–171.

[27] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801.

[28] Adam Kortylewski, Bernhard Egger, Andreas Morel-Forster, Andreas Schneider, Thomas Gerig, Clemens Blumer, Corius Reyneke, and Thomas Vetter. 2018. Can synthetic faces undo the damage of dataset bias to face recognition and facial landmark detection? *arXiv preprint arXiv:1811.08565* (2018).

[29] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 513–528.

[30] Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9572–9581.

[31] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4185–4194.

[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[33] Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 444–460.

[34] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. 2018. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*. IEEE, 158–165.

[35] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 28–29.

[36] Hongwei Qin, Junjie Yan, Xiu Li, and Xiaolin Hu. 2016. Joint training of cascaded CNN for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3456–3465.

[37] Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 189–198.

[38] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9301–9310.

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

[40] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. 2010. Adapting visual category models to new domains. In *European conference on computer vision*. Springer, 213–226.

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

[42] Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

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

[44] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[45] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[46] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*. Springer, 506–523.

[47] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 547–558.

[48] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[50] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and Evaluating Racial Biases in Image Captioning. *arXiv preprint arXiv:2106.08503* (2021).

[51] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.