Fairness Matters – A Data-Driven Framework Towards Fair and High Performing Facial Recognition Systems

YUSHI CAO* and DAVID BEREND*†, Nanyang Technological University, Singapore
PALINA TOLMACH, Nanyang Technological University, Singapore
MOSHE LEVY, Ben-Gurion University, Israel
GUY AMIT, Ben-Gurion University, Israel
ASAF SHABTAI, Ben-Gurion University, Israel
YUVAL ELOVICI, Ben-Gurion University, Israel
YANG LIU, Nanyang Technological University, Singapore

Facial recognition technologies are widely used in governmental and industrial applications. Together with the advancements in deep learning (DL), human-centric tasks such as accurate age prediction based on face images become feasible. However, the issue of fairness when predicting the age for different ethnicity and gender remains an open problem. Policing systems use age to estimate the likelihood of someone to commit a crime, where younger suspects tend to be more likely involved. Unfair age prediction may lead to unfair treatment of humans not only in crime prevention but also in marketing, identity acquisition and authentication. Therefore, this work follows two parts. First, an empirical study is conducted evaluating performance and fairness of state-of-the-art systems for age prediction including baseline and most recent works of academia and the main industrial service providers (Amazon AWS and Microsoft Azure). Building on the findings we present a novel approach to mitigate unfairness and enhance performance, using distribution-aware dataset curation and augmentation. Distribution-awareness is based on out-of-distribution detection which is utilized to validate equal and diverse DL system behavior towards e.g. ethnicity and gender when predicting age. In total we train 24 DNN models and utilize one million data points from five benchmark datasets to assess the state-of-the-art for face recognition algorithms performance and fairness. We evaluated our proposed methodology, demonstrating an improvement in mean absolute age prediction error from 7.70 to 3.39 years and a 4-fold increase in fairness towards ethnicity when compared to related work. Utilizing the presented methodology we are able to outperform leading industry players such as Amazon AWS or Microsoft Azure in both fairness and age prediction accuracy and provide the necessary guidelines to assess quality and enhance face recognition systems based on DL techniques.

CCS Concepts: • **Computing methodologies → Scene understanding**; *Neural networks*; • **Human-centered computing**;

Additional Key Words and Phrases: Out-of-distribution detection, learning generalization, facial recognition, age prediction

1 INTRODUCTION

Deep learning-based facial recognition systems are widely adopted by governments and industry due to its high accuracy. In many cases, a publicly available face recognition service such as Amazon's AWS is used to identify faces and derive estimates to the gender, ethnicity as well as age. Out of these tasks, age classification is considered one of the most challenging, given the high number of possible ages and the small apparent physical difference among them [18]. At the same time, accurate age classification is a very important task given its application in policing [4], surveillance [9, 26], marketing [3, 9] and identity acquisition [4, 9]. In 2019 alone, spending for facial recognition and age prediction services exceeded 3.2$ billion with defense and governmental sectors as main contractors [56]. The public is exposed to the recognition services, whose information is

*Both authors contributed equally to this research.
†David Berend (bere0003@e.ntu.edu.sg) is the corresponding author.

processed and used in majority by governments and defense sector. Hence, assurance for quality and fairness is of high importance.

However, recent public announcements [5, 10, 25] have shown that DL systems for age prediction discriminate users regarding their gender or ethnicity. Here, services are considered unfair as the average prediction is not similar for all ages when comparing ethnicity or gender. This results in applications which use age prediction services to treat e.g. Caucasian different from Asian (protected features). On the contrary, a fair age prediction service would on average predict age similarly among those protected features, which refer to those features that can be used to discriminate an individual by law [57]. This has led big enterprises such as IBM to shut down their facial recognition and age prediction services [41]. Amazon AWS and Microsoft Azure followed this trend, limiting the usage of their services for governments and more specifically police institutions [6, 58]. Similar concerns are raised in academia, too, where unfairness is either detected or overlooked in applications of age prediction research [20]. Joined by global calls for racial equality [42], these recent developments motivate the analysis and improvement of fairness of DL systems used for age prediction.

One strongly debated example is in the field of policing, where the likelihood to commit a crime is higher for a younger person than an older one. When, e.g., Afro-American faces are classified several years younger than Asian faces for the same actual age, policing system relying on DL-based age prediction models will raise more warnings for Afro-American ethnicity groups than for Asian ethnicity groups. Thereby, the age prediction service causes unfair treatment.

In the field of age prediction, recent work has extensively studied deep neural network (DNN) architectures [66, 68, 69] or learning optimizations during training [45, 69] to enhance performance, which is commonly measured by mean absolute error of predicting age (MAE). However, the results and approaches demonstrate a concentrated focus on performance while neglecting the issue of fairness. For example, related work evaluates performance on heavily imbalanced datasets in regards to gender or ethnicity [48, 66, 83, 84], while others have shown that training DNN models with ethnicity-imbalanced datasets leads to performance-differences among ethnicity up to 34.7% [14]. In addition, one to two datasets are used to evaluate related work approaches. Hence, comparing approaches with reported performance is an infeasible task, when imbalanced datasets for evaluation differ. Therefore, we identify two main areas which motivate and guide this work. The first area is about evaluating performance and fairness of current state-of-the-art (SOTA) in industry and academia for age prediction on the same dataset. The second area targets the mitigation of unfairness and the creation of high performing and fair age prediction systems.

Following the presented areas, this work is structured in two parts. First, we conduct a large-scale empirical study integrating the most recent works in academia [18, 66] as well as evaluating the main industry services provided by Amazon AWS and Microsoft Azure. We further integrate a human study as comparison, which contains 7,500 samples. In total, we train 24 DNN models and utilize one million data points from five benchmark datasets to answer four guiding research questions (RQs). *"RQ1: How balanced are commonly used benchmark datasets regarding age and protected features?"* motivates the analysis of the commonly used age prediction datasets with special focus on sample-size distribution of age, gender and ethnicity. *"RQ2: How does balanced testing impact reported SOTA performance?"* assesses how well related work performs in different settings and different protected feature at equal proportion. Thereby, we minimize potential advantage to systems which are trained in majority on one setting which is also existing in majority in the test set for evaluation e.g. one particular ethnicity. *"RQ3: How fair and accurate are SOTA industry, academia and human perception?"* builds on the findings of RQ2 and enlarges the study to industry, further academia, and human perception for comparison. This provides a quantitative foundation for *"RQ4: What are the implications of unfairness?"*, in which we infer the weaknesses and strengths of related work's performance towards overall fairness and ethnicity or gender individually. Our results enable to identify unfair behavior towards age prediction in both academia and industry, with AWS being the most unfair system mis-classifying e.g., female faces on average 4.03 years younger than male faces and Afro-American faces on average 5.22 years younger than Asian faces. Similar patterns are observed in academia. Overall, we show that all systems deviate significantly from their reported performance when evaluated on data which integrates diverse settings and equal proportions of protected features which is commonly encountered in real-world setting. These findings give us strong motivation to develop a new a approach to mitigate unfair behavior and create high performing DL systems for age-prediction along two data-driven enhancements. First, we present an algorithmic approach to curate a balanced dataset and further integrate DNN-based distribution-awareness to understand how certain the DNN model predicts data regarding different features. Second, we optimize the data balancing efforts by introducing a novel distribution-aware data augmentation approach to fix remaining inequalities of size for age. For the data augmentation we utilize distribution-awareness, namely out-of-distribution detection, to filter too similar and unrealistic augmentations to increase the generalization capabilities of the DNN model.

To summarize, our main contributions of this study are:

- We evaluate performance and fairness on related work and industry, identifying differences in predicting age among ethnicity and gender of up 7.59 years.
- We show how to improve the training data such that performance and fairness increase, reducing mean absolute error from 7.70 to 3.39 years and showing an 4-fold increase in fairness score compared to related work.
- Finally, we introduce a novel way for data balancing and data augmentation using out-of-distribution detection for too similar and unrealistic data, demonstrating its effectiveness by increasing the DNN model's ability to predict unknown age prediction benchmark datasets by 20.0%.

This work is organized as follows. Section 2 provides the necessary background and compares our approach with related work. Section 3 introduces the design of our empirical study and technical methodology for enhancement. Section 4 evaluates the study and technical contribution. In Section 5, we discuss the results and provide an overall research guidance. Section 6 concludes the work.

The code together with DNN models, datasets and methodologies is available on our public repository[1].

2 BACKGROUND & RELATED WORK

2.1 Age Prediction

Predicting age using facial images has been studied for years. Early work mainly focuses on extracting facial features using principal component analysis (PCA) [74] or Eigen Faces [44] to predict real age. Similarly, Kwon et al. [48] extracted features such as nose and chin in combination with wrinkle analysis. Others use human aging patterns [29] and bio-inspired features [30] to predict age.

Recently, with the success of deep learning and more specifically convolutional neural networks (CNN), the field could advance as shown throughout various contests [2, 15, 23, 77]. Similar to initial approaches, Yi et al. [53, 80] utilize CNNs to extract facial features to estimate age. Chen et al. [16] transform age prediction into a ranking estimation problem. In another study, Zhu et al. [89] proposed a network called Conditional Multi-Adversarial AutoEncoder with Ordinal Regression (CMAAE-OR), which is based on complex generative adversarial network (GAN) data enhancement

[1] Our repository is available at https://github.com/davidberend/FairnessMatters

to estimate age as well as future or past appearance of a person. This approach achieves an overall mean absolute error (MAE) of 3.27 on commonly used MORPH-2 dataset [65]. The best performance on MORPH-2 is achieved by Zhang et al. [84] who utilize a network called AL-RoR which combines long- and short-term memory (LSTM) architecture [37] and residual networks [33] to achieve 2.36 MAE. Another study from Rothe et al. [66] presents a method based on common classification networks such as VGG-16 [71] and transferable to ResNet-50 [33] or DenseNet-121 [38] which makes it easy-reproducible while maintaining comparably good performance of 2.68 MAE. Thereby, the approach of Rothe et al. ranks among the top works [60] and provides a general approach which this work follows.

We identify that the majority of related work trained DNN models on a single benchmark dataset and reported their performance on the corresponding testset of the same benchmark. These common benchmark datasets differ in size and differ in the diversity of scenes. For example, related work [29, 66, 84] use a dataset which contains only images from similar human pose with white background while others [53] evaluate their approach using a dataset collected from social media sources which integrates significantly higher diversity in overall settings [24]. Thereby, approaches from related works are difficult to compare by their reported results as the data for evaluation is fundamentally different. We hypothesize that the same dataset should be utilized to compare different approaches. Furthermore, discussed approaches and DL systems are commonly used on a variety of applications after deployment. Therefore, we further hypothesize that a DL system should be evaluated with a dataset which inherits different scenes and settings to confidently evaluate performance and fairness for deployment.

2.2 Fairness & Mitigation

Generally, an unfair model can be defined as making skewed decisions due to different protected features, such as gender, ethnicity, religious/political belief or age [57].

In related literature, several unfairness mitigation techniques exist, which assume protected features to be given directly as model input during training [20, 63, 85] or prevent a system to be exploited by visual attacks [32, 40, 62, 78]. Given protected features explicitly as input, loss-functions can be optimized towards fair representations [20, 63]. Additionally, having the explicit protected features, the fairness of the model can be measured quantitatively using a commonly used metric called p-rule [12, 39, 47, 82]. Furthermore, white-box fairness testing techniques [86] assess the fairness of a DL system using gradient computation and clustering techniques.

In age prediction, unfair behavior exists when the average predicted age differs among protected features. However, with facial recognition, protected features are not passed directly as input, instead they are naturally inherited in the picture. This makes the network unable to directly identify a protected feature (e.g., gender or ethnicity) as a matter of ground truth which makes it common to observe inaccuracies in facial recognition and age prediction systems. One recent work regarding unfairness analysis and mitigation for age prediction is performed by Clapés et al. [18], which is used in this work as baseline comparison. Here unfairness in age estimation is introduced through protected features e.g., age, gender, and ethnicity as well as non-demographic characteristics of a person such as make-up, hair or facial expression. In addition, the difference between the real age of a person and their apparent age is analyzed. The proposed unfairness correction involves shifting the apparent age towards the actual age during training. Nevertheless, the results remain questionable as the reported mean absolute error (MAE) varies on average by 12 years, meaning that a face of a person of age 25 is predicted on average wrong on a scale from 13 - 37 years old. In addition, the results are reported based on one DNN model architecture and use MAE as single performance metrics. Das et al. [20] propose a MultiTask-CNN to mitigate unfairness in age prediction. Here, the DNN model architecture is enlarged to age, gender and ethnicity classification at the same time. The results of age prediction are reported in accuracy and vary from 69.5% to 80.1% among ethnicity. The evaluation, however, also focuses only on the performance without assessing unfairness individually.

Overall, several advancements have been made when protected features can be given the DNN model as direct input, which is not the case for age prediction based on face images. Here, mitigation techniques for unfair DNN behavior remain less studied and are performed on DL system which show low performance. Therefore, the reliability of previous methods is questioned and calls for a mitigation approach based on high performing DL systems.

2.3 Distribution Awareness

As mentioned in the previous section, datasets used by related work to evaluate their approaches differ in size among age and overall diversity of scenes captured in the data. Facial recognition and thereby age prediction are used in a variety of settings in respect of different angles, backgrounds, shapes, etc. DNN models trained on a less diverse data may perform badly when encountering a broader range of such scene. Hence, it remains a viable task to verify that the training set integrates the necessary diversity of scenes for deployment. In many cases datasets consists of hundreds of thousands of samples, making it impossible to check its diversity manually. Therefore, it calls for a systematic approach to analysis the overall diversity and distribution of the data.

One way to analyze diversity is directly on the data itself using unsupervised clustering (e.g. k-means) [43] or stratified comparisons (e.g. splitting data by features) [75]. However, these mentioned techniques are analyzing data directly without addressing how it is perceived by the DNN model. For example, two independent DNN models have a different architecture and are trained with the same data. After training, the performance between the two DNN models may be different, meaning they have learned the data distribution in different ways. Therefore, analyzing the data alone regarding its diversity and distribution is not enough, as the decision is ultimately by the DNN model. To analyze data by the DNN model behavior out-of-distribution (OOD) detection mechanisms exist. Here, data is analyzed through the perspective of the DNN model, revealing the perceived diversity of different populations in the data and granting a novel perspective on facial recognition and age prediction datasets.

For definition of OOD, related work [11, 35, 36] distinguishes two totally different datasets from another. Here dataset A has a data distribution D_A another dataset B has a distribution D_B. If they have a totally different distribution (e.g., dataset A contains face images and dataset B contains animal images), the DNN is not expected to handle the data from B. If however A and B have similar distributions, the DNN model is more likely to handle data from B correctly. The distributions are based on OOD scores which are extracted from the DNN model and vary in calculation depending on the OOD method.

Various techniques have been recently proposed to address the challenge of OOD detection, such as [17, 35, 36, 49, 51, 52, 54, 59, 64, 70, 73, 76]. These techniques provide different ways to evaluate the distribution of training data. Hendrycks et al. [34, 36] use the maximum Softmax probability of the DNN model output as OOD score. Liang et al. [52, 54] add additional input manipulation to retrieve the score. Others [64, 70, 76] utilize separately trained DNN models or DNN model architecture-change to calculate the OOD-score distribution behavior.

To the best of our knowledge, we are the first to employ OOD techniques in facial recognition and age prediction domain which enables validating that a DNN model has learned to predict data from various scenes and to validate that a DNN model predicts such scenes similar among protected features. Age prediction brings three main challenges to OOD. First, age prediction includes many output classes (typically 100 different ages), for which techniques such as [35, 36, 54, 59] have been under performing as tested by Amit & Levy et al. [7]. Second, the goal is to provide a lightweight

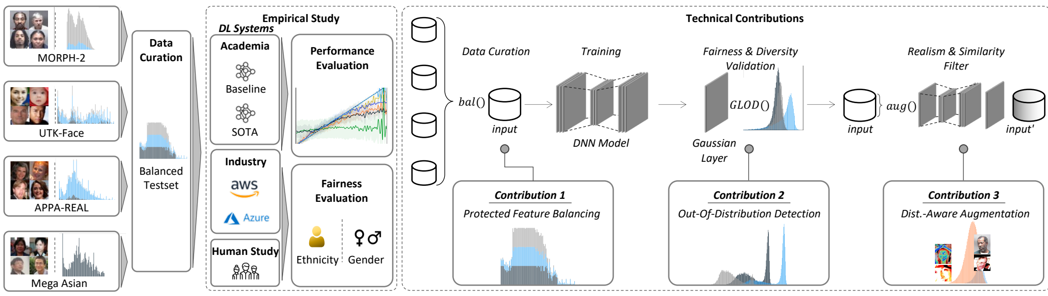


Fig. 1. Workflow of paper

and reproducible approach, which conflicts with OOD techniques requiring separately trained models [17, 36, 49, 70], additional OOD data [36, 51, 70], or are computationally intensive [52]. Hence, our choice falls on Gaussian Likelihood OOD Detector [7]. Furthermore, compared to Lee et al. [52], GLOD is computationally efficient and can be adjusted as presented in this paper to not require additional retraining of the DNN model.

3 METHODOLOGY

3.1 Overview

We present this work in two parts. Following Figure 1, we first conduct an empirical study to evaluate how fair and accurate the current state-of-the-art age prediction algorithms used in both industry and academia are when evaluated on data which stems from various scenes and balanced among gender and ethnicity. The analysis guides in identifying current weakness and strengths of related work and concludes with the drivers for potential unfairness and performance. Based on our findings, the second part addresses the identified weaknesses and proposes two enhancement techniques and one validation technique, namely protected feature balancing, distribution-aware data augmentation, and out-of-distribution detection.

One of the main findings in related work is that there is a high imbalance of gender and ethnicity in popular benchmark datasets which are used to train and evaluate DNN models. Hence, our first enhancement proposes an algorithmic approach towards merging and curating balanced and diverse datasets for training and evaluation of age prediction DL systems. We then utilize a tailored OOD-technique to validate the trained DNN model on perceiving data similarly among protected features. We identify that even with our best efforts, a perfect balance among protected features and age cannot be achieved given the available public benchmark datasets which is a common problem especially in industry. Therefore, we propose distribution-aware data augmentation as the second enhancement to further balance data using most recent AutoAugment [19] and random augmentation techniques. In addition, we study the impact of OOD-detection when filtering out augmentations that are perceived unrealistic or too similar compared to the trained dataset.

3.2 Study Design

This section explains the general setup of the empirical study. The goal is to evaluate performance of the existing state-of-the-art age prediction algorithms and identify its implications towards fairness. The empirical study follows four research questions:

RQ1: How balanced are commonly used benchmark datasets regarding age and protected features? We take a look at the benchmark datasets used by related work, namely APPA-REAL [18], MORPH-2 [65], UTK-Face [88], and Mega Asian [87] to identify their individual sample-size towards different ages from 1 to 100. Then, we identify the proportion of protected features in each sample. Thereby, we can understand the distribution used for training of related work's DL systems and curate a balanced testset which enables a comprehensive evaluation and comparison among SOTA DNN models.

RQ2: How does balanced testing impact reported SOTA performance? We hypothesize that approaches for age prediction by related work are difficult to compare since they use datasets with significantly different distributions and style. Therefore, we evaluate two settings on a trained DNN model from related work [66] which is trained on commonly used imbalanced dataset MORPH-2. For the first setting the corresponding MORPH-2 testset is used to evaluate performance. For the second setting the balanced testset from RQ1 is used to evaluate performance and compared with setting one. Furthermore, fairness is assessed, in particular the differences in predicted age for ethnicity or gender.

RQ3: How fair and accurate are SOTA industry, academia and human perception? We aim at validating potential findings from RQ2 by extending the study to industry, including Amazon AWS and Microsoft Azure DL systems, as well as one further related work from academia [18] which is considered as baseline. In addition, a human study from APPA-REAL [1] is used as comparison with 7500 age predictions to provide a conclusive perspective on the DL system performance regarding actual and apparent age.

RQ4: What are the implications for unfairness? Answering the aforementioned research questions, we acquire a fine-grained understanding of the overall SOTA DNN model performance and the individual performance regarding ethnicity and gender. Thereby, we are able to study the differences between the protected features to formulate a conclusion on what fairness means in context of age prediction and on how fair the protected features are classified by current SOTA.

3.3 Enhancement Approach

As described in Section 2 regarding age prediction, the focus of related work is in majority on improving performance in MAE using a single dataset, which is similar in style and imbalanced among age, gender and ethnicity. Hence, the problem of unfair prediction is mostly neglected. Building on the findings of the empirical study, we propose two algorithmic approaches to (1) curate balanced datasets regarding protected features and (2) optimize the balancing efforts with data augmentation. For both enhancements, we further introduce distribution-awareness using out-of-distribution detection as validation technique and data selection technique. Thereby, we approach performance and fairness enhancement from a data-driven perspective and evaluate each technique with the proposed evaluation criteria of Section 4.2.

3.3.1 ***Distribution Awareness.*** Distribution awareness lies at core in both of the enhancements. We acknowledge its relevance from two perspectives. First, a dataset may be balanced in quantity regarding protected features, while the diversity between features and scenes is low. Therefore, distribution-awareness enables validating that a DNN model has learned to predict data from various scenes and validating that a DNN model predicts such scenes similar among protected features. For example, a dataset may consist of 10,000 faces from both Caucasian and Asian faces. However, for Caucasian faces, 2000 images are taken from 5 subjects while for Asian faces, 10 images are taken from 1000 subjects. Thereby, the number of different subjects for Asian faces is 200 times higher, which has higher diversity and gives the DNN model a better ability to predict age for Asian faces, as it is familiar with more styles of shapes and sizes. The second perspective focuses on filtering data. When data augmentation is used, not all augmentations may be considered relevant to the application or may be considered as outliers. We hypothesize that too similar and too different (unrealistic) augmentations in respect to the training set, may be harmful to DNN model performance or the ability to generalize to unknown inputs. Hence, we utilize OOD-detection techniques to identify the right thresholds of OOD scores (based on the Gaussian likelihood of a prediction) to filter out too similar as well as too different augmentations to validate our hypothesis.

In related work in Section 2, we identify that current SOTA OOD techniques have several drawbacks: (1) difficulty on handling many output classes, (2) require additional data to familiarize the DNN system with application-irrelevant data or (3) require additional model adjustments or training of separate DNN models for OOD detection. The identified technique Gaussian Likelihood out-of-distribution Detector (GLOD) achieves highest reported accuracy among related work and addresses the first two drawbacks. However, the original method requires DNN model adjustments to retrieve the Gaussian likelihood for OOD score calculation. Hence, we integrate a variant of the GLOD technique, which is able to directly calculate the OOD score, without any further retraining of the DNN model.

The full GLOD implementation replaces the last fully-connected layer with a Gaussian likelihood layer and requires further training such that the data representation is learned. In our case we resort to an adjusted version of GLOD that only replaces the final layer for score calculation without training a specific data representation for the layer beforehand.

For the score calculation, GLOD utilizes a Gaussian likelihood layer in order to model the likelihood of each class in the penultimate representation. The penultimate layer is commonly used for analysis as it contains the most processed information without limiting the feature space to the relatively small number of output classes. Thereby, the use of this layer supplies the DNN with the capability to estimate certainty at the highest granularity; the closer the sample to the class center is, the higher the confidence that the input belongs to a certain class and thereby to the trained distribution. With the help of the Gaussian layer the data is represented as a multivariate Gaussian. The multivariate Gaussian [8] has two parameters, a center vector and a co-variance matrix. The Gaussian layer includes those two parameters for each class, which are directly calculated based on the training data via the variant presented in this paper. For class c and penultimate representation of the dataset X, we calculate the center μ_c and the co-variance Σ_c as follows:

$$\mu_c = \frac{1}{|c|} \sum_{x_i \in c} x_i \tag{1}$$

$$\Sigma_c = \frac{1}{|c|} \sum_{x_i \in c} (x_i - \mu_c)(x_i - \mu_c)^T \tag{2}$$

With d-dimensional penultimate representation, where $\mathcal{N}$ stands for multivariate Gaussian distribution as:

$$f(x|\Sigma_c; \mu_c) = \log\left(\mathcal{N}(x|\mu_c; \Sigma_c)\right) = \tag{3}$$

$$-\frac{d}{2} log(2\pi) - \frac{1}{2} log(|\Sigma_c|) - \frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c)$$

To calculate the score, Equation 4 takes the probability ratio between the log of the predicted classes and the not-predicted classes. Compared to other OOD techniques, which only utilize the probability of the predicted class, GLOD is capable of integrating the overall DNN model behavior as all class outcomes are integrated in calculation. Thereby, a more fine-grained analysis of OOD-score distribution behavior is possible. Given the trained DNN model and the corresponding distribution scores of the training data, an OOD score can be retrieved for a new input sample. K represents the group of the k class indices with the top likelihood scores, and does not contain $\hat{y}$.

$$\mathcal{LLR} = log(p_{pred}) - \log(p_o|\hat{y} = c; k) \tag{4}$$

For detection of outliers or out-of-distribution data, GLOD uses the $\mathcal{LLR}$ statistical test. This test provides an estimate that measures how far away the sample is from its predicted class in the penultimate representation. Samples that are too far away from their predicted class relatively to other classes are given a low score and tagged as OOD. The $\mathcal{LLR}$ test is calculated as a subtraction of two log-likelihood scores. The first one is the log-likelihood of the predicted class:

$$log(p_{pred}(x)) = \max_{c \in \{1,\ldots,C\}} f(x|\mu_c; \Sigma_c) \tag{5}$$

and the second one is an estimate for the log-likelihood outside of the predicted class:

$$\log(p_o(x|\hat{y} = c; k) = \frac{1}{k} \sum_{i \in K} f(x|\mu_k; \Sigma_k) \tag{6}$$

The parameter K may be commonly set to the number of available output classes. Nevertheless, for applications such as age prediction, it is not guaranteed that data exists for all 100 classes which influences overall score calculation as the Gaussian clusters are then placed randomly in the overall calculation space contributing unproportionally to each calculation. Hence, the flexibility of adjusting K enables the OOD technique to adjust to available classes or represent the data in the appropriate application-specific way.

To identify that two OOD-score distributions are significantly different, we use the t-value as it is a common measure used to distinguish two OOD-score distributions from each other [81].

Using the presented OOD score provided by the adjusted GLOD technique, we are able to distinguish diversity between ethnicity or gender using the overall OOD-score distribution and are able to filter augmentations which are perceived too similar or different by the DNN model using thresholds regarding OOD-score. Furthermore, the higher the overlap of the individual protected feature OOD-score distributions the more similar they are predicted. Thereby, we hypothesize that similarly perceived feature OOD-score distributions contribute to fair learned representation between such features.

3.3.2 ***Protected Feature Balancing.*** The first enhancement is based on balancing data among all ages and between protected features. We hypothesize that balanced training data will result in a trained DNN model which predicts protected features at similar accuracy and age to address the fairness problem. Our novelty in the balancing approach lies in validating the effectiveness of balancing by analyzing the diversity of the dataset using OOD-technique GLOD. We hypothesize that a similar diversity of OOD scores among protected features points towards similar performance and higher fairness. Since data shortages are nothing unrare, perfect balancing is not always guaranteed, for which diversity validation techniques offer a way to estimate deployment-readiness. As a result, datasets are balanced in overall quantity of protected features, e.g., 1000 Asian face images and 1000 Caucasian face images, and diversity, e.g., 100 subjects of each ethnicity from which the images were taken from.

The main challenge lies on curating a dataset which inherits a diverse amount scenes to help the DNN model learn the variety of settings it will encounter in the real world after deployment. Since age prediction is a human-centric task, the DNN model requires to learn the representations of protected features equally to predict them with the same average age. Our proposed Algorithm 1 lies emphasis on protected features, such as ethnicity or gender while extracting the data from different sources with the goal to maximize diversity and similarity among average age prediction. First, data is segmented for each class c in all the classes C (which are 1-100 ages in this case), with the total

Algorithm 1: Curating protected feature balanced dataset

```
   Result: Processed balanced dataset
1  num_sample ←Sum of number of samples from all datasets by class C and state S;
2  sort(num_sample by s);
3  max_sample ← min{quantile(num_sample_{c,s}, 0.8)|s ∈ S};
4  min_sample ← max{quantile(num_sample_{c,s}, 0.2)|s ∈ S};
5  for all c ∈ C do
6      threshold ← min{num_sample_{c,s}|s ∈ S};
7      threshold ← min(max_sample, max(min_sample, threshold));
8      ds_num ← number of datasets;
9      select_size ← threshold/ds_num;
10     for all s ∈ S do
11         sort(D,c,s);
12         for all d ∈ D do
13             num ←length of d_{c,s};
14             if num < select_size then
15                 select all data in d_{c,s};
16                 remain ← select_size − num;
17                 update(select_size, remain) ;
18             else
19                 random_select(d_{c,s}, select_size);
20             end
21         end
22     end
23 end
```

number of samples for each state s of protected features S (which are ethnicity and gender in our case) for all datasets D (Line 1). Then, the minimum number of samples among all states for class c is identified as the sampling threshold to make sure each state has the same number of samples after balancing (Line 6). Also, maximum and minimum sample sizes (*max_sample*, *min_sample*) are defined to serve as cutoff threshold to avoid overfitting issues where samples are highly concentrated on only a few ages. Grouping the data by age for each state, the minimum amount *min_sample* is identified by selecting a threshold at the 0.2 quantile of the age sample sizes and the 0.8 quantile as the maximum *max_sample* (Line 2-4). With *min_sample* all data is taken into consideration if the available sources together do not possess the necessary amount for balancing. Thereby, a necessary base performance is enabled while maximizing fairness to the best ability. With *max_sample* the ratio among individual sample size per age can be maintained to reduce risks regarding overfitting. For each age from 1 to 100 a age threshold is retrieved based on the smallest state of the available age sample. The threshold is adjusted to *min_sample* or *max_sample* if it smaller or higher respectively (Line 7). Afterward, it is divided by the number of available datasets which serve as data sources to retrieve *select_size* (Line 8-9). Finally, *select_size* of data is extracted from each state s from each source dataset d to maximize balance among protected features S and diversity among datasets d. In case d has insufficient data available the remainder *remain* is calculated and taken from the other sources (Line 19). If sufficient data exists, *select_size* of data is random sampled from $d_{c,s}$ (Line 15-17). The detailed approach for dataset creation is shown in Algorithm 1.

Following the balancing approach, we present a systematic way to curate datasets with similar occurrence of protected features among age and high diversity among source datasets. Thereby, a DNN model may be able to capture information equally for all classes and features, which will be evaluated by the validation step using OOD detection.

3.3.3 ***Distribution-Aware Data Augmentation***. The second enhancement is distribution-aware data augmentation. Augmenting data is a common practice to enhance datasets such that the DNN model is able to learn additional features, which may contribute to overall performance, generalization, and fairness. For performance, we hypothesize that balanced datasets achieve better results than imbalanced datasets when evaluated in the real-world setting. For age prediction in particular, benchmark datasets are commonly imbalanced among age. Hence, being able to augment samples from minority classes contributes to the balancing approach. For generalization, the DNN models may become more robust to unknown data as new features are learned, which may enable prediction of the unknown inputs with higher accuracy. For fairness, data augmentation enables further balancing between protected features, which may contribute positively to similar average age prediction between, e.g., ethnicity.

One drawback from augmentations however is that even under careful consideration of augmentation parameters there is no guarantee that all augmentations can be considered realistic or feasible to the trained distribution of the DNN model. At the same time, some augmentations may not produce a significant change to the initial image and may thereby be considered as too similar, which may cause overfitting to one specific data representation.

Therefore, we apply out-of-distribution detection to identify at what point augmentations exceed realism and at what point augmented change is too low and therefore the result remaining too similar. We define cut-off thresholds based on OOD score to filter out the undesired augmentations. Thereby, the model can be enhanced with highly diverse yet realistic augmentations.

For the augmentation approach, max_num is defined by the maximum number of samples among classes and states and the max_ratio is calculated to limit the number of augmentations in respect to the overall dataset size and the individual age sample size (Line 3-4). Next, for each class c and state s, $aug\_ratio_{c,s}$ is calculated specifically to identify the size of necessary augmentations to enhance balance without overfitting to a specific data representation (Line 7-10). Afterward, OOD analysis is employed, calculating the OOD-scores for all augmentations and sorting the data by the score (Line 13). Using defined thresholds for unrealistic and too similar for the OOD-score distribution, the augmentations are random sampled from the defined area following the same approach as in the balancing algorithm indicated as $Balanced\_sample()$ by calculating $select\_size$ (Line 16) as the minimum number of samples among all protected_features S after augmentation (Line 17-19). Our overall approach is formalized in Algorithm 2.

To produce augmentations, we utilize and compare AutoAugment [19], which represents one of the most recent contributions for augmentation creation, and random augmentation guided by defined parameters. Auto-Augment integrates several pre-defined settings which are used depending on the feedback of the DNN-model. Thereby, it claims to produce high diversity, yet model compatible inputs. For random augmentation we define parameters following previous work [11] and validate 100+ of the outcomes manually to make sure realism of produced images is maintained. For further validation, we manually assess behavior by visualizing augmentations of different OOD score.

We hypothesize that augmentations with high OOD score may harm the overall DNN model performance in MAE as their diversity exceeds the trained OOD-score distribution and may be considered unrealistic towards real-world scenarios. For generalization, both face images with low and high OOD score may contribute, whereas it is likely that the more diverse (i.e., with a

Algorithm 2: Creation of augmentation dataset

```
Result: Augmented dataset for finetuning
median_num ← median number of samples in dataset;
mean_num ← average number of samples in dataset;
max_num ← maximum number of samples in dataset;
max_ratio ← ceil(median_num/mean_num);
for all c ∈ C do
    for all s ∈ S do
        num ← number of samples in dataset_{c,s};
        aug_ratio_{c,s} ← ceil(max_num/num);
        aug_ratio_{c,s} ← min(aug_ratio_{c,s}, max_ratio);
        augmenting data according to aug_ratio_{c,s};
    end
end
calculating OOD_scores from OOD_analysis(augmented_data);
splitting augmented_data according to OOD_scores;
for all c ∈ C do
    get(select_size);
    for all s ∈ S do
        balanced_sample(augmented_data_{c,s}, select_size);
    end
end
```

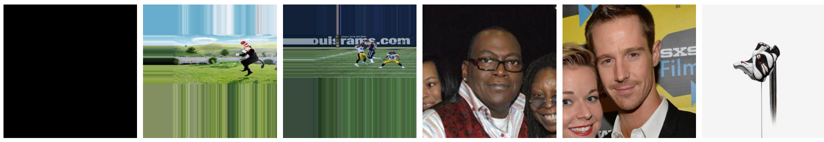
higher OOD score) the augmentation is, the better the DNN model may generalize. To test our hypothesis, we utilize a fifth benchmark dataset FG-NET [50] which is non-integral to the training or testing set, differs in style and therefore unrelated to the learned data OOD-score distribution of the trained DNN model. The dataset contains images of 82 people from which 10-15 pictures were taken at different age. In total, it consists of 1000 face images, which are used for validating the DNN model's ability to generalize towards unknown data from various scenes and settings.

4 EVALUATION

Following the proposed methodology, we structure the evaluation into two parts. First, we conduct an empirical study of the existing state-of-the-art to analyze performance and fairness on both imbalanced and balanced data. Addressing the strengths and weaknesses of current SOTA, we evaluate distribution-aware balancing and distribution-aware data augmentation following Section 3.3. Finally, we conclude by comparing our approach to related work.

4.1 Experimental Setup

Data. One important aspect concerning the evaluation is the testing dataset used to analyze the performance of the individual systems. Hence, we curate a testing set presented in Figure 3 which inherits face images from four commonly used benchmark datasets with the goal of diversifying faces. Most importantly, we combine the datasets to the best of our ability to equally balance each ethnicity (Asian, Caucasian and Afro-American) following Algorithm 1. This enables fine-grained analysis of the DNN model in a setting it would face after deployment encountering images of different style with faces varying in shape and size.



(a) Only black (b) No face (c) No face (d) 2+ faces (e) 2+ faces (f) Anomaly (g) Anomaly

Fig. 2. Pre-processing examples

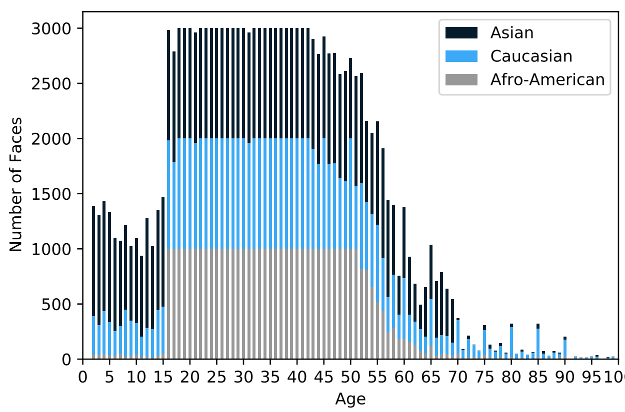


Fig. 3. Balanced dataset sample-size distribution

Pre-Processing. To carefully analyze common benchmark datasets, we identify potential sources for invalid face images. To further accelerate the discovery of such images, a face recognition DNN model is used to filter out images where zero faces or more than one face are identified. The discovery shows four common sources of invalid samples: (1) totally black images, (2) abstract and unrelated images without faces (anomalies), (3) comic or fiction related faces, and (4) multiple faces of a group of people (examples are presented in Figure 2). With the help of face recognition DNN model, we filtered out over 10,000+ out of 636,022 images, where most stemmed from IMDB-WIKI benchmark dataset which is used for pre-training the DNN model.

DNN Model Architecture & Training. In this work a ResNet-50 architecture [33] is used consisting of residual components and 2,560,000 parameters. In addition, VGG-16 [71] and DenseNet-121 [38] architectures are utilized for validating the data-based hypothesis by observing similar behavior between the architectures.

The DNN models are trained using Transfer Learning [72], which has shown significant performance improvement in related work [66]. Here, the DNN models are pre-trained on large and noisy data from ImageNet [21] (provided by PyTorch library [61]) and further pre-trained on commonly used dataset IMDB-WIKI [67]. Thereby, the DNN model weights become pre-adjusted to the overall domain, which can then be finetuned to a given application using a dataset with higher granularity.

Hyper-parameters. For the DNN model optimizer, the approach of Rothe et al. [66] is followed who use an Adam optimizer with learning rate at 0.001 for pre-training on ResNet-50 and DenseNet-121 architecture. For DNN model with VGG-16 architecture an SGD optimizer with learning rate at

0.001 is utilized instead, as previous work has shown low learning effectiveness with Adam [46]. During finetuning, the learning rates are all set to 0.0001. In addition, the learning rate is reduced every 40 epochs by a factor of 5 for all settings.

Technical Setup. The experiments are conducted on a high performance computer cluster with each cluster node running a GNU/Linux system with Linux kernel 4.4.0 on 2 18-core 2.3GHz Intel Xeon CPU E5-2699 with 190 GB RAM equipped with two NVIDIA Tesla P/M40 GPU.

4.2 Evaluation metrics

4.2.1 ***Fairness Metrics***. For fairness evaluation, two metrics are utilized which are based on existing principles of fairness assessment and tailored towards age prediction. The first fairness testing criteria is perception distance, which measures the average difference between actual and perceived age. This metric is provides insight in assessing by how much each, e.g., ethnicity is perceived younger or older compared to the actual age of the sample.

For the second fairness metric, the common fairness metric, namely mean-distance presented in Section 2.2 is utilized in combination with the p-rule as estimate when an application is considered fair. The p-value, guided by the *Uniform Guidelines On Employee Selection Procedures* [12], is set at 80%, which implies that if 80% of the samples are predicted similar, we consider the DL system fair.

In detail, the fairness score provides an aggregated metric to compare how different protected features are predicted, which provides a decision support for any safety-critical or ethical-relevant application, such as public surveillance.

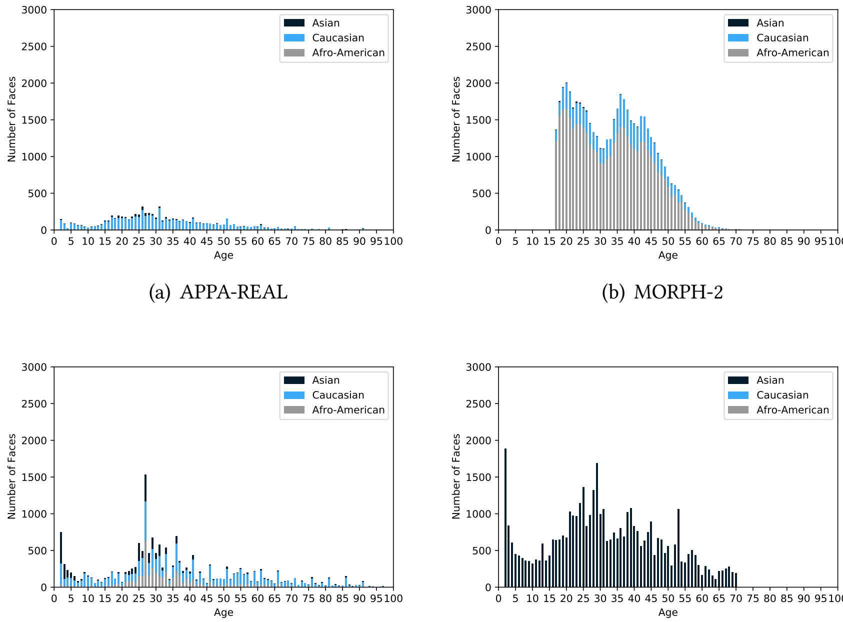
To calculate the mean-distance for the fairness score, the mean distance in average predicted age y_i between all candidates $\mathbf{s}$ is calculated and validated if it is close enough to threshold t (shown as F in Function 8), which is set to 3 years following average error reported by SOTA DL system [66] and the severity of safety-critical applications such as policing systems [4]. In this case 3 years of difference in average age prediction is the limit at which fair treatment between protected features is considered. To retrieve F, the perceived age P from one candidate s_1, e.g., Asian, is taken and then compared via K (Function 9) to all the perceived ages P (Function 9) of all other candidates, e.g., Afro-American and Caucasian. If K satisfies the requirement between all candidates, F returns a positive value following the indicator function $\mathbb{1}$. Since F can either be either 0 or 1, we divide the total positive results for F by the total number of ages, which gives us the fairness score p. If p is above the p-value of 0.8 we consider the DL system fair. In Equation 7-9, $\hat{y_i}$ is denoted as the predicted age at the i-th sample, y_i is denoted as the real age at the i-th sample, n is the total number of the samples.

$$p = \frac{1}{n} \sum_{i=1}^{n} F(y_i|\mathbf{s}) \tag{7}$$

$$F(y_i|\mathbf{s}) = \mathbb{1}\left(\left(\sum_{j=2}^{m} K(y_i|s_1, s_j)\right) = m - 1\right) \tag{8}$$

$$K(y_i|s_1, s_j) = \mathbb{1}\left(|P(y_i|s_1) - P(y_i|s_j)| < t\right) \tag{9}$$

4.2.2 ***Performance Metrics***. To evaluate performance in classification tasks, the most commonly used metric is top-1 accuracy, which calculates the mean of correct predictions. Given the difficulty of age prediction and the number of classes, top-1 accuracy remains important for comparison to other facial recognition applications, however, further performance metrics are required for thorough evaluation of our study and enhancements.



(a) APPA-REAL (b) MORPH-2

(c) UTK-Face (d) Mega Asian

Fig. 4. Common benchmark dataset sample-size distributions stratified by ethnicity

Three commonly used metrics for age prediction are evaluated. In particular: MAE, $CS(\theta)$ and AMAE [28, 55]. MAE is the mean absolute error in years over all samples, $CS(\theta)$ is the cumulative score measuring the proportion of predictions which have an absolute error lower than θ. AMAE measure the MAE for each age first and then takes the average over all the means of ages. $CS(\theta)$ is sensitive to the choice of θ which may cause inaccuracy when comparing different DL system approaches. AMAE may be considered effective when the number of samples per age are similar. However, this is not the case for age prediction which is why for this study MAE is the key metric to evaluate performance

Besides MAE and top-1 accuracy, we further choose 1-off accuracy as it evaluates performance with higher tolerance given the difficulty of the prediction task. Here, a prediction is considered correct when predicting age correctly and ±1 years older or younger.

4.3 Empirical Study

RQ1 motivates the analysis of the commonly used age prediction datasets with special focus on sample-size distribution of age, gender and ethnicity. RQ2 assesses how well related work performs in different settings and different protected feature at equal proportion. Thereby, we minimize potential advantage to systems which are trained in majority on one setting or e.g. ethnicity which may exist in majority in the test set, too. RQ3 builds on the findings of RQ2 and enlarges the study to industry, academia, and human perception for comparison. This provides a quantitative foundation for RQ4 in which we infer the weaknesses and strengths of related work towards fairness ethnicity or gender specific performance.

RQ1: How balanced are commonly used benchmark datasets regarding age and protected features? For RQ1, the goal is to identify the sample-size distribution of protected features in the four commonly used benchmark datasets which are used for training the DNN models by related work and SOTA. We hypothesize that training a DNN model with imbalanced data among age and ethnicity or gender has consequences on the individual accuracy of classes or features which may result in unfair predictions. Therefore, we want to understand the sample-size distribution of the data on which related work trained their DNN models.

Figure 4 presents the sample-size distribution split into three ethnicity Asian, Caucasian and Afro-American. We identify that MORPH-2, one of the most popular datasets [65], consists of 80% Afro-American, 19.8% Caucasian and 0.2% Asian faces. Similar patterns are observed for APPA-REAL and UTK-Face. Mega Asian is meant to focus on one ethnicity only. Age wise, data is distributed with higher concentration from age 16 to 55.

Answer to RQ1: Careful analysis of the datasets show that benchmark datasets contain data which largely differs among age and ethnicity. Most work optimizes and compares DNN models focusing on only one dataset, which is used for both training and testing. Hence, reported performance is mostly constrained by an age range and a single ethnicity, which may impose significant difficulties when evaluating the different approaches for deployment or when comparing results between each other.

RQ2: Balanced testing on SOTA DNN Models. Given the results RQ1, all datasets show imbalanced sample-size distributions among age and protected features. In RQ2 the goal is to identify the difference in performance and fairness when testing an imbalanced DNN model with two different settings. The first setting is testing the DNN model on the original data, which was used by related work to report results. The second setting is testing the DNN model on the curated balanced dataset presented in Section 4.1 which integrates multiple datasets with each ethnicity in equal proportion. In case the difference between the evaluation settings is high, it emphasizes that reported results of related work are difficult to compare and require further quality assessment. First, the performance of the DNN model is analyzed for setting one based on original imbalanced MORPH-2 dataset to recreate results of related work. The results are presented for each ethnicity individually in Table 1 in row 'Unbalanced'. Overall, the MAE is 2.21 years, which is similar to the performance reported by related work. Overall a fairness score of 0.78 is retrieved which is slightly below the p-value of 0.8 and therefore may almost be considered fair. We conclude that the DNN model performs under unbalanced setting at state-of-the-art levels similar to the reported results by related work [57, 66].

For the second setting, Figure 5 presents the perceived age between ethnicity on the balanced testset. The results in Table 1 in row 'balanced' show that once the imbalanced DNN model is evaluated on the balanced testset, the performance as well as fairness drop significantly. The mean absolute error (MAE) increases from 2.21 to 7.70 years and fairness score decreases from 0.78 to 0.18. Furthermore, following Figure 5 a large perception gap can be observed between ethnicity from age 30 to 60. One reason for the performance drop may be the imbalance in available data per age by the MORPH-2 dataset. Thereby, the DNN model is trained well for some ages while it has not learned enough features for others. Nevertheless, the results also demonstrate, that even for those ages with highest available data e.g. age 20 or 40 (Figure 4), the performance between ethnicity varies 5 to 10 years (Figure 5), hinting towards an overall unfair treatment among protected features and an underperforming DNN model.

Table 1. Comparison of imbalanced and balanced testset evaluation

Testset	MAE	Acc (%)	1-off Acc (%)	Perc.	Fairness
unbalanced	2.21	21.01	46.52	-0.32	0.78
balanced	7.70	6.17	20.07	0.78	0.18

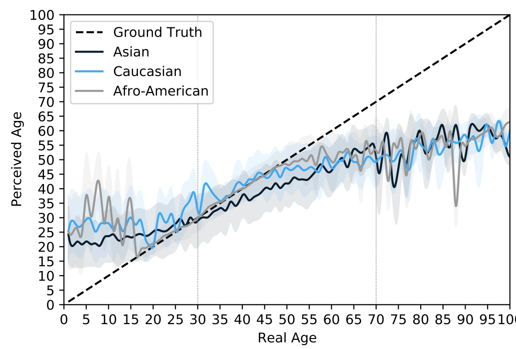


Fig. 5. MORPH-2 DNN model on balanced testset

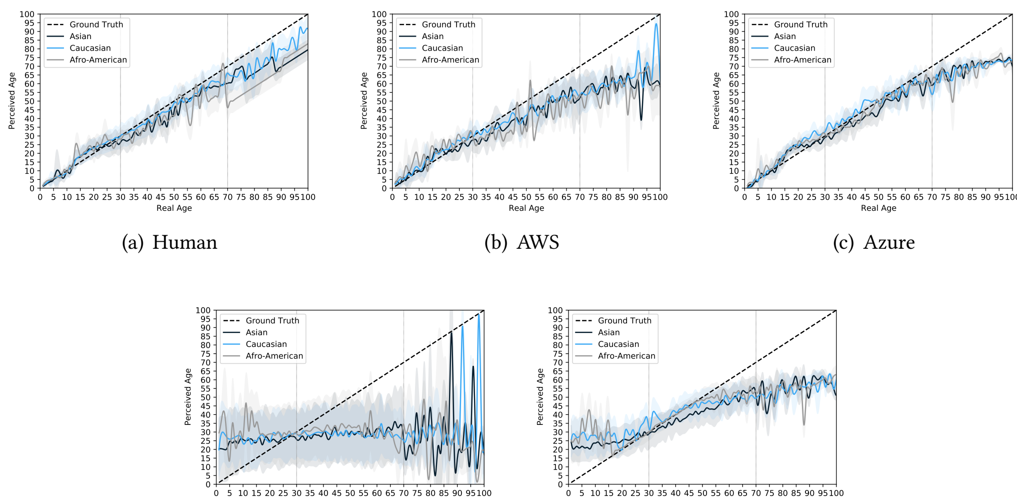
Answer to RQ2: The DNN model trained on commonly used imbalanced datasets, such as MORPH-2, shows high performance on the original testing set. However, when evaluated with the curated balanced testset, the DNN model starts to underperform with an increase in mean absolute error from 2.21 to 7.70 years and fairness score dropping from 0.78 to 0.18. The results question the ability to compare related work by their reported results, as training and testing is done on datasets which vary in their distribution of age and protected features. When considering deployment based on related work approaches their reported results may lead to inaccurate and unfair behavior.

RQ3: Performance and fairness of SOTA academia, industry and humans. RQ2 questioned the ability to compare related work on their reported results, as the DL systems are evaluated on different and imbalanced datasets. Having identified differences in performance when evaluating on a balanced dataset, the study is extended to the two industry DL systems of AWS and Azure and one further DL system of academia based on AlexNet [18] which serves as baseline to validate our finding. In addition, we compare the results to a human study given by Agustsson et al. [1].

The results in Table 2 show the testing criteria for each DL system. Out of the four DL systems, industry is leading in performance and fairness where Microsoft Azure shows an MAE of 6.38 and a fairness score of 0.32. The worst MAE is retrieved on AlexNet, which serves as baseline given its focus on fairness [18]. One further finding is that the fairness score does not necessarily correlate with the accuracy or MAE. The reason may be that only the difference between the individual protected features, such as ethnicity, plays a role, without integrating the difference to the actual age. Thereby, in case all ethnicity are perceived equally bad, the fairness may still remain high. A good example is the baseline AlexNet with the lowest performance but the highest fairness score of 0.32. Compared to human performance all DL systems underperform with humans showing an MAE of 5.26. However, humans tend to be the second worst in fairness score.

Table 2. Balanced testset evaluation on SOTA

DL System		MAE	Accuracy (%)	1-off Accuracy (%)	Perception Distance	Fairness Score
Industry	AWS	10.57	5.55	16.89	-7.68	0.25
	Azure	6.39	6.91	20.30	-2.45	0.29
Academia	Baseline	18.01	2.21	6.31	-6.26	0.32
	DEX	7.70	6.17	20.07	0.78	0.18
Human Study		5.26	10.75	24.52	-0.10	0.20



(a) Human (b) AWS (c) Azure

(d) AlexNet (e) DEX

Fig. 6. Ethnicity specific perception on related work

> **Answer to RQ3**: Our initial finding of RQ2 is further reflected in the identified results, where also AlexNet reported an MAE on dataset APPA-REAL between 11-12 which drops to 18 when evaluated on the balanced testset. All DL systems including human are not capable of achieving the p-value in fairness score of 0.80, which emphasizes unfair behavior among the state-of-the-art. The best DL system is Microsoft Azure with an MAE of 6.28.

RQ4: Implications towards unfairness. In RQ3 we identified significant differences in performance and fairness among SOTA DL systems in both industry and academia. Now, we take a closer look at the performance and fairness towards ethnicity and gender. This will give the necessary insight to identify the implications for encountering unfair DL systems in the real-world.

Figure 6 presents the ethnicity perception for all SOTA models. We observe different behavior between age ranges, which is why we choose to assess performance and fairness for the identified ranges individually. We define the ranges from 1-30 years old and 30-70 years old. We do not integrate results after 70 years as a data shortage is experienced given the benchmark datasets. The fine-grained testing criteria results for both age ranges can be found in the appendix under Tables 5, 6.

The results for the first age range (1-30 years) show that for industry, AWS classifies age the most different between ethnicity, perceiving Caucasian faces on average 2.05 years younger than

Afro-American faces. For academia, DEX shows the highest difference in age-perception, with perceiving Caucasian faces on average 7.59 years older than Afro-American faces. Furthermore, DEX perceives female faces on average 5.41 years older than male faces. Azure and AlexNet DNN models as well as human perception do not show significant differences (below 2 years) for 1 to 30 year old age range. This behavior, however, changes dramatically for the second age range.

For range 31 to 70 years old, age is perceived between ethnicity and gender different than for the previous age range. While human perception still performs best with around 3 years of difference between both, ethnicity and gender, industry and academia DNN models differ at a much higher rate. The highest difference in industry is again found for AWS, where Afro-American faces are classified on average 5.22 years younger than Asian faces. In addition, female faces are classified 3.98 years younger than male ones. For Azure, similar tendencies are observed for gender, with female faces being classified on average 4.03 years younger. Academia DNN models again show the highest difference for ethnicity with around 6 years, however, show slightly lower difference for gender, perceiving female faces around 3 years younger.

Answer to RQ4: High differences in perceiving average age between ethnicity and gender is observed for both industry and academia DNN models. Given the fact that evaluated systems are used for policing and large-scale human affecting applications, unfair behavior may be experienced on a daily basis. Crime prevention systems tend to raise more warnings for younger people. As AWS classifies Afro-American faces 5.22 years younger than Asian faces the crime prevention system will raise more warnings for Afro-American ethnicity just because of the inaccurate age prediction service. As a consequence, a crime prevention system will treat Afro-American unfair. The results demonstrate an unfair age prediction system landscape and call for improvement.

4.4 Data-Driven Enhancements

The results of the empirical study demonstrate the need for an improvement in performance and fairness for age prediction. Having identified a large gap between reported results of related work and retrieved results by this paper based on data similar to deployment scenario the enhancement section focuses on data-driven enhancements to create DL systems with high performance and fair predictions under deployment conditions.

4.4.1 ***Distribution-Aware Balancing of Protected Features.*** The empirical study shows that DNN models behave significantly different when facing balanced and diverse testdata. One of the identified reasons is that DL systems of related work are trained on imbalanced and varying datasets, such as MORPH-2. Hence, the first enhancement is about curating a balanced and diverse training set, train a SOTA DNN model architecture and compare the testing criteria on the same balanced and diverse testing set used for the empirical study. Thereby, the impact of balanced data is compared from training perspective.

For curating the dataset, we follow Algorithm 1 and utilize all four commonly used datasets presented in Figure 4 and explained in Section 4.1. Having analyzed their individual sample-size distribution towards protected features we combine the pre-processed datasets to create one large and balanced finetuning dataset containing 122,599. The end result is illustrated in Figure 3 which is split into 80% for training and 20% for testing while maintaining balance towards ethnicity. We then compare the performance and fairness criteria of MORPH-2-based DNN model and the DNN model trained on the balanced dataset following the training procedure outlined in the experimental setup.

Table 3. Improvement of distribution-aware balancing (indicated as **Ours**) compared to related work

DL System		MAE	Accuracy (%)	1-off Accuracy (%)	Perception Distance	Fairness Score
Industry	AWS	10.57	5.55	16.89	-7.68	0.25
	Azure	6.39	6.91	20.30	-2.45	0.29
Academia	Baseline	18.01	2.21	6.31	-6.26	0.32
	DEX	7.70	6.17	20.07	0.78	0.18
Human Study		5.26	10.75	24.52	-0.10	0.20
Ours		**3.39**	**37.91**	**50.28**	**0.67**	**0.89**

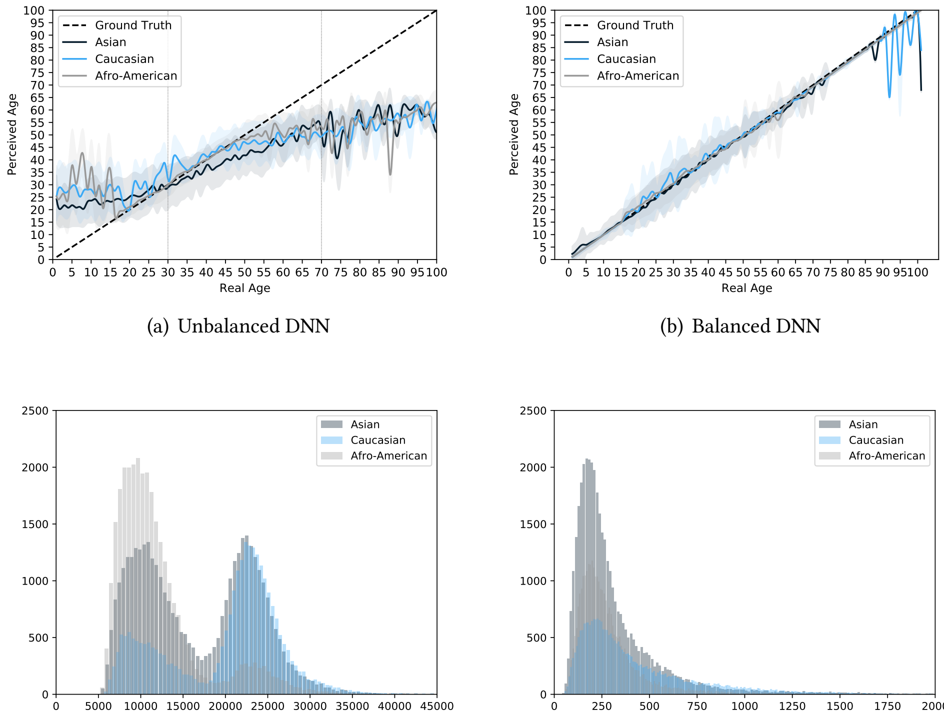
The results in Figure 7 and Table 3 show that the DNN model trained on the balanced dataset achieves an MAE of 3.39 which is a 227.24% increase in performance compared to SOTA in academia trained on imbalanced dataset with an MAE of 7.70. In addition, top-1 accuracy and 1-off accuracy improve from 6.17% to 37.91% and from 20.07% to 50.28%, respectively. Fairness wise, the perception distance from age 1 to 30 is 8.63 and from age 31 to 70 is 6.32 on the unbalanced model. For the balanced model the distance is 3.34 for age 1 to 30 and 3.53 for age 31 to 70. The overall fairness score for unbalanced and balanced DNN model are 18.68% and 87.91%, which shows a 370.61% improvement.

Effectiveness of protected feature balancing enhancement: Overall, balancing the training data shows an improvement in MAE by 227.24% when compared to the previous SOTA in academia. Compared to the best performing DL system in industry the balanced version outperforms Microsoft Azure by 188.49%. In addition, the balanced DNN model is the first system to achieve a fairness score of 87.91% above the p-value which indicates a fair system.

We further want to validate our results on balanced and imbalanced DNN model regarding the overall diversity of the datasets. For example, face images of MORPH-2 dataset are taken in portrait mode and share a similar white background style. Therefore, the diversity of the dataset is low as only one scene is captured. In a real-world scenario faces are observed in various shape, angle and style. Therefore, the diversity of MORPH-2 may be insufficient to train the DNN model towards real-world deployment settings. Analyzing the diversity for each protected feature individually, we are able to assess how similar the feature, e.g. ethnicity is perceived, which may hint to implications towards fairness.

We retrieve the OOD scores from the evaluation testset using GLOD on the imbalanced and balanced DNN model. The OOD-score distributions are visualized independently for each ethnicity in Figure 7. The results show a difference in predicting ethnicity between the balanced and unbalanced DNN model. While the balanced DNN model predict each ethnicity similarly, represented by a similar OOD-score distribution, the unbalanced DNN-Model varies in its prediction performance. Since the unbalanced version is trained in majority on Afro-American faces, it's ability to predict Caucasian and Asian faces declines, resulting in a lower certainty and therefore higher OOD-score, which is reflected by the second peak in the unbalanced distributions. Therefore, a visual assessment of two different distributions of OOD-scores helps in identifying imbalance and guiding criteria towards equal and diverse perception without the need for assessing the training data.

To statistically validate the effectiveness, we perform two T-tests [81] in order to check for significant differences between the ethnicity OOD scores. We use two-sided T-tests with a significance level of 0.95, and use Levene's test [27] if variances are not similar in order to check the following: (1) difference between Caucasian scores and Afro-American scores and (2) difference between Caucasian scores and Asian scores. The T-tests result in significant differences in both of our comparison's: (1) $T(51565.6) = 131.6, p = 0.00$ and (2) $T(56801.21) = 52.7, p = 0.00$. Those results follow our hypothesis and demonstrate that the DNN perceives ethnicity different, and



(a) Unbalanced DNN (b) Balanced DNN

(c) Unbalanced DNN (d) Balanced DNN

Fig. 7. Impact of balancing for ethnicity perception (top) and ethnicity OOD-score distribution (bottom)

thus failing the T-test. This leads to the conclusion that models that were trained on unbalanced datasets will have similar unbalanced OOD-scores distributions and are likely to result in errors after deployment making the validation technique a viable option to test DNN models, especially in scenarios where the training data is unknown.

Effectiveness of distribution-aware validation: Comparing the analysis of the models trained on unbalanced and balanced dataset demonstrates the ability to systematically quantify deployment readiness by analyzing OOD-score distribution behavior regarding protected features. Thereby, besides the improvement in performance and fairness, the presented balancing approach increases deployment-readiness, too as the DNN model handles diverse settings among protected features with similar certainty. In addition, we identify distribution-aware validation supports curating datasets for training fair and accurate DNN models by analyzing dense OOD-score areas as observed in imbalanced setting which hints towards further required data to fill the gaps of overall diversity and equal certainty in predicting data of varying protected features.

4.4.2 ***Distribution-Aware Data Augmentation***. In this study, we curate a dataset utilizing commonly used benchmark datasets for facial recognition and preserve balance among protected-features, namely ethnicity and gender to the best of our ability. We identify that available benchmark datasets do not enable a similar sample size among ages. Regardless, following Algorithm 1 a dataset is curated which integrates dataset size maximizing balance and diversity among protected features based on available data sources as shown in Figure 3. One main motivation of data augmentation is to further balance the dataset by augmenting data of minority areas with higher priority following algorithm 2. For distribution-awareness we select different OOD-score thresholds to compare the effectiveness of similar, diverse and too different or unrealistic augmentations. To measure the DNN model's ability to generalize to unknown data and settings we evaluate the retrained DNN models with the FG-NET dataset, which is unknown and thereby approximates a real-world setting after deployment.

Utilizing the balanced dataset curated in this work, we compare two common approaches in generating augmentations, namely AutoAugment [19] and random augmentation. AutoAugment integrates DNN model for feedback and thereby may have the chance to further increase diversity. Random augmentation uses manually defined parameters following related work [11, 31, 79] and randomly generates augmentations in the defined bounds. For both augmentation styles we generate 480,000 face images from which we sample 40,000 for each evaluation setting in defined regions.

Next, we retrieve the OOD scores on the balanced DNN model for all augmentations. Figure 8 presents OOD-score distribution of augmentation dataset compared to the original balanced dataset. Using different thresholds, we identify that the larger the gap between original dataset and augmented dataset, the more different and diverse augmentations become. For example, at the very right (quantile $q = 0.80$) we showcase two augmentations, with AutoAugment (top) and random augment (bottom). Both augmentations look similar to the original dataset. However, when going to the very left of the augmentation OOD-score distribution (quantile $q = 0.03$) the augmentations show a large difference to the original dataset and therefore integrate the highest diversity. Nevertheless, realism may be questioned for some augmentation results.

Having analyzed 200+ augmentations from OOD score perspective we identify three thresholds to reflect *unrealistic*, *diverse*, and *too similar* setting. For each area we sample 40,000 augmentations and retrain the DNN model in addition to the original training set. As further comparison and baseline, 40,000 augmentations are sampled randomly from the whole range without any threshold. The defined areas based on visual observation are defined as follows: For filtering too different samples the range of OOD-score quantiles 0.03 and 1 is selected. In addition, to filter samples that are too similar, an upper bound at OOD-score quantile 0.20 is defined.

We assess each area from performance, fairness, and generalization perspectives and compare AutoAugment with random augmentation approach. For performance, MAE is measured based on balanced testset. Similarly, the fairness score is retrieved from the same testset. For generalization MAE is measured on the unknown FG-NET dataset.

Table 4 shows the results of the different augmentation types in combination with OOD-score areas. From the performance perspective, all augmentation strategies slightly decrease MAE on the balanced testset and do not show any significant difference among each other, which hints towards sufficient features in the original training set to handle the balanced testset. From the fairness perspective, the score decreases by $0 - 8\%$ among augmentation settings, with AutoAugment as exception, showing a higher ability to preserve fairness among protected features. The higher fairness observed by AutoAugment hints towards its ability to produce augmentations similar to each protected feature whereas random augmentation randomly selects the augmentation influence without retrieving DNN model feedback. From the generalization perspective, the biggest influence by distribution-aware data augmentation can be observed. the results show that the MAE on the unknown dataset FG-NET improves for all augmentation settings up to 20.03%. The best setting for generalization is the one which filters out too similar and unrealistic augmentations using random augment.

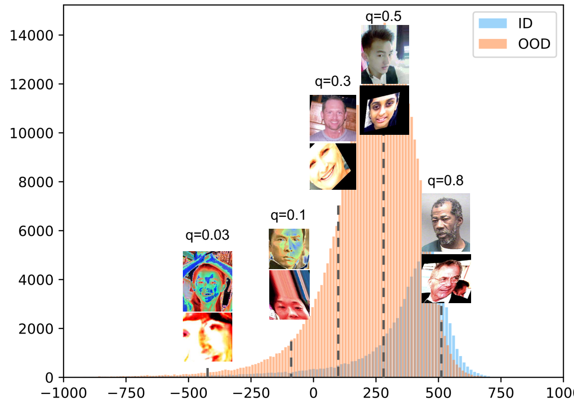


Fig. 8. Augmentation OOD-score distribution of augmentation and original dataset with examples of Auto Augment (top) and random augment (bottom)

Table 4. Results for distribution-sensitive augmentation settings (in MAE)

Testing Criteria	No Augmentation	Baseline 0-1		Unrealistic Filter 0.03-1		+ Similarity Filter 0.05-0.2	
		AutoAug	StandAug	AutoAug	StandAug	AutoAug	StandAug
Performance	**3.39**	3.62	3.64	3.66	3.70	3.64	3.74
Fairness	0.87	0.85	0.81	**0.88**	0.79	0.85	0.80
Generalization	7.55	6.84	6.93	6.57	6.55	6.57	**6.29**

Effectiveness of distribution-aware data augmentation: Overall, adding data augmentation to the training process affects performance and fairness on balanced testset slightly but significantly enhances generalization, showing the ability to predict unknown inputs at 20.03% better MAE. Furthermore, integrating distribution-awareness into the augmentation process provides the ability to filter too similar and unrealistic augmentations, which further enhances generalization by 10.38% compared to random sampling. Finally, random augmentation with defined bounds tends to slightly outperform Auto-Augment in regards to generalization, however underperforms in performance and fairness evaluation. Therefore, we conclude that Auto-Augment is good for large datasets, where the ability to generalize can be naturally higher, and random augment is suitable for smaller datasets with small diversity. Overall, distribution-aware data augmentation indicates a 10.38% increase in generalization over related augmentation approaches to create DL systems which behave at higher accuracy to unknown inputs which is often perceived in deployment settings.

4.5 Threat to Validity

The balanced testset could be a threat to validity since its complexity is limited to the public available data which integrates the necessary information on ethnicity and gender. We try to counter this by maximizing the ethnicity sample-size distribution for each age following Algorithm 1 and utilize four different datasets to enhance diversity as well as demonstrate real-world impact. OOD detection is a very challenging problem as there is no perfect ground truth, which could be a threat to validity. To this end, we assess the face images at different OOD scores for empirical validation and six DNN models with different data source and OOD score areas.

5 DISCUSSION AND RESEARCH GUIDANCE

Based on our results, we pinpoint the following research guidance:

- **Evaluation of state-of-the-art.** The empirical study shows that commonly used datasets are imbalanced towards ethnicity or gender. The reported results of related work drop in performance up to 348.42% when tested with a balanced and diverse testset. When enlarging the study to industry and an additional baseline from academia, similar low performance is retrieved on the balanced testset. We conclude as root cause that current SOTA may be trained with imbalanced and low-diversity training sets, which makes it difficult to compare related work by their reported results and that most related work is unlikely to perform as reported in their evaluation setting in a deployment setting when facing unknown and diverse data inputs.
  Research guidance: This work takes a first step towards a diverse and balanced testset among protected features for comparing related work. We identify a large gap between reported and actual results when evaluating on diverse and balanced setting. A possible direction is to establish a common facial recognition dataset benchmark for fair evaluation and comparison. Considering the potential risk in deploying facial recognition applications, a benchmark testset provides a standardized way to compare most recent advancements and analyze which ones benefit performance and fairness. Considering the masses of data private and public sector have acquired, it is likely that necessary resources exist to curate such dataset.
- **Implications of unfairness** Our results show that the existing DNN systems for age prediction have significantly different results towards performance and fairness between ethnicity and gender. None of the evaluated DNN models can be considered fair, with the highest fairness score by AlexNet of 0.32. We observe that on average female faces are classified 4 years younger (Azure) and ethnicity are perceived with 5 years difference (AWS). Facial recognition services, such as Amazon AWS are used by companies and governments around the globe [5]. When policing warning systems use age prediction services, differences in age play a large role as, e.g., younger people tend to commit more crime. Thereby, the relevance of fairness should be high in DL systems which however is currently not sufficiently addressed in industry and academia.
  Research guidance: Following the research direction of the empirical evaluation, quantitative standards may be developed to verify that facial recognition systems meet quantified fairness thresholds and therefore classify for sensitive applications. To further stimulate the field towards performance, fairness and generalization pre-trained DNN models may be publicly accessed, which are trained on large amounts of balanced and diverse faces following presented enhancements of this work. In natural language processing domain (NLP) this method has shown already great success in almost all use cases with BERT [22] or GPT-3 [13] which enhanced various applications as quality level foundation. Similarly, in facial recognition DNN models could be accessed via public libraries, e.g., Keras or Pytorch, and made available for individual fine-tuning for applications. This would enable a high baseline for fairness and performance and could especially help applications with data shortages.
- **Effectiveness of enhancements.** Given the presented SOTA in academia, balancing the training data improved MAE by 227.24% and achieved a 4-fold increase in fairness. This improvement is visualized and compared to SOTA in Figure 9.
  Furthermore, the effectiveness of distribution-awareness using OOD-techniques is demonstrated through integration in the balancing efforts, ensuring that protected features are predicted similarly and are trained with similar diversity which can be statistically validated using T-test.

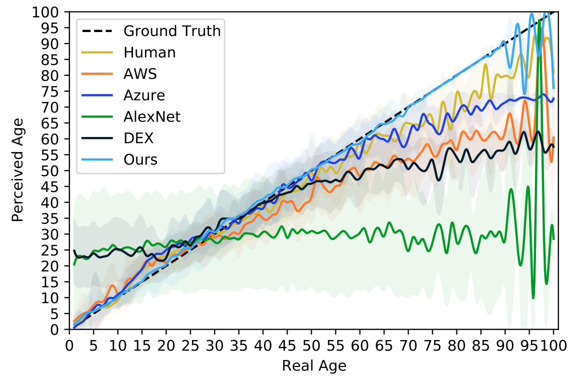


Fig. 9. Comparison of enhancement approach to SOTA and human perception

Thereby, OOD serves as a guiding, control and validation measure for data diversity and DNN model perception, especially for those cases where the training data is unknown.
Despite our high balancing efforts, not all ages are able to represent a similar amount of protected features. Therefore, we identify data augmentation as systematic approach to further stimulate balancing efforts. In addition, we closely analyze OOD-score distribution diversity behavior and identify that distributions with too high and too low OOD score do not benefit generalization as they are perceived either too similar or unrealistic. Hence, OOD detection enables filtering such negatively impacting augmentations to maximize diversity and improve generalization. Overall, we are able to increase the ability to generalize on unknown data with e.g. new scenes by 20.03%.
Research guidance: A possible direction is further integrating OOD score in the augmentation process. Currently OOD detection serves as a filtering mechanism. Instead, it may serve augmentation generation as a guiding mechanism, which controls the diversity and integrates direct DNN model feedback in the approach. Thereby, we assume to produce augmentations with highest possible diversity while saving storage.
Commercial guidance: Our data-driven enhancement can be taken as guidance for developing fair commercial face recognition services. For development, given a dataset, it is important to analyze the sample-size distribution among protected features and balance the dataset among these features while maintaining maximum diversity levels. Furthermore, to validate the diversity of the dataset and the balancing operation, diversity analysis helps to analyze if the DNN model is able to predict protected features with similar certainty and learned similar diverse data representations. Due to the difficulty in perfectly balancing and maximizing diversity of a dataset, missing data can be identified with data augmentation. Using OOD-techniques the most effective augmentations can be identified to further enhance the ability to generalize. After retraining with augmentations, the DNN model can be reassessed with a distribution validation to verify that OOD-score distributions of protected features show similar diversity and certainty. Finally, a new diverse dataset should be retrieved to test the DNN model's ability to generalize when encountering unknown scenes which is likely to be the case after deployment.

6 CONCLUSION

In this paper, we conduct a large-scale empirical study on the state-of-the-art (SOTA) DL-based age prediction systems from both academia as well as industry and include a human study as comparison, too. We use the identified strengths and weakness of current SOTA to present distribution-aware balancing and data augmentation to enhance performance and fairness. Our results show that the existing DL systems vary in performance with the best system below human performance of 5.26 years of average error. In addition, ethnicity and gender are predicted on average up to 7.59 years differently which has consequences in fair treatment for human-centric applications which use age prediction services. Utilizing publicly available benchmark datasets and the proposed distribution-aware balancing and data augmentation methodology, we are able to enhance the SOTA DL systems from predicting age with a mean absolute error from 7.70 to 3.39 years. Furthermore, the enhancements enable the presented DL system to predict age similar between ethnicity and gender, showing a 4-fold increase in fairness score. This work makes one of the first steps in age prediction to integrate fairness into the enhancement process, which enables as well as validates fair and high performing system in the human-centric field of facial recognition.

7 ACKNOWLEDGMENTS

We thank Professor Lei Ma from Kyushu University, Japan and Dr. Xiaofei Xie from Nanyang Technological University, Singapore for the helpful discussions. We also thank Professor Xavier Bresson from Nanyang Technological University, Singapore for his guidance during the idea development stage. This work was supported by Singapore Ministry of Education Academic Research Fund Tier 1 (Award No. 2018-T1-002-069), the National Research Foundation, Prime Ministers Office, Singapore under its National Cybersecurity R&D Program (Award No. NRF2018NCR-NCR005-0001), the Singapore National Research Foundation under NCR Award Number NSOE003-0001 and NRF Investigatorship NRFI06-2020-0022. We also gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC) to our research.

REFERENCES

[1] E Agustsson, R Timofte, S Escalera, X Baro, I Guyon, and R Rothe. 2017. Apparent and real age estimation in still images with deep residual regressors on APPA-REAL database.. In *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE.

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

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

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

[5] Amazon. 2020. "Amazon Rekognition Customers". Retrieved June 12 2020 from https://blog.aboutamazon.com/policy/we-are-implementing-a-one-year-moratorium-on-police-use-of-rekognition

[6] Amazon. 2020. "The Amazon blog: We are implementing a one-year moratorium on police use of Rekognition". Retrieved June 12 2020 from https://aws.amazon.com/rekognition/customers/

[7] Guy Amit, Moshe Levy, Ishai Rosenberg, A. Shabtai, and Yuval Elovici. 2020. GLOD: Gaussian Likelihood Out of Distribution Detector. *arXiv preprint arXiv:2008.06856*.

[8] Theodore W Anderson. 1958. *An introduction to multivariate statistical analysis*. Technical Report.

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

[10] BBC. 2019. "Apple AI accused of leading to man's wrongful arrest - BBC News". Retrieved June 16 2020 from https://www.bbc.com/news/technology-48022890

[11] David Berend, Xiaofei Xie, Lei Ma, Lingjun Zhou, Yang Liu, Chi Xu, and Jianjun Zhao. 2020. Cats Are Not Fish: Deep Learning Testing Calls for Out-Of-Distribution Awareness. In *The 35th IEEE/ACM International Conference on*

Automated Software Engineering. Association for Computing Machinery, New York, NY, USA.
[12] Dan Biddle. 2006. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd.
[13] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html
[15] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao. 2017. Using Ranking-CNN for Age Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 742–751.
[16] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. 2017. Using ranking-cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5183–5192.
[17] Hyunsun Choi and Eric Jang. 2019. Generative Ensembles for Robust Anomaly Detection. https://openreview.net/forum?id=B1e8CsRctX
[18] A. Clapés, G. Anbarjafari, O. Bilici, D. Temirova, E. Avots, and S. Escalera. 2018. 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)*. 2436–243609.
[19] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. 2018. AutoAugment: Learning Augmentation Policies from Data. *CoRR* abs/1805.09501 (2018). arXiv:1805.09501 http://arxiv.org/abs/1805.09501
[20] Abhijit Das and Antitza Dantcheva. 2018. Mitigating Bias in Gender, Age, and Ethnicity Classification: a Multi-Task Convolution Neural Network Approach.
[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805
[23] Mingxing Duan, Kenli Li, Canqun Yang, and Keqin Li. 2018. A Hybrid Deep Learning CNNELM for Age and Gender Classification. *Neurocomput.* 275, C (Jan. 2018), 448–461. https://doi.org/10.1016/j.neucom.2017.08.062
[24] E. Eidinger, R. Enbar, and T. Hassner. 2014. Age and Gender Estimation of Unfiltered Faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179.
[25] FaceWatch. 2020. "Facial recognition security system UK | Facewatch". Retrieved June 12 2020 from https://www.facewatch.co.uk/
[26] Yun Fu, Guodong Guo, and Thomas S Huang. 2010. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence* 32, 11 (2010), 1955–1976.
[27] Joseph Gastwirth, Yulia Gel, and Weiwen Miao. 2010. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statist. Sci.* 24 (10 2010). https://doi.org/10.1214/09-STS301
[28] Tobias Gehrig, Matthias Steiner, and Hazım Kemal Ekenel. 2011. Draft: evaluation guidelines for gender classification and age estimation. *Technical report, Karlsruhe Institute of Technology* (2011).
[29] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. 2007. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence* 29, 12 (2007), 2234–2240.
[30] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang. 2009. Human age estimation using bio-inspired features. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 112–119.
[31] Jianmin Guo, Yu Jiang, Yue Zhao, Quan Chen, and Jiaguang Sun. 2018. 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* (Lake Buena Vista, FL, USA) *(ESEC/FSE 2018)*. Association for Computing Machinery, New York, NY, USA, 739–743. https://doi.org/10.1145/3236024.3264835
[32] Qing Guo, Xiaofei Xie, Lei Ma, Zhongguo Li, Wei Feng, and Yang Liu. 2019. Spatial-aware Online Adversarial Perturbations Against Visual Object Tracking. *arXiv* (2019). arXiv:arXiv:1910.08681v1
[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385
[34] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *CoRR* abs/1610.02136 (2016). arXiv:1610.02136 http://arxiv.org/abs/1610.02136

[35] Dan Hendrycks and Kevin Gimpel. 2017. 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*.

[36] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. 2019. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HyxCxhRcY7

[37] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[38] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR* abs/1608.06993 (2016). arXiv:1608.06993 http://arxiv.org/abs/1608.06993

[39] Lingxiao Huang and Nisheeth K Vishnoi. 2019. Stable and fair classification. *arXiv preprint arXiv:1902.07823* (2019).

[40] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. 2020. FakePolisher: Making DeepFakes More Detection-Evasive by Shallow Reconstruction.

[41] IBM. 2020. "IBM CEO's Letter to Congress on Racial Justice Reform". Retrieved June 12 2020 from https://www.ibm.com/blogs/policy/facial-recognition-susset-racial-justice-reforms/

[42] Diaconu Ion, Augusto Lindgren Alves José, Malhotra Ajai, Omar Mona, Salmón Elizabeth, Seetulsingh Dheerujlall, and Hennebel Ludovic. 2019. "A global call for the total elimination of racism, racial discrimination, xenophobia and related intolerance". https://www.ohchr.org/EN/HRBodies/HRC/AdvisoryCommittee/Pages/GlobalcallTotalEliminationRacism.aspx

[43] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. An efficient k-means clustering algorithms: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2002). https://doi.org/10.1109/TPAMI.2002.1017616

[44] Dr.Hlaing Htake Khaung Tin. 2013. How Old Are You?: Age Prediction using Eigen Face.

[45] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. 2019. Learning Not to Learn: Training Deep Neural Networks With Biased Data. 9004–9012. https://doi.org/10.1109/CVPR.2019.00922

[46] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[47] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*. 853–862.

[48] Young H Kwon and Niels da Vitoria Lobo. 1999. Age classification from facial images. *Computer vision and image understanding* 74, 1 (1999), 1–21.

[49] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6402–6413. http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf

[50] A. Lanitis, C. J. Taylor, and T. F. Cootes. 2002. Toward Automatic Simulation of Aging Effects on Face Images. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 4 (April 2002), 442–455. https://doi.org/10.1109/34.993553

[51] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. 2018. Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ryiAv2xAZ

[52] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 7167–7177. http://papers.nips.cc/paper/7947-a-simple-unified-framework-for-detecting-out-of-distribution-samples-and-adversarial-attacks.pdf

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

[54] Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1VGkIxRZ

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

[56] Markets and Markets. 2020. "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". Retrieved June 12 2020 from https://www.marketsandmarkets.com/Market-Reports/facial-recognition-market-995.html

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]

[58] Microsoft. 2020. "Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM". Retrieved June 16 2020 from https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/

[59] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. 2020. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. https://openreview.net/forum?id=r1lnxTEYPS

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

[61] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alchebuc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[62] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, and Jianjun Zhao. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms.

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

[64] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. 2019. Likelihood Ratios for Out-of-Distribution Detection. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 14707–14718. http://papers.nips.cc/paper/9611-likelihood-ratios-for-out-of-distribution-detection.pdf

[65] K. Ricanek and T. Tesafaye. 2006. MORPH: a longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 341–345.

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

[67] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision* 126, 2-4 (2018), 144–157.

[68] Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. 2017. Inclusivefacenet: Improving fañe attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193* (2017).

[69] Ignacio Serna, Alejandro Peña, Aythami Morales, and Julian Fierrez. 2020. InsideBias: Measuring Bias in Deep Networks and Application to Face Gender Biometrics.

[70] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. 2020. Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SyxIWpVYvr

[71] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[72] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-01424-7_27

[73] Engkarat Techapanurak and Takayuki Okatani. 2019. Hyperparameter-Free Out-of-Distribution Detection Using Softmax of Scaled Cosine Similarity. *CoRR* abs/1905.10628 (2019). arXiv:1905.10628 http://arxiv.org/abs/1905.10628

[74] Hlaing Htake Khaung Tin. 2012. Subjective age prediction of face images using PCA. *International Journal of Information and Electronics Engineering* 2, 3 (2012), 296.

[75] Jan E. Trost. 1986. Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology* (1986). https://doi.org/10.1007/BF00988249

[76] Sachin Vernekar, Ashish Gaurav, Taylor Denouden, Buu Phan, Vahdat Abdelzad, Rick Salay, and Krzysztof Czarnecki. 2019. Analysis of Confident-Classifiers for Out-of-distribution Detection. *CoRR* abs/1904.12220 (2019). arXiv:1904.12220 http://arxiv.org/abs/1904.12220

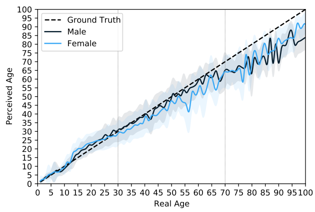
[77] J. Wan, Z. Tan, Z. Lei, G. Guo, and S. Z. Li. 2018. Auxiliary Demographic Information Assisted Age Estimation With Cascaded Structure. *IEEE Transactions on Cybernetics* 48, 9 (2018), 2531–2541.

[78] Run Wang, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yihao Huang, Jian Wang, and Yang Liu. 2020. FakeSpotter: A Simple yet Robust Baseline for Spotting AI-Synthesized Fake Faces. https://doi.org/10.24963/ijcai.2020/476 arXiv:1909.06122

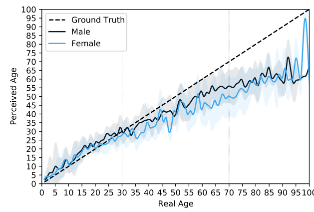
[79] Xiaofei Xie, Lei Ma, Felix Juefei-Xu, Minhui Xue, Hongxu Chen, Yang Liu, Jianjun Zhao, Bo Li, Jianxiong Yin, and Simon See. 2019. DeepHunter: A Coverage-Guided Fuzz Testing Framework for Deep Neural Networks. In *Proceedings*

of the 28th ACM SIGSOFT International Symposium on Software Testing and Analysis (Beijing, China) *(ISSTA 2019)*. Association for Computing Machinery, New York, NY, USA, 146–157. https://doi.org/10.1145/3293882.3330579
- [80] Dong Yi, Zhen Lei, and Stan Z Li. 2014. Age estimation by multi-scale convolutional network. In *Asian conference on computer vision*. Springer, 144–158.
- [81] Sandy Zabell. 2008. On Student's 1908 Article "The Probable Error of a Mean". *Journal of The American Statistical Association - J AMER STATIST ASSN* 103 (03 2008), 1–7. https://doi.org/10.1198/016214508000000030
- [82] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.
- [83] Chao Zhang, Shuaicheng Liu, Xun Xu, and Ce Zhu. 2019. C3AE: Exploring the limits of compact model for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12587–12596.
- [84] Ke Zhang, Na Liu, Xingfang Yuan, Xinyao Guo, Ce Gao, Zhenbing Zhao, and Zhanyu Ma. 2019. Fine-grained age estimation in the wild with attention LSTM networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).
- [85] Peixin ZHANG, Jingyi WANG, Jun SUN, Guoliang DONG, Xinyu WANG, Xingen WANG, Jin Song DONG, and Dai TING. 2020. White-box fairness testing through adversarial sampling. (2020).
- [86] Peixin ZHANG, Jingyi WANG, Jun SUN, Guoliang DONG, Xinyu WANG, Xingen WANG, Jin Song DONG, and Dai TING. 2020. White-box fairness testing through adversarial sampling. (2020).
- [87] Yunxuan Zhang, Li Liu, Cheng Li, and Chen Change Loy. 2017. Quantifying Facial Age by Posterior of Age Comparisons. In *British Machine Vision Conference (BMVC)*.
- [88] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.
- [89] Haiping Zhu, Qi Zhou, Junping Zhang, and James Z Wang. 2018. Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression. *arXiv preprint arXiv:1804.02740* (2018).

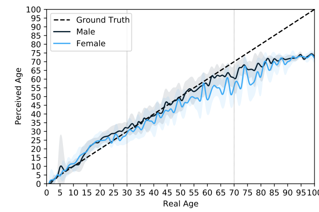
APPENDIX



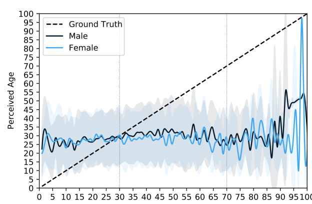
(a) Human



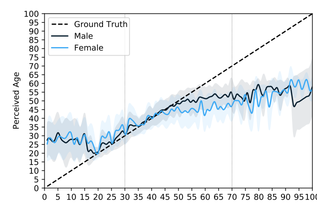
(b) AWS



(c) Azure



(d) AlexNet



(e) DEX

Fig. 10. Gender specific perception

Table 5. Age Range 1 - 30 years - Testing Criteria Results for SOTA Analysis on Ethnicity and Gender

		MAE	Acc	1-off Acc	Perc.
Human	Asian	4.35	13.71%	29.38%	0.66
	Caucasian	4.09	14.17%	31.46%	1.79
	Afro-American	4.35	11.54%	29.86%	1.91
	Male	3.76	15.66%	34.28%	1.61
	Female	4.41	12.78%	28.79%	1.72
AWS	Asian	4.12	8.52%	30.33%	0.60
	Caucasian	3.45	12.85%	37.08%	2.02
	Afro-American	5.50	10.29%	30.88%	4.51
	Male	4.38	8.98%	34.24%	2.77
	Female	3.82	12.66%	33.51%	1.58
Azure	Asian	3.38	9.07%	33.17%	1.27
	Caucasian	4.18	7.03%	29.25%	3.12
	Afro-American	3.69	13.49%	32.29%	2.96
	Male	3.96	8.12%	27.77%	2.96
	Female	3.53	11.71%	35.79%	1.89
AlexNet	Asian	14.42	3.59%	9.52%	8.94
	Caucasian	13.19	2.95%	8.33%	8.11
	Afro-American	14.50	1.61%	4.54%	6.46
	Male	13.49	2.37%	6.69%	6.87
	Female	14.72	2.22%	6.28%	9.18
DEX	Asian	10.56	2.32%	8.04%	8.65
	Caucasian	10.32	5.66%	18.64%	9.72
	Afro-American	2.96	13.39%	42.95%	1.71
	Male	5.75	10.88%	34.71%	4.83
	Female	11.16	4.25%	15.01%	10.37

Table 6. Age Range 31 - 70 years - Testing Criteria Results for SOTA Analysis on Ethnicity and Gender

		MAE	Acc	1-off Acc	Perc.
Human	Asian	6.93	5.59%	14.05%	-4.60
	Caucasian	6.55	6.29%	15.90%	-1.91
	Afro-American	7.79	4.61%	11.95%	-5.17
	Male	6.09	6.61%	16.77%	-0.90
	Female	7.18	5.78%	14.51%	-3.51
AWS	Asian	9.35	1.96%	8.31%	-7.58
	Caucasian	9.36	4.67%	11.68%	-8.16
	Afro-American	13.33	0.97%	5.31%	-12.80
	Male	8.70	4.15%	11.87%	-7.30
	Female	12.28	1.68%	5.67%	-11.24
Azure	Asian	5.55	5.21%	18.48%	-3.55
	Caucasian	6.18	7.81%	18.34%	-0.45
	Afro-American	5.05	6.55%	17.48%	-2.49
	Male	5.08	7.32%	19.09%	-0.97
	Female	7.06	4.80%	15.11%	-5.00
AlexNet	Asian	21.86	1.08%	3.79%	-18.03
	Caucasian	19.72	1.36%	4.46%	-16.26
	Afro-American	17.48	2.92%	7.89%	-12.05
	Male	17.98	2.39%	6.89%	-13.25
	Female	20.42	1.42%	4.12%	-16.76
DEX	Asian	8.21	4.28%	12.00%	-6.57
	Caucasian	6.69	5.92%	19.82%	-1.62
	Afro-American	3.66	9.67%	32.88%	-0.79
	Male	4.39	8.40%	28.70%	-0.59
	Female	7.40	6.06%	19.67%	-2.97