Fair and accurate age prediction using distribution aware data curation and augmentation

Yushi Cao[1,a*], David Berend[1,a†], Palina Tolmach[1], Guy Amit[2], Moshe Levy[2]
Yang Liu[1], Asaf Shabtai[2], Yuval Elovici[2]
[1]Nanyang Technological University, [2]Ben-Gurion University of the Negev
{yushi002, bere0003, palina0001}@e.ntu.edu.sg, {guy5, moshe5}@post.bgu.ac.il
yangliu@ntu.edu.sg, {shabtaia, elovici}@bgu.ac.il

Abstract

Deep learning-based facial recognition systems have experienced increased media attention due to exhibiting unfair behavior. Large enterprises, such as IBM, shut down their facial recognition and age prediction systems as a consequence. Age prediction is an especially difficult application with the issue of fairness remaining an open research problem (e.g., predicting age for different ethnicity equally accurate). One of the main causes of unfair behavior in age prediction methods lies in the distribution and diversity of the training data. In this work, we present two novel approaches for dataset curation and data augmentation in order to increase fairness through balanced feature curation and increase diversity through distribution aware augmentation. To achieve this, we introduce out-of-distribution detection to the facial recognition domain which is used to select the data most relevant to the deep neural network's (DNN) task when balancing the data among age, ethnicity, and gender. Our approach shows promising results. Our best-trained DNN model outperformed all academic and industrial baselines in terms of fairness by up to 4.92 times and also enhanced the DNN's ability to generalize outperforming Amazon AWS and Microsoft Azure public cloud systems by 31.88% and 10.95%, respectively.

1. Introduction

Due to their high accuracy, deep learning-based facial recognition systems have been widely adopted by governments and industry, with $3.8 billion in governmental spending on such systems in 2020 [43]. Despite this government investment and their accuracy and rate of adoption, there have been numerous cases in which deep learning (DL) systems have been shown to behave unfairly. More specifically,

---

*Yushi Cao and David Berend cobntributed equally.
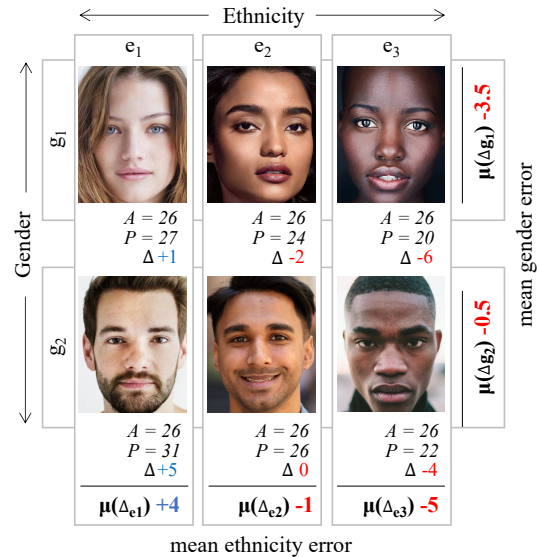†David Berend is the corresponding author.

Figure 1: Comparison of actual (A) and predicted age (P). Illustrates existence of systematic unequal prediction among gender and ethnicity which can be avoided.

in some cases, DL systems have demonstrated poor accuracy for a specific ethnicity or gender, e.g., male, while demonstrating high accuracy for other feature representations, e.g., female [9, 5, 20]. This inconsistent accuracy has led large enterprises, such as IBM, to eliminate the use of facial recognition systems [33]. Amazon Web Services (AWS) and Microsoft Azure have followed this trend, limiting the use of their facial recognition systems to governments and more specifically, to law enforcement agencies [5, 45]. However, the development of facial recognition systems continues and their adoption is increasing [43]. Therefore, there is a need to ensure the fairness of existing facial recognition systems and those developed in the future.

Age prediction is considered one of the most challenging facial recognition system tasks. The difficulty stems

from the high number of predictable classes representing ages, among which there are small apparent physical differences [15]. Age prediction systems have also been widely adopted, with applications in law enforcement [4], surveillance [7, 21], marketing [7, 3], and identity acquisition [4, 7]. These systems are considered unfair when the mean of the predicted age is different for, e.g., male and female [44].

To provide an example of unfair DL age prediction systems and their devastating impact, we present a law enforcement use case. In this setting, age is a factor, as younger people are more likely to commit a crime than older people. Law enforcement agencies have increased their use of DL-based age prediction systems [59]. These systems may be inherently unfair and consistently predict that African Americans are younger than Caucasians, and as a result, African Americans may be unfairly punished by law enforcement agencies solely due to their ethnicity. In fact, just recently news has reported that COMPAS, a tool used in many US jurisdictions to help make decisions about pretrial release and sentencing, was twice as likely to favor arrests for the African American ethnicity than for Caucasian ethnicity [26, 48].

Age prediction belongs to the field of computer vision (CV) where the issue of fairness is a challenging task due to the difficulty in creating diverse and sensitive feature balanced datasets. Most features and camera settings are embedded into the image as a combination of pixels, rather than labels. One simple solution is to combine all available data from existing benchmarks into one data pool. However, while this large data pool may be more diverse, it can still be imbalanced in its distribution of features and inherit the bias of an individual dataset. This is a common challenge faced by large cloud systems [34, 10] and by academia, reflected by the imbalanced distribution behavior in academic age prediction benchmark datasets used in prior studies [55, 16]. These benchmark datasets show fundamentally different distribution of features and camera settings which makes comparison of age prediction approaches using different benchmark datasets an infeasible task.

To address the challenge of curating diverse and balanced datasets, we propose a novel dataset curation approach that increases diversity by maximizing the balance between sensitive features. In addition, we introduce distribution awareness [27] to the age prediction domain. Distribution awareness provides a high-dimensional certainty estimate from the neural network for each input. Therefore, we aim to improve data augmentation which is commonly used to increase the amount of data to ensure that the DNN model has learned sufficient representations of sensitive features. We hypothesize that not all augmentation may be beneficial to a DL system's learning process and may in fact perpetuate existing bias in the data. This benefits our balancing efforts as we can enhance the dataset with images related to features that have insufficient samples while ensuring the added augmentation

are indeed beneficial to the age prediction system.

In summary, our contributions are as follows:

- We propose feature-aware dataset curation, an approach that aims to increase fairness by maximizing the balance among sensitive features.

- We introduce distribution awareness to data augmentation to further increase diversity and sufficiency of sensitive features. We find that not all augmentations benefit performance and that these harmful augmentations are in majority far from the trained data distribution and can be identified and filtered out using our lightweight out-of-distribution detection technique.

- We are the first in age prediction research to evaluate our approaches on a large scale, using 24 DNN models, six benchmark datasets, and one million data points. This enables a fair comparison in which we outperform approaches of prior research [55, 15] and those of AWS and Azure in fairness and performance[1].

2. Background & Related Work

2.1. Age Prediction

With the increase in computing capacity and the rise of deep learning, new approaches rise to advance the age prediction field [13, 61, 19, 2]. Yi et al. [64, 40] was one of the first to utilize convolutional neural networks to extract facial features to estimate age. The best performance on the benchmark dataset MORPH-2 was achieved by Zhang et al. [65] who utilized a network called AL-RoR which combines long- and short-term memory (LSTM) architecture [30] and residual networks [25] to achieve a 2.36 MAE. A study performed by Rothe et al. [55] presented a method based on classification networks, such as VGG-16 [57], which is transferable to ResNet-50 [25] or DenseNet-121 [31], making it easily reproducible and which obtained state-of-the-art performance for age prediction [50] on the MORPH dataset, with an MAE of 2.68. We use the approach from [55] as baseline.

2.2. Fairness & Mitigation

Addressing the fairness of age prediction systems is challenging. This may be due to the fact that sensitive features are not always available as labeled data. Instead, they are inherent in the picture and represented by pixels. A study performed by Clapés et al. [15] aimed at fair age prediction; the authors also analyzed the difference between the real age of a person and their apparent age. Their proposed fairness improvements involve shifting the apparent age towards the actual age during training. The results of the study are somewhat questionable, as the reported performance measured in

---

[1]Our code, DNN models, and public-accessible data are available in our public repository: https://github.com/ForBlindRev/AIBias
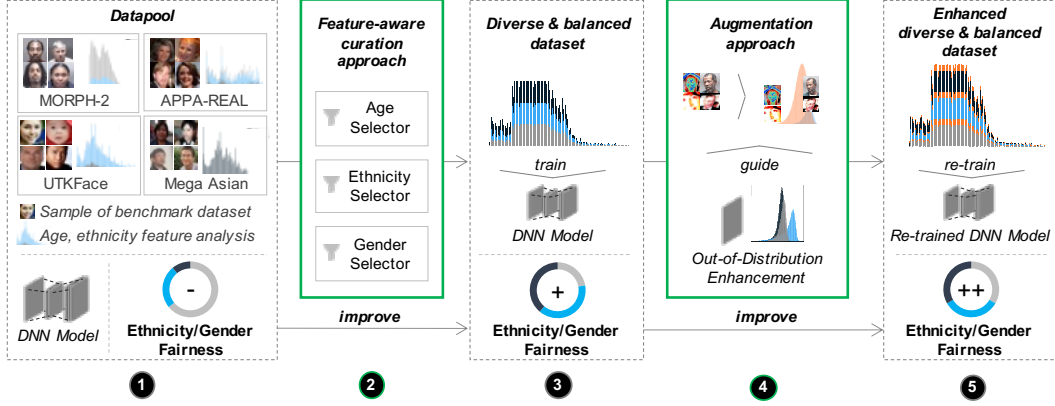
Figure 2: Overview of our approaches.

terms of the MAE varied by 12 years, on average. In other DL domains several mitigation techniques for addressing fairness exist [16, 66, 53, 62, 24, 32, 52, 17, 49]. Given the additional features, e.g., DL loss functions can be optimized towards fair representations [16, 53]. However, such methods remain so far limited in the field of age prediction. This calls for new ways to evaluate and optimize such systems which goes beyond previous efforts focusing primarily on performance, e.g., by NIST using the FERET dataset [51].

2.3. Distribution Awareness

Distribution awareness, or also referred to as Out-Of-Distribution (OOD) detection [27] is the ability to compare what the DNN model has learned to what the DNN model receives as input. The first goal is to filter out data that is not suited for the DNN application, both during training and later during testing and deployment. The second goal is to select data that fits the DNN application and inherits new information the DNN model has not learned yet, e.g., balance ethnicity or gender by selecting diverse yet realistic mutations. Recently, various techniques have been proposed as OOD scoring functions based on the DNN model's activation, such as [54, 56, 41, 14, 37, 60, 38, 28, 58, 46, 39, 29]. These techniques provide different ways of calculating a score on how far or close a given input is to the trained data distribution. Hendrycks et al. [27, 29] used the maximum softmax probability of the DNN model output. Liang et al. [41] added an additional input perturbation to retrieve the OOD score. Others [60, 54, 56] utilized separately trained DNN models or changed the DNN model architecture to calculate the OOD score.

3. Methodology

3.1. Overview

Figure 2 presents the overview of our approaches. We present a novel dataset curation approach and a novel data augmentation approach (both indicated in green). Overall,

the methodology consists of five steps. In step ❶, we analyze the most recent age prediction benchmark datasets which tend to be unbalanced and poorly distributed; to do this, we cross-validate four benchmark datasets along with two DNN training approaches of prior research. The four benchmark datasets have varying diversity and varying quantities of sensitive features which will help in identifying the root cause of unfair behavior. In order to curate a diverse and balanced dataset, we present step ❷, where we employ our novel dataset curation approach using the four benchmark datasets as a combined data pool; this simulates a data scenario commonly faced by large enterprises. Once the dataset is curated in ❸, we further enhance the DL system in step ❹ where we apply our novel distribution-aware augmentation approach. Here, we aim to assess how fairness and performance can be enhanced by filtering augmentations with very high (too different from the trained distribution) and very low (too similar to the trained distribution) OOD scores. Therefore, we re-train the DNN model in step ❺ with the respective augmentation sets to identify optimal data distributional setting for maximizing fairness and performance. To validate the effectiveness, we evaluate our final age prediction system to those following approaches of the previous state of the art in academia and to the leading facial recognition systems from Amazon AWS and Microsoft Azure, using two individual benchmark datasets for assessing generalization.

3.2. Feature Aware Dataset Curation

We aim to curate a dataset which is both diverse and well-balanced among sensitive features. To accomplish this, we propose a dataset curation approach. Given a set containing all the data sources $\{d_1, \ldots, d_n\}$, we denote the union of all data sources by $\mathcal{D}$, the set of sensitive features S, and L the set of labels, we select the samples where x_i^s is the subset of samples containing the feature s with value i, and x_{i,d_t}^L are the samples labeled as a, a subscript d_t denotes that the samples originated from data source d_t:

$$\arg\max_{x \subseteq \mathcal{D}} |x| \text{ s.t. } \forall t, z \leq n :$$
$$\forall a, b \in L : |x_a^L| = |x_b^L| \wedge |x_{a,d_t}^L| = |x_{a,d_z}^L| \qquad (1)$$
$$\forall s \in S, \forall i, j \in s : |x_i^s| = |x_j^s| \wedge |x_{i,d_t}^s| = |x_{i,d_z}^s|$$

Thereby, we aim to balance data among sensitive features for each age while extracting data from as many data sources to maximise diversity. In case insufficient data exists for a given age we define a global maximum (to prevent overfitting) and minimum (to prevent bias) for selecting samples for a given age. The pseudocode of the data curation can be found in the appendix A.5.

3.3. Distribution Aware Augmentation

Effective dataset curation relies on data availability; the ability to address data insufficiency and increase the amount of available data and help improve the balance between sensitive features. This improves fairness, and/or improve the uniformity of available data per age, thereby improving performance. Therefore, data augmentation is commonly used to address data insufficiency when creating a DL system. However, one drawback of augmentation is that even with careful consideration of the augmentation parameters, there is no guarantee that all augmentations will be considered realistic or feasible for the trained distribution or application task of the DNN model [11, 63]. Furthermore, some augmentations may not modify the original image enough, producing too many similar images which may result in overfitting. Different ways to create augmentations have been studied. We illustrate and compare the most prominent approaches in the appendix A.4. and continue with the best performing approach which employs both affine and color augmentations [11, 63, 23].

One of our key novelties is the utilisation of OOD scores when selecting augmentations. Here, we opted for the base methodology of the fast-out-of-distribution detection method (FOOD) [6], which is computationally efficient, does not need to train separate approaches unlike other approaches [29, 56, 14, 37] nor requires additional data [29, 56, 38]. The main idea behind FOOD is replacing the last-fully connected layer with a Gaussian likelihood layer to represent the training data of the model as a multivariate Gaussian with a center vector and a co-variance matrix as parameters. As a result, a likelihood ratio (LLR) can be extracted for each data point which serves as OOD score. We optimize this approach, such that the center μ_c and the co-variance Σ_c can be directly calculated without requiring additional training of the Gaussian layer. Thereby, increasing overall applicability of the technique and enabling faster integration into ML workflows (the mathematical foundations can be found in the appendix A.1).

Algorithm 1: Enhancing a dataset by distribution-aware augmentation.

Result: Curated dataset enhanced by distribution-aware augmentation

1 $median\_num \leftarrow$ median samples in dataset;
2 $mean\_num \leftarrow$ average samples in dataset;
3 $max\_num \leftarrow$ maximum samples in dataset;
4 $max\_ratio \leftarrow ceil(max\_num/mean\_num)$;
5 **for** *all* $c \in C$ **do**
6   **for** *all* $s \in S$ **do**
7          $num \leftarrow$ number of samples in $dataset_{c,s}$;
8          $aug\_ratio_{c,s} \leftarrow ceil(median\_num/num)$;
9          $aug\_ratio_{c,s} \leftarrow$
           $min(aug\_ratio_{c,s}, max\_ratio)$;
10   augmenting data according to $aug\_ratio_{c,s}$;
11   **end**
12 **end**
13 calculating OOD_scores from
    OOD_analysis($augmented\_data$);
14 filter $augmented\_data$ by OOD_scores;
15 **for** *all* $c \in C$ **do**
16   get($select\_size$);
17   **for** *all* $s \in S$ **do**
18   sample($augmented\_data_{c,s}, select\_size$);
19   **end**
20 **end**

We integrate the augmentation and OOD scoring capabilities in Algorithm 1. First, $median\_num$, $mean\_num$, and max_num are defined by the median, average, and maximum number of samples among classes and states (lines 1-3). Then, max_ratio is calculated, which serves to limit the number of augmentations with respect to the overall dataset size and the individual age sample size (line 4). Next, for each class c and sensitive feature s, $aug\_ratio_{c,s}$ is calculated specifically to identify and generate the required number of augmentations to enhance the balance among s without overfitting to a specific representation (lines 5-12). Finally, the OOD scores are calculated by which all augmentation are sorted and filtered (line 13-14). This enables filtering augmentations with a large OOD score, far from the trained distribution, and filtering augmentations with a low OOD score, close to the trained distribution. Thereby, we aim to evaluate under which settings, too different or too similar augmentations can be identified and how they impact fairness and performance. Then, the augmentations are sampled following the same approach as in Section 3.2 (lines 15-20).
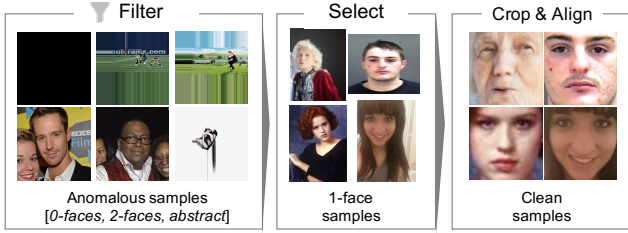
Figure 3: Pre-processing procedure.

4. Evaluation

The evaluation follows Figure 2, in which we first cross-analyze commonly used benchmark datasets; this is followed by the dataset curation and augmentation evaluation. Finally, the overall outcome of our approaches is evaluated with regard to fairness and performance, and compared with the state of the art in research and in industry.

4.1. Experimental Setup

Data. We utilize the IMDB-WIKI dataset to pretrain our DNN models as done in related work [55]. Four benchmark datasets are utilized for cross-evaluation, namely APPA-REAL, MORPH-2, UTKFace and Mega Asian. To compare model fairness and performance to prior work in academia and industry, we employ two independent benchmark datasets (AFAD[47], CACD[12]). Details of the datasets can be found in the appendix A.2.

Pre-processing. For a fair comparison and to prevent dataset bias, all data follows the pre-processing procedure presented in Figure 3, which builds on the preprocessing procedure of [35]. With the help of the DNN facial recognition model, we filtered out over 10,000 of the total 758,613 images, most of which stemmed from the more noisy IMDB-WIKI benchmark dataset. Then, two common facial recognition procedures, namely crop and align, are employed on the resulting clean dataset.

DNN Models. In this work, a total of 24 DNN models are trained. We employ the state of the art in age prediction and computer vision for DNN model architectures. Previous research on age prediction has used the DEX-VGG (VGG-16 based architecture) [57, 55] and the AlexNet architecture [36]. Drawing from the general computer vision domain, we also employ the ResNet-50 [25] and DenseNet-121 [31] DL architectures. In doing so, we aim to cover both domain-specific and general architectures to showcase the potential generalization of our approach.

Evaluation metrics. Fairness and performance are at the core of DL systems exposed to human-centric applications like age prediction. For performance, we opt for the commonly used mean absolute error (MAE) [42, 22], which calculates the mean of how many years the actual age has been mispredicted in absolute terms. For fairness, we build

on commonly used mean distance [55, 66] which takes the mean distance between ages per sensitive features. For age prediction in particular, this distance has to be calculated for each age. Hence, we introduce a *Fairness score* for further evaluation, which takes the mean distance between sensitive features and checks if the distance is lower than a pre-defined threshold $t = 3$, which stems from age discrimination based on the human rights act [1]. Given 100 ages for evaluation, we calculate the fairness score by the per cent of ages, where the mean distance between sensitive features is below the defined threshold (a mathematical definition can be found in the appendix A.3).

4.2. Preliminary Cross-Analysis

DL-based age prediction has relied on four commonly used benchmark datasets which differ regarding their camera setting and distribution of sensitive features. Figure 4 shows the individual distributions by age segmented by ethnicity for each of the datasets. After preprocessing, MORPH-2 is the largest dataset, with predominantly faces of African American ethnicity (80%) and a minority of Caucasian faces (18%); the smallest group is that of Asian faces (2%). The APPA-REAL and Mega Asian benchmark datasets are similarly imbalanced in terms of ethnicity. UTKFace is the only dataset that is somewhat balanced. We train two DNN model architectures for each benchmark dataset following two approaches used in prior research, namely DEX-VGG [55] and AlexNet [36]. Each DNN model is cross-evaluated with all four benchmark datasets. Table 1 presents the average MAE of the DNN models evaluated on the training and testing data from the same benchmark dataset as well as the average of the three other benchmark datasets. We can see that the performance varies significantly depending on the benchmark dataset. For example, for MORPH-2, the average MAE across both approaches is 2.69. However, the average MAE on the other datasets is 10.10, which is the worst of the four datasets. This hints at the similar camera setting used for MORPH-2 dataset's images and the large proportion of one ethnicity in the dataset, which results in the best performance when evaluated by its equivalent testing portion but poor results when evaluated on other datasets. In contrast, when evaluating the DNN models trained on the more diverse UTKFace dataset individually, a slightly lower MAE of 5.23 is obtained. However, the average performance on the other datasets improves, decreasing the average MAE by 20.09%. Detailed results are shown in appendix A.6.

Prior research did not compare the proposed approaches on multiple datasets. We are the first to do this and based on our comparison we observe that a fair comparison between approaches using only one but different benchmark dataset is not possible and requires diverse and balanced data.
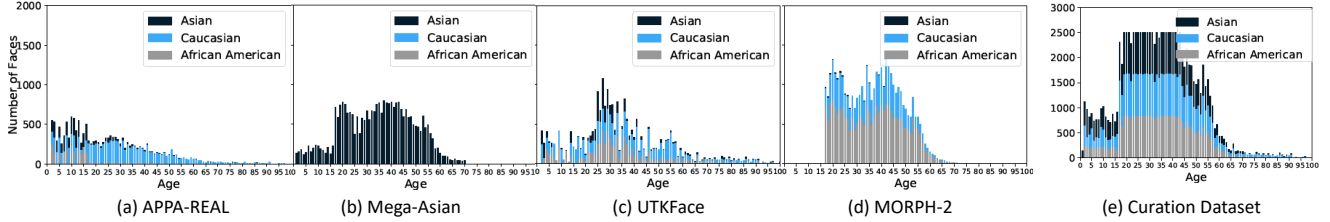
Figure 4: Sample-size per age stratified by ethnicity of the four benchmark datasets (a-d) and the curation dataset (e)

Table 1: Cross-evaluation of prior research.

Training Data	Testing Data	MAE↓
APPA-REAL	APPA-REAL	7.27
	Others (average)	8.74
Mega Asian	Mega Asian	5.03
	Others (average)	9.77
MORPH-2	MORPH-2	**2.69**
	Others (average)	10.10
UTKFace	UTKFace	5.23
	Others (average)	**8.07**

Table 2: Evaluation of dataset curation setting (Ethnicity, Gender in Fairness-Score).

Curation Setting	MAE↓	Ethnicity↑	Gender↑
None	5.05	23.38	61.25
Random	4.03	50.25	76.50
Age only	4.06	63.50	78.50
Ours	**4.00**	**67.50**	**81.00**

4.3. Feature Aware Dataset Curation

We present a balanced and feature-aware dataset curation approach. To create our data pool, we combine the four previously mentioned benchmark datasets, whose individual data distribution by age is presented in Figure 4 (a-d). Three dataset curation baselines are used for comparison. The first baseline is no curation, called *None*, for which the average results are obtained when the DNN models from Section 4.2 are evaluated individually. In the second baseline of random curation the total dataset size of the proposed algorithm's outcome is used to select data randomly from the pool, named *Random*. The third baseline is balanced with regard to its age distribution but selects samples without focusing on balancing sensitive features, named *Age only*. Our approach, named *Ours*, balances both age and sensitive features. The final data distribution by age is illustrated in Figure 4 (e).

The results presented in Table 2 show that the dataset curated using our approach performs best in all categories (row 4, *Ours* setting). This shows the effectiveness when balancing age and sensitive features from all dataset sources, diversity is maximally increased; this benefits MAE to be well maintained across different camera setups. More importantly, the highest levels of fairness are achieved, with a fairness score of 67.50 for ethnicity and 81.00 for gender. Compared to the baseline setting *None*, this represents a 188.71% improvement in fairness for ethnicity and a 32.24% improvement in fairness for gender.

4.4. Distribution Aware Augmentation

Following the dataset curation we are able to increase performance above presented baselines, however, Figure 4

(e) shows various insufficiencies of available data per age. Hence, to address the remaining data points, we augment the data (using fine-grained affine and color mutations based on [11, 63, 23]) and add it to the dataset, which is a common practice in DL system development. We hypothesize that augmentations with a high OOD score may harm the overall DNN model's performance. A high OOD score may indicate an augmented image far from trained distribution and may be considered unrealistic in real-world scenarios. Figure 5 demonstrates this phenomenon, where similar and very different augmentations can be identified based on their OOD score. Those augmentations with a likelihood-ratio (LLR, FOOD's OOD score metric) below the 0.05 quantile (indicating a high OOD score) differ greatly from the original samples. On the other hand, those with a large LLR above 0.95 quantile (a low OOD score) have great similarity to the original samples. This is also reflected by the overlap of the training data (Train) and augmentation data (Aug) distributions. Thus, the results are in line with qualitative human-based perspective.

To quantify the impact of distribution aware augmentation, we employ two independent datasets (CACD [12] and AFAD [47]). In this way, we aim to conduct a fair comparison. Furthermore, utilising the two datasets is close to simulating a real-world scenario in which unknown scenes with all kinds of camera settings and sensitive features are encountered, which helps assessing the DNN model's ability to generalize. The goal is to assess for which distribution aware filter setting, the best fairness and performance is achieved. Table 3 presents the results for five filter settings. For all settings in which mutations are used, we generate 100,000 mutations and select 20,000 of them based on the strategy of the setting. Then, we add them to the 107,404 image large curated dataset. Afterward, the corresponding DNN models are retrained. The first setting, **1**, serves as comparison when
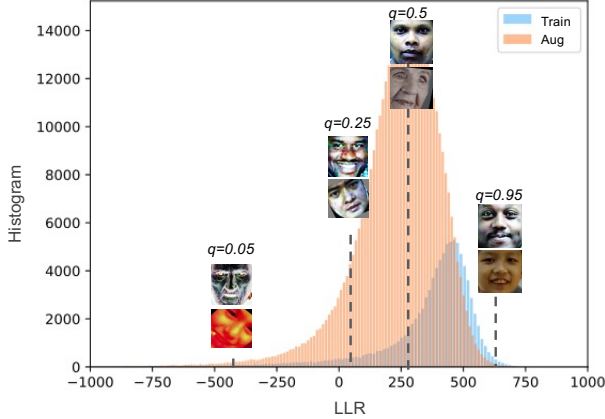
Figure 5: OOD-score distributions of augmentation data (Aug) and training data (Train) with augmentation examples based on LLR quantile.

no augmentations are utilised. Here, the DNN models are trained using data from the previous data curation evaluation with the 107,404 images only (first baseline). In the second setting, **2**, the mutations are randomly sampled without awareness of the data distribution, meaning data of LLR quantiles 0.00 - 1.00 are excepted (second baseline). Settings **3-5** are used to assess our approach when integrating distribution-awareness into the data augmentation process. Here, we filter all augmentations based on their OOD score first, before selecting 20,000 of them for retraining. Setting **3**, filters augmentations far from the trained distribution, which tend to look different, meaning data of LLR quantiles 0.00 - 0.05 are filtered out. Setting **4**, filters augmentations far and close to trained distribution, meaning images which tend to look different (LLR quantiles 0.00 - 0.05) and very similar (LLR quantiles 0.95 - 1.00). Finally, setting **5** serves as contrast to setting **4**, filtering all augmentations except those which are far and close to the trained distribution (LLR quantiles 0.05 - 0.95). Thereby, the goal is to analyze their potentially harmful impact. In the table, column 2 contains the quantile range of the LLR used for augmentation selection. Columns 3-4 contain the MAE for the CACD and AFAD datasets respectively, and columns 5-6 contain the fairness score for ethnicity and gender respectively.

The results show that the best setting is **3**, when augmented data far from the trained distribution is filtered. Compared to the first and second baseline, for the CACD, the MAE improved by decreasing from 4.77 for setting **1** (no augmentation) and 4.65 for setting **2** (no filter) to 4.53 (setting **3**, filtering very different data). Similar behavior is observed for the AFAD benchmark dataset. One further observation when comparing setting **3** to setting **1** is that the fairness score for ethnicity increases from 70.50 to 73.50 and is nearly retained for gender (81.00 vs. 80.00). Filtering out very different and similar data (setting **4**) tends to under

Table 3: Evaluation of distribution aware augmentation approach under different OOD-score filter settings defined by LLR ranges (CACD, AFAD in MAE, Ethnicity, Gender in Fairness-Score).

Setting	Trained LLR Range	CACD↓	AFAD↓	Ethnicity↑	Gender↑
1	No Augmentation	4.77	7.11	70.50	**81.00**
2	[0.00 - 1.00]	4.65	7.08	69.00	75.00
3	[0.05 - 1.00]	**4.53**	**7.01**	**73.50**	80.00
4	[0.05 - 0.95]	4.64	7.05	69.50	78.00
5	[0.00-0.05;0.95-1.00]	4.62	7.03	67.00	80.00

perform in comparison to filtering very different data only. While there is a risk of overfitting with similar data, the DNN model seems to benefit from the increase in data availability more at this point. Finally, setting **5** validates our results, showing that integrating very different and similar augmentations only harms DNN model fairness and underperforms in comparison to all other filter settings.

4.5. Comparison to Related Work

The curation and the augmentation approach of this work are both evaluated on four different DNN model architectures. Having assessed the overall results, we take the best performing DNN model, namely DenseNet-121, and compare it with the state-of-the-art in academia, DEX-VGG and in industrial sector, AWS and Microsoft Azure. To retrieve the results for industrial sector, we created accounts for Microsoft Azure and Amazon AWS gave the test data as input to their age prediction system, which gave the predicted age as output. Microsoft Azure provides a discrete age estimate, while Amazon AWS provides an age interval for which we take the median age. Table 4 presents the results and shows that our age prediction DL system is superior, especially in terms of fairness, increasing ethnicity- and gender-based fairness respectively by 4.92 and 1.88 times when compared to AWS. Also in terms of performance, our DL system performed the best, obtaining the highest MAE with the DenseNet-121 DNN model (an MAE of 4.13 on CACD and an MAE of 6.75 on AFAD). The academic baseline DEX-VGG performed worst, which however, showed slightly better fairness scores when compared to industrial sector. For the evaluated industrial systems, Microsoft Azure's system was better in both, fairness and performance when compared to Amazon AWS' system. It achieved an MAE of 4.17, which is very similar to the MAE obtained by our DL system on CACD. However, it underperformed when compared to our DL system on AFAD.

Our results demonstrate the effectiveness of balanced dataset curation and distribution aware data augmentation, and demonstrate the importance of feature diverse datasets.

4.6. Threat to Validity

Dataset bias may be encountered when curating a new dataset out of four benchmark datasets. To address this,

Table 4: Evaluation of our best DL system, DEX-VGG, and Microsoft Azure and Amazon AWS (CACD, AFAD in MAE, Ethnicity, Gender in Fairness-Score).

DL System	CACD↓	AFAD↓	Ethnicity↑	Gender↑
Ours	**4.13**	**6.75**	**77.00**	**88.00**
DEX-VGG [55]	6.24	9.70	24.00	57.00
Microsoft Azure	4.17	7.58	34.00	56.80
Amazon AWS	5.57	9.91	13.00	30.50

we individually assessed in their distribution to ensure diversity of sensitive features, camera settings, and age and trained a seperate classifier which was unable to destinct between the datasets. For evaluation purposes, we further introduced to two independent age prediction benchmark datasets, namely the CACD and AFAD, to mitigate potential sources for dataset bias.

Using only one DNN model may also present a threat to validity. To ensure confident measures, we trained four different DNN models for each experiment, and the average across these models is presented in Tables 1 - 4.

The ground-truth for determining what data is in fact in-distribution and what data is out-of-distribution is unobtainable. When evaluation the augmentations and OOD techniques we presented a statistical evaluation and empirically assessed the visualisation of the OOD score distribution including 100,000+ augmentations respective to their distribution in Figure 5.

Finally, the non-public training sets used by public cloud age prediction systems may inherit samples used in our evaluation which would dramatically increase their fairness and performance. However, we found that despite this possibility, the performance and fairness of the systems evaluated were low compared to our approach.

5. Discussion and Research Guidance

Proposing age prediction techniques. As our first step, we assessed prior research and observed that single dataset are used to evaluate prior research. In our assessment, we found that there is an imbalance of age and sensitive features in these datasets. Given this, existing approaches for age prediction are difficult to compare using a single benchmark. *Research guidance:* When developing new approaches, multiple benchmark datasets should be used for comparisons. In addition, different DNN architectures should be trained for robust estimation.
Balanced benchmarks. We note that the great diversity of camera settings and sensitive features found in the real world has not been well addressed by existing benchmarks. *Research guidance:* Our dataset curation approach has been shown effective and could serve as a valuable resource to industry and academia, both of which often have access to large sets of samples. To support ongoing development of age prediction methods, we encourage the development of large-scale and diverse benchmark datasets. Such benchmarks can be found in other domains (e.g., ImageNet [18] is used in the computer vision domain for classification tasks). **Integrating distribution awareness.** One of the main contributions of this work is to introduce distribution awareness to data augmentation and thereby improve fairness, with regard to sensitive features. *Research guidance:* To further facilitate distribution-aware DL system development and contribute to future research and improvements in the field, we have made our code publicly available [8]. In particular, there is room for improvement in the data augmentation generation step where distribution awareness could be directly integrated in guiding the mutation criteria.

6. Conclusion

In this paper we presented feature aware dataset curation and introduce distribution awareness to data augmentation. Our approaches aim to increase diversity and maximize balance among sensitive features, such as ethnicity and gender. We compared our novel dataset curation approach, to three baselines which our approach all outperformed. For data augmentation, we analyzed three different augmentation techniques and assessed the different and similar augmentations individually. We found that our modified lightweight OOD-technique is indeed helpful to identify augmentations which are harmful to DNN model fairness and performance, namely those augmentations which are far from the trained distribution. The largest contribution could be made for fairness. Here, our final DL age prediction system outperformed the state of the art system from Amazon AWS or Microsoft Azure by predicting age among individual ethnicity up to 4.92 times more accurate. In addition, our DL age prediction system yielded an MAE of 4.13 on CACD benchmark dataset and 6.75 on AFAD benchmark dataset, thereby outperforming prior research and outperforming DL age prediction systems of Amazon AWS and Microsoft Azure. Overall, our work stressed the importance of curating datasets that consider the sensitive features and data diversity criteria. We gave research guidance along with future directions and provide the technique and code needed to assess both criteria to encourage further work in this important field.

7. Acknowledgments

References

[1] Universal declaration of human rights, Dec. 1948. Online; accessed 16 June 2020.

[2] Arwa S Al-Shannaq and Lamiaa A Elrefaei. Comprehensive analysis of the literature for age estimation from facial images. *IEEE Access*, 7:93229–93249, 2019.

[3] Ibrahim Mousa Al-Zuabi, Assef Jafar, and Kadan Aljoumaa. Predicting customer's gender and age depending on mobile phone data. *Journal of Big Data*, 6(1):18, 2019.

[4] Kanar Alkass, Bruce A Buchholz, Susumu Ohtani, Toshiharu Yamamoto, Henrik Druid, and Kirsty L Spalding. Age estimation in forensic sciences: application of combined aspartic acid racemization and radiocarbon analysis. *Molecular & Cellular Proteomics*, 9(5):1022–1030, 2010.

[5] Amazon. The amazon blog: We are implementing a one-year moratorium on police use of rekognition, 2020.

[6] Guy Amit, Moshe Levy, Ishai Rosenberg, Asaf Shabtai, and Y. Elovici. Food: Fast out-of-distribution detector. 2020.

[7] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, 2018.

[8] Anonymous. Aibias. https://github.com/ForBlindRev/AIBias, April 2021.

[9] BBC. "Apple AI accused of leading to man's wrongful arrest - BBC News", 2019.

[10] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018.

[11] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. Cats are not fish: Deep learning testing calls for out-of-distribution awareness. In *The 35th IEEE/ACM International Conference on Automated Software Engineering*, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[13] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192, 2017.

[14] Hyunsun Choi and Eric Jang. Generative ensembles for robust anomaly detection, 2019.

[15] A. Clapés, G. Anbarjafari, O. Bilici, D. Temirova, E. Avots, and S. Escalera. From apparent to real age: Gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2436–243609, 2018.

[16] Abhijit Das and Antitza Dantcheva. Mitigating bias in gender, age, and ethnicity classification: a multi-task convolution neural network approach. 10 2018.

[17] Thais de Castro Moraes and Xue-Ming Yuan. Data-driven solutions for the newsvendor problem: A systematic literature review. In *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems*, pages 149–158, Cham, 2021. Springer International Publishing.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[19] Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li. A hybrid deep learning cnnelm for age and gender classification. *Neurocomput.*, 275(C):448–461, Jan. 2018.

[20] FaceWatch. "Facial recognition security system UK — Facewatch", 2020.

[21] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976, 2010.

[22] Tobias Gehrig, Matthias Steiner, and Hazım Kemal Ekenel. Draft: evaluation guidelines for gender classification and age estimation. *Technical report, Karlsruhe Institute of Technology*, 2011.

[23] Jianmin Guo, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. Dlfuzz: Differential fuzzing testing of deep learning systems. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2018, page 739–743, New York, NY, USA, 2018. Association for Computing Machinery.

[24] Qing Guo, Xiaofei Xie, Lei Ma, Zhongguo Li, Wei Feng, and Yang Liu. Spatial-aware Online Adversarial Perturbations Against Visual Object Tracking. *arXiv*, 2019.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[26] Will Douglas Heaven. "Predictive policing algorithms are racist. They need to be dismantled.", 2020.

[27] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016.

[28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[29] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[31] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.

[32] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction, 06 2020.

[33] IBM. "IBM CEO's Letter to Congress on Racial Justice Reform", 2020.

[34] IBM. Ibm watson is ai for smarter business, Oct. 2020.

[35] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[37] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

[38] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[39] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7167–7177. Curran Associates, Inc., 2018.

[40] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42, 2015.

[41] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

[42] Khoa Luu, Tien Dai Bui, Ching Y Suen, and Karl Ricanek. Spectral regression based age determination. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 103–107. IEEE, 2010.

[43] Markets and Markets. "Facial Recognition Market by Component (Software Tools (2D Recognition, 3D Recognition, and Facial Analytics) and Services), Application Area (Emotion Recognition, Access Control, and Law Enforcement), Vertical, and Region - Global Forecast to 2024", 2020.

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2019.

[45] Microsoft. "Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM", 2020.

[46] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality, 2020.

[47] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.

[48] US Department of Justice. "Statistical Briefing Book", 2020.

[49] Mustafa Umit Oner, Jianbin Chen, Egor Revkov, Anne James, Seow Ye Heng, Arife Neslihan Kaya, Jacob Josiah Santiago Alvarez, Angela Takano, Xin Min Cheng, Tony Kiat Hon Lim, Daniel Shao Weng Tan, Weiwei Zhai, Anders Jacobsen Skanderup, Wing-Kin Sung, and Hwee Kuan Lee. Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study. *bioRxiv*, 2021.

[50] paperswithcode. "Age Estimation". https://paperswithcode.com/task/age-estimation, 2020. Online; accessed 9 August 2020.

[51] P. Jonathon Phillips, Hyeonjoon Moon, Syed Rizvi, and Patrick Rauss. The feret evaluation methodology for face-recognition algorithms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22:1090 – 1104, 10 2000.

[52] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, 06 2020.

[53] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017.

[54] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14707–14718. Curran Associates, Inc., 2019.

[55] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, December 2015.

[56] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[58] Engkarat Techapanurak and Takayuki Okatani. Hyperparameter-free out-of-distribution detection using softmax of scaled cosine similarity. *CoRR*, abs/1905.10628, 2019.

[59] Jennifer Valentino-DeVries. "How the Police Use Facial Recognition, and Where It Falls Short - The New York Times", 2020.

[60] Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and Krzysztof Czarnecki. Analysis of confident-classifiers for out-of-distribution detection. *CoRR*, abs/1904.12220, 2019.

[61] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li. Auxiliary demographic information assisted age estimation with cascaded

structure. *IEEE Transactions on Cybernetics*, 48(9):2531–2541, 2018.

[62] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. 2020.

[63] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See. Deephunter: A coverage-guided fuzz testing framework for deep neural networks. In *Proceedings of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2019, page 146–157, New York, NY, USA, 2019. Association for Computing Machinery.

[64] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*, pages 144–158. Springer, 2014.

[65] Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao, and Zhanyu Ma. Fine-grained age estimation in the wild with attention lstm networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[66] Peixin ZHANG, Jingyi WANG, Jun SUN, Guoliang DONG, Xinyu WANG, Xingen WANG, Jin Song DONG, and Dai TING. White-box fairness testing through adversarial sampling. 2020.