CAST: Conditional Attribute Subsampling Toolkit for Fine-grained Evaluation

Wes Robbins*†1 Steven Zhou*†1 Aman Bhatta†2 Chad Mello†1
Vítor Albiero†2 Kevin W. Bowyer†2 Terrance E. Boult†1

[1]University of Colorado, Colorado Springs
[2]University of Notre Dame

{wrobbins,szhou,cmello}@uccs.edu, tboult@vast.uccs.edu, {abhatta,valbiero,kwb}@nd.edu

Abstract

Thorough evaluation is critical for developing models that are fair and robust. In this work, we describe the Conditional Attribute Subsampling Toolkit (CAST) for selecting data subsets for fine-grained scientific evaluations. Our toolkit efficiently filters data given an arbitrary number of conditions for metadata attributes. The purpose of the toolkit is to allow researchers to easily to evaluate models on targeted test distributions. The functionality of CAST is demonstrated on the WebFace42M face Recognition dataset. We calculate over 50 attributes for this dataset including race, image quality, facial features, and accessories. Using our toolkit, we create over a hundred test sets conditioned on one or multiple attributes. Results are presented for subsets of various demographics and image quality ranges. Using eleven different subsets, we build a face recognition 1:1 verification benchmark called C11 that exclusively contains pairs that are near the decision threshold. Evaluation on C11 with state-of-the-art methods demonstrates the suitability of the proposed benchmark. The toolkit is publicly available at `https://github.com/WesRobbins/CAST`.

1. Introduction

Benchmark datasets such as Imagenet for image classification [17], MS-COCO for object detection [33], and IJB-C for face recognition [36] have been pivotal for the progress of deep learning. By using standardized metrics and datasets, researchers can quickly compare methods and training sets. However, benchmarks often only capture a narrow view of performance. For example, improving a benchmark score does not answer the following: *Which classes or distributions did the model perform better on? Was performance sacrificed on other classes or distributions?* Answering such questions can be critical for estimating the performance and fairness in real-world conditions. Thus, it is worth considering during model development. A first step toward more detailed evaluations can be to examine the performance of each class with statically significance testing with multiple

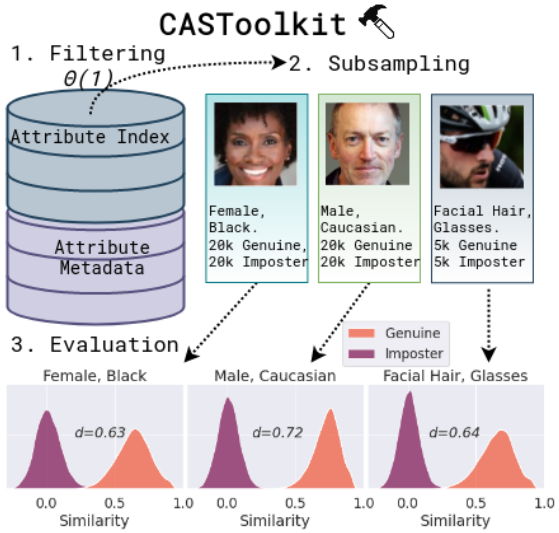

Figure 1. **C**onditional **A**ttribute **S**ubsampling **T**oolkit (CAST) is our open-source tool for train and test set sampling conditioned on data attributes. CAST provides efficient filtering and sampling from large datasets and automatic evaluation for face recognition—allowing for easy comparisons between attributes, such as between demographics The plots show genuine and imposter distributions from subsets generated by CAST. Lower distance between distributions is worse.

*Co-first Author

†This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100003]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

sets. In addition, it may be desirable to understand performance between test distributions that are not strictly captured by a class label. Such test distributions may relate to properties such as image quality, lighting conditions, or other metadata that can be task specific (such as race in face recognition). As the number of properties (or combination of properties) increases, it can be burdensome to generate and evaluate performance differences for each evaluation set.

To aid with the generation of test sets, we develop the **C**onditional **A**ttribute **S**ubsampling **T**oolkit (CAST). CAST creates evaluation sets by conditioning sampling over requested attributes. By using a pre-built index for metadata, CAST can efficiently subsample million-scale datasets to create hundreds of test sets. For face recognition, CAST also includes pre-built test sets and an evaluation module. While some research tools are complicated to use, CAST is designed to minimize overhead for researchers. By simply providing a path to the data and metadata, evaluation sets are created and saved to a structured directory. Figure 1 shows an overview of the functionality of CAST. Detailed workings of the toolkit are described in Section 3.

Along with CAST, we implement a probabilistic approach for inter-category evaluation. Inspired by classic works on softmax as probability [12, 20], network softmax outputs are used as probabilities for evaluation categories. Probabilistic inter-category evaluation attributes the performance of each sample relative to the sample's probability of belonging to a test property. For example, if we are comparing model performance between high- and low-quality samples (scored by some numerical metric), a sample between two bins would partially contribute to each of these bins. The benefit of probabilistic inter-category evaluation is that more samples can be used to evaluate each category. The probabilistic inter-category evaluation method is provided to supplement standard disjoint test sets, rather than replace them; and both are implemented within CAST. Further discussion and formal implementation details can be found in Section 4.

To demonstrate the functionality of CAST, we use the WebFace42M face recognition dataset [72]. This work focuses on face recognition with CAST because there is an ongoing need for fine-grained evaluation between demographic groups to understand the fairness of face recognition models. An additional contribution of this work is an extensive evaluation of face recognition models on the evaluation sets generated by CAST. On the WebFace42M dataset—which contains 42 million images—we calculate over 50 attributes for each image using numerous open-source repositories. The attributes collected include race, sex, age, accessories, blind image-quality, and face-image quality. A full list of the collected attributes can be found in Section 3. These attributes are passed to CAST to create numerous test sets. Figure 1 highlights the genuine and imposter distributions on three different test sets generated by CAST. In our evaluation with new subsets and previous benchmarks, we observe that most face-pairs (both impostor and genuine) are far from a decision threshold, and thus there is near zero risk of misclassifying these pairs for modern networks. Motivated by this observation, we filter datasets created with CAST to only include hard pairs, thus saving the time and compute cycles wasted on trivial comparisons. Using only near-threshold 1:1 verification pairs, we create a benchmark titled **CAST-11 (C11)**, which contains 11 sub-benchmarks with different attribute categories. In Section 5, we provide results on numerous CAST test sets including an evaluation on C11 with several models.

In summary, this work makes the following contributions:

- Provides an open-source toolkit for subsampling data to create training or evaluation sets conditioned on any number of numerical or categorical metadata fields.
- Proposes Probabilistic Inter-Category Evaluation as a supplemental approach for inter-category performance comparisons (e.g., between demographics).
- Presents extensive attribute and performance evaluation on the WebFace42M face recognition dataset.
- Creates the CAST-11 (C11) face recognition benchmark, which only contains hard verification pairs.

2. Related Work

Our toolkit aims to provide a system to conditionally sample over datasets in order to create test sets to allow for better improvement on models with newfound information on inter-category performances. Ideas and practices surrounding subsets have been implemented [71] [70] before. These studies found an improvement in facial recognition with the implementation of their methods, but both preformed analysis on smaller, controlled datasets. [18] expands upon this by implementing a system to sort larger scale datasets into two classes, clean and noisy, with the usage of a new sub-center. By dropping noisy samples, they achieved comparable performance compared to a manually cleaned dataset. Similarly, [32] produced a toolkit that has the ability to correct imbalanced datasets with different sampling techniques. Our toolkit implements a way to create balanced datasets with specified traits of images in each on a much larger scale than these toolkits and with more versatility in attribute selection. While CAST allows for benchmark analysis through the creation of subsets, tools like [61] [63] [42] provide ways to analyze the effects of changes in the evaluation pipeline in order to more efficiently run benchmark studies.

The accuracy of face recognition models is tested on several benchmark datasets in order to ensure the robustness of these systems. These benchmark datasets are usually constructed to encompass a wide variety of variations that could be prevalent in real-world operational scenarios. The benchmark datasets like CFP [50], AgeDB [41], LFW [27], CPLFW [68], CALFW [69] are designed to evaluate recog-

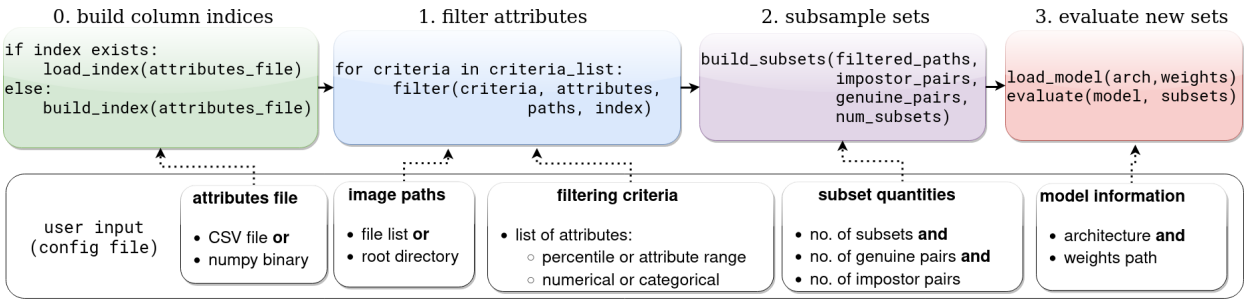

Figure 2. A flowchart of CAST functionality. The top row shows the main functionality of each step with pseudo-code. The bottom row show the user input, which is read from a configuration file. The entire functionality can be run by calling the `build_and_evaluate.py` script.

nition accuracy mostly for varying pose and age intra-class distributions. Other larger benchmark datasets such as IJB-C [35] and MegaFace [29] include the images with varying poses, illumination, demographics, quality, etc. and thus, are designed to evaluate the models for more operationally realistic and large scale recognition accuracy. None of these datasets contain multiple test subtests for computing the statistical significance of difference. Different from previous benchmarks, our proposed C11 benchmark only has pairs that are near the genuine vs. impostor decision boundary and enough subsets for statistical comparison.

Though recent advances in deep learning have enabled FR systems to achieve higher scores on several performance metrics across several benchmark datasets, face recognition accuracy differences across demographics is prevalent and widely acknowledged by academic researchers. The general consensus across several research results is that the FR accuracy is worse for females, young, and black/darker skin toned cohorts at a fixed global threshold [1, 9, 24, 30, 22, 45, 34, 53, 59, 23, 5, 31]. Past researchers have speculated causes such as the use of cosmetics [30, 34, 15], more varied hairstyles [5], or differences in average height, leading to non-optimal camera angle [15, 24] for the varying FR accuracy across demographics. Since the advent of deep learning, imbalanced training data is often suggested as the go-to cause [59, 21, 38]. Few works attempt to mitigate the differences by using one of three methods: improved algorithm and training pipeline [54, 2, 52, 56, 57, 16], balanced training datasets [59, 62, 49], and dynamic decision thresholding across demographics [58, 47]. These accuracy differences across demographics are well known, but there is relatively little work that attempts to identify the cause or causes [5, 3, 7, **?**, 10, 64] and thus, substantial research effort is required to understand the causes for the accuracy differences across different demographics groups.

3. CASToolkit

Conditional **A**ttribute **S**ubsampling **T**oolkit (CAST) is our tool for subsampling data. CAST can be used to subsample for training or evaluation, and we focus on using CAST for evaluation sets in this work. The three qualities that make CAST advantageous for a researcher are that it is fast, extensible and easy to use.

Usability and Extensibility. Any array with rows of images and columns of attributes can be used – csv and numpy are easily used. A user then specifies subsets and generates files for experiments. A user can import a PyTorch model to easily run our C11 benchmark consisting of 11 separate validation sets; or create a new benchmark based on attributes of the user's choice.

Speed. For WebFace42M, a pre-built index over 50+ attributes is provided, which allows filtering to be performed in $\mathcal{O}(1)$. If new attributes or data are provided, an index is automatically created, allowing subsequent attribute filtering to be done in constant time.

In the remainder of this section we provide details on our adoption of the WebFace42M dataset for building subsets (Section 3.1), description of the attributes collected for WebFace42M (Section 3.2), filtering & subsampling implementation details (Section 3.3), and last we introduce the Cast-11 (C11) benchmark which is provided as an extension of CAST (Section 3.4).

3.1. WebFace42M Dataset

WebFace42M is a large scale dataset for face recognition with 42 million images and 2 million identities [72]. Due to the size and training cost of WebFace42M, prior work (including the original paper) [8, 72] have presented results on two subsets of WebFace42M: WebFace4M and WebFace12M. These subsets have respectively 4 million and 12 million images, and WebFace12M is a superset of WebFace4M. To create test sets, we only use images and identities that are not included in WebFace4M/12M. Thus, WebFace4M/12M are still viable training sets for the evaluation sets created in the work. The relationship between different versions of WebFace is visualized in Figure 4. For our experiments, some of our models are trained on WebFace4M.

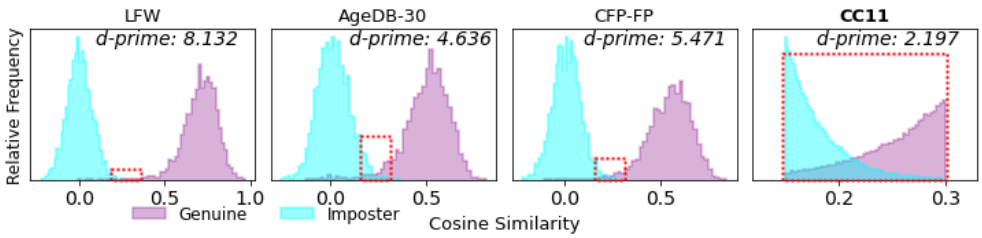

Figure 3. Rather than spending compute on 95+% trivial 1:1 verification pairs, the C11 benchmark only contains hard pairs. In the left three plots, it can be seen previous benchmarks contain a majority easy pairs, some hard pairs (in red box), and some impossible pairs. The impossible pairs are pairs that are being misclassified and are far from the threshold (e.g., the purple area around 0.0 for AgeDB-30). C11 filters out easy and impossible pairs, and thus only contains hard pairs. The red box highlight pairs that fall with in a cosine similarity of 0.15–0.30, which is the hard pair range. The abrupt edges of the C11 impostor and genuine distributions is due to removing pairs outside of the hard pair range. The C11 benchmark has 110,000 pairs.

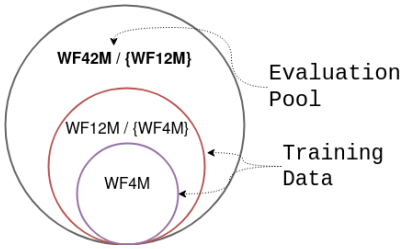

Figure 4. The data used for training and evaluation in this work. The WebFace12M and WebFace4M (a subset of 12M) sets are held out for training. The remaining 30 million images are used as a pool for sampling evaluation sets.

3.2. Attribute Calculation

Using a selection of methods described below, we create an array of attribute values for each image. Values in the first 40 columns are generated by AFFACT [25] which scores facial attributes (e.g., mustache, sideburns, glasses, attractiveness, etc.). We then use three ResNet-50 models from [6] to predict race, age, and gender. A ResNet50 model trained on AFAD [43], IMFDB [51], MegaAsian [66], MORPH3 [46], and UTKFace datasets [67] predicted race. A ResNet-50 model trained on AAF [14], AFAD, AgeDB [41], CACD [13], IMDB-WIKI [48], IMFDB, MegaAgeAsian, MORPH3, and UTKFace datasets predicted gender, and age was predicted by a ResNet-50 model trained on AAF, AFAD, AgeDB, IMDB-WIKI, IMFDB, MegaAgeAsian, MORPH3, and UTKFace datasets. We help verify and compare these values with regression values from FairFace [28] which gives a model confidence score for seven races, binary genders, and age ranges. For FairFace, we resize the WebFace images to 224x224. The softmax operation is then performed over each output category (e.g., 7 races, 2 genders, 9 age bins). Race and Gender information calculated with FairFace can be found in Figure 5. Finally, we use BRISQUE [40], NIMA [55], Paq2Piq [65] to give general image quality scores and SDD-FIQA [44], CR-FIQA [11], and MagFace[1] [37] to provide Face-Image quality assessments. img2pose [4] is then run which regresses 6Dof and estimates a 3D pose for the face in the image. These resulting values are concatenated to form a general attribute array. We create a sorted list of indexes for the array to aid in selecting bounds when filtering. A full list of attributes and their correlation are in the supplemental material.

3.3. Implementation Details

Users first declare the number of subsets they would like to create. From there, the attribute value array, path array, and the indexes of the sorted attribute array are loaded in as shown in the blue section of Figure 2. The user picks an attribute to filter by as shown in the green-colored section of Figure 2. If the attribute is numerical, a range can be selected based directly on attribute value or with a percentile range. For attributes such as age, absolute attribute values may make more sense than the percentile range. If a percentile range is chosen, the percentiles are then translated into indices that bound the percentile range (filtering in $\mathcal{O}(1)$). The upper and lower indices are used to create a mask as shown in the red section of Figure 2. If the option for filtering is based on attribute value (as opposed to percentile range), the array is searched to find the beginning and end of the requested range (filtering in $\mathcal{O}(n \log n)$). Users also have the ability to view reference images that hold similar scores relative to the declared bounds. If the attribute is scored with discrete, classification values, then the user is asked which class to filter for, and then a boolean statement setting the array equal to the class is created. The toolkit can filter by many attributes as the booleans of each specified attribute are combined with logical and. All bounds and classes used are logged into a .txt file. A mask with a value of True at indexes where images fit the boolean logic and False otherwise is

[1]The MagFace model used for quality assessment scores is a pre-trained network from the MagFace repository, which was trained on MS1Mv2. This is separate from the MagFace network we train for evaluation. The MagFace model is retrained for evaluation in order to match evaluation model settings.

created. This mask is paired with the paths list to return the paths of the images which fit the boolean statements. The user inputs the number of images they would like in their dataset and the image paths are randomly sampled into a .list file. The result is a directory with the specified number of datasets in the form of .list files and a .txt file describing the bounds and classes used to create the dataset. This process works for both validation and training sets, but of course, training sets are written into the .list files differently.

3.4. CAST-11: Only Challenging Pairs

With the help of our toolkit, we create a benchmark called CAST-11 (C11). The name includes 'CAST' because it is used for subsampling and the 11 is for the 11 sub-benchmarks for different categorical attributes. The first motivation for C11 is to offer more fine-grained evaluation with sub-benchmarks. Specifically, the sub-benchmarks are: Black, Caucasian, East Asian, Latinx, Middle Eastern, Young, Female, Male, Glasses & Facial Hair, Low-paq2piq, and random.

The second motivation for C11 is that most pairs in Face Recognition benchmarks are far from the decision threshold and have near-zero risk of being misclassified by a modern network. On the other extreme, there are some pairs that are mislabeled that they are impossible for a model to correctly classify. In Figure 3, we show that the genuine/impostor distributions for three previous benchmarks. It can be seen that most pairs in these datasets are trivial for a deep learning model to classify. We contend that rather than spending time and cycles validating or testing on trivial or impossible pairs, it is more efficient and useful to test on challenging pairs that are near the decision threshold.

When building the C11 benchmark we explicitly reject pairs that are classified to be far from the decision boundary (either too easy or too hard). To create each sub-benchmark, we first use CAST to obtain an available pool of images conditioned on the attribute that the category is named. As pairs are sampled from the pool of available images, they are passed through a ResNet100 to calculate the cosine similarity of the features. If the cosine similarity falls outside of the 'hard-pair threshold' the pair is rejected. In practice, we use 0.15-0.30 for the hard pair range. The range is tuned to 1) encompass the test-time threshold and 2) be wide enough such that enough pairs exist to create a benchmark. In Figure 3, it can be seen that *all* pairs are strictly within our hard pair threshold. This selection range is based on ResNet100 features, so the functional range may be different for other networks. However, in our experiments (Section 5.5) we find that the test sets created with this procedure are challenging for all models. In Section 5.5, results are presented on each sub-benchmark along with overall scores. Additionally, pseudo-code for creating the C11 benchmark is provided in the supplementary.

4. Probabilistic Inter-Category Evaluation

In this section, we introduce probabilistic inter-category evaluation as a supplemental method for comparing performance differences across categorical groups. Our motivation for this evaluation method it that attributes are often labeled such that each sample each belongs to one and only one class. In some cases, one-sample-per-category is a poor ontology. Consider the case of classifying an individual's race. While race is often viewed as a categorical attribute (e.g., Asian, Black, Middle Eastern, White, etc.), many people belong to multiple categories to varying degrees. For this reason, it can be worthwhile to account for samples that belong to multiple categories when validating performance between attributes. Furthermore, when subsampling data, the pool of available samples exponentially decreases as conditions are added. However, if hard filtering is not required, the pool of available samples does not decrease.

Probabilistic inter-category is founded on using softmax as probabilities [12, 20], which we obtain through attribute classifiers. The implementation is as follows. Let $\mathcal{S}$ be a set of test samples. Consider sample $x \in \mathcal{S}$ which belongs to attribute class $l \in L$ with probability $\mathcal{P}(x = l)$. First, let's consider the standard method for comparing performance on subsets. To test performance on each attribute class, $||L||$ disjoint sets are created,

$$S_l = \{x \in N | l = \underset{l' \in L}{\operatorname{argmax}} \mathcal{P}(x = l')\}.$$

For each attribute class, test accuracy A_l is calculated as

$$A_l = \frac{\sum\limits_{x \in S_l} \begin{cases} 1, & \text{if } y_x = \mathcal{F}(x) \\ 0, & \text{otherwise} \end{cases}}{||S_l||} \tag{1}$$

where $\mathcal{F}$ is a function and y_x is a ground truth label.

As an additional method for comparing performance between categories, we propose removing the disjoint subsets. Instead, test accuracy A_l is calculated such that each sample $x \in S$ contributes to A_l proportional to probability $\mathcal{P}(x = l)$,

$$A_l = \frac{\sum\limits_{x \in S} \begin{cases} \mathcal{P}(x = l), & \text{if } y_x = \mathcal{F}(x) \\ 0, & \text{otherwise} \end{cases}}{\sum\limits_{x \in S} \mathcal{P}(x = l)}. \tag{2}$$

In this work, we refer to Equation 1 as disjoint set evaluation and Equation 2 as probabilistic inter-category evaluation. A comparison of results with each method on demographic categories can be found in Section 5.4.

5. Experiments

For our experiments, we use CAST to generate test sets from the WebFace42M dataset. Procedures for generating

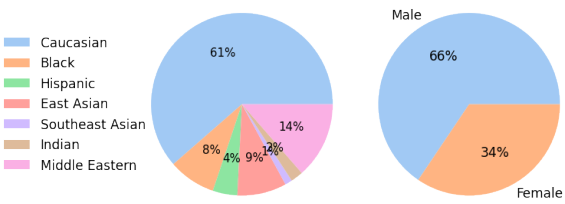

Figure 5. Demographic information from the WebFace42M dataset, computed with FairFace [28].

test sets are discussed in Section 3. Our experiments are presented in several following subsections. In Section 5.1, we discuss training details. In Section 5.2, we present results on demographic test splits. Then, in Section 5.3 we show results on image-quality test splits. We compare results between probabilistic inter-category evaluation and standard disjoint sets in Section 5.4. Last, we provide results from twelve models on our CAST-11 (C11) benchmark in Section 3.4 results.

5.1. Experimental Settings

We use a total of 12 models for our experiments. We develop 6 models trained on WebFace4M using combinations between three loss functions and two backbone architectures. For losses we use margin-based softmax losses:ArcFace [19], CosFace [60], and MagFace [37]. For backbone architectures we use ResNet50, and ResNet100 [26]. For fair comparison, each loss function and backbone architecture is implemented into a single repository with uniform settings (described below) and trained on WebFace4M. Additionally, we use three pre-trained models on Glint360k and three pre-trained models on MS1Mv3 from the InsightFace repository [19] each with the two backbones above and ResNet34 [26].

We use a batch size of 256 per GPU on each of 3xRTX3090,1xA6000 GPUs. For more efficient training we use mixed-precision floating point [39]. We follow prior work for setting hyperparameters. Training is completed over 20 epochs with Stochastic Gradient Descent (SGD) optimizer and polynomial weight decay. A base learning rate 0.1 is used. Horizontal-flip is adopted as augmentation. Our primary experiments are on validation sets created using CAST, however, we provide performance on common benchmarks LFW [27], AgeDB-30 [41], CFP-FP [50], and IJB-C [36] as a reference point for the difficulty of our proposed subsets. As a reference point, results on previous benchmarks for each of the twelve models used in our experiments can be found in Table 1.

5.2. Demographic Subset Results

Table 2 shows results on 14 different subsets on race and gender. Of the 14 subsets, 'Southeast Asian Female' is has the lowest score with 98.20 and 'Caucasian Male' has the highest score with 99.90. Additionally, Table 2, the average for each race and each gender. 'Middle Eastern' and 'Southeast Asian' have the lowest scores which are almost 1% lower than 'Caucasian'. For gender, Male has 0.74% higher scores. The average for all sets in this experiment is 99.18%.

Data	Model	Loss	LFW	CFP	AgeDB	IJB-C
WF4	R50	CosFace	99.82	99.11	97.92	96.89
WF4	R50	ArcFace	99.78	99.11	97.92	96.78
WF4	R50	MagFace	99.78	98.89	97.78	96.73
WF4	R100	CosFace	99.82	99.14	98.15	97.26
WF4	R100	ArcFace	99.83	99.23	98.07	97.10
WF4	R100	MagFace	99.82	99.03	98.07	97.01
G-Pre	R34	CosFace	99.80	98.76	98.32	96.56
G-Pre	R50	CosFace	99.80	99.14	98.20	96.97
G-Pre	R100	CosFace	99.80	99.24	98.28	97.32
M-Pre	R34	ArcFace	99.77	98.19	97.87	95.91
M-Pre	R50	ArcFace	99.80	98.40	98.20	96.46
M-Pre	R100	ArcFace	99.82	98.93	98.47	96.81

WF4=WebFace4M;MS-pre=MS1Mv3[19]; G-Pre=Glint360k[19]

Table 1. For reference, results on previous benchmarks with each of the 12 models used in our experiments.

Race	Female	Male	Average
Black	99.01±0.17	99.34±0.07	99.18
Caucasian	99.69±0.03	99.90±0.02	99.80
East Asian	98.41±0.10	99.71±0.05	99.06
Hispanic Latino	99.04±0.11	99.69±0.05	99.37
Indian	98.93±0.10	99.42±0.08	99.18
Middle Eastern	98.35±0.07	99.32±0.06	98.87
Southeast Asian	98.20±0.12	99.44±0.07	98.82
Average	98.80±0.48	99.54±0.20	99.18

Table 2. Results on demographic evaluation sets. 10 test sets of 10,000 pairs are used for each race and gender combination. It can be seen performance is highest on Caucasian's and Males. ArcFace ResNet100 trained of WebFace4M is used here.

5.3. Image-Quality Subset Results

Here, results are presented for subsets sampled according to image quality. Three blind image-quality metrics: NIMA, BIRSQUE, and paq2piq, and two face-specific image-quality metrics: SDD-FIQA and MagFace. Using CAST, we subsample 10 datasets—each with 5,000 genuine and 5,000 imposter pairs—for each of 10 quartile ranges. Subsets are created for each quality metric for a total of 500 datasets. In Figure 6, results are plotted for increasing quartile ranges. In Figure 6, it can be seen that the Nima blind image-quality measure corresponds poorly with performance. However, paq2piq and BRISQUE are found to be useful. Upon manual inspection of the images, we find low paq2piq scores are often blurry images. Unsurprisingly, we find the face-specific image-quality measures to be more representative of face recognition performance. A final observation from Figure 6,

is that 1:1 verification face recognition performance is high even on lower quality images.

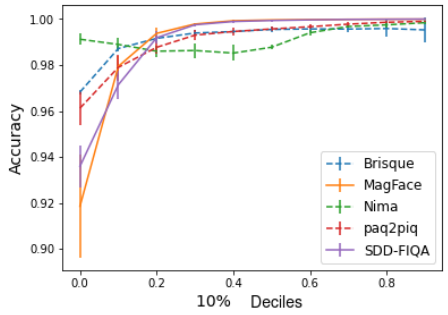

Figure 6. Performance on test sets sampled from 10 quartile ranges for each of 5 quality assessment measures. Face-specific quality assessments are shown with solid lines, and blind image quality assessments are show with dashed lines. The face-specific image quality measures show the best ResNet100 trained on WebFace4M with ArcFace loss is used here.

In Table 3, we show results with each of the twelve models on datasets sampled from low (0-10%), medium (45-55%), and high (90-100%) quartile ranges of the SDD-FIQA metric. We use SDD-FIQA because we find it to be correlated with face recognition performance. We do not use MagFace since it is also one of the training methods. Overall, it can be seen that performance is much lower on the low range than the medium range. However, performance is nearly identical on the medium quartile range and the high quartile range. This indicates that face recognition performance can become saturated even for medium-quality images. Within the WebFace4M models, it can be seen that the CosFace loss incurs the least drop in performance from the medium to the low range. It can also be seen that the ResNet50 models have nearly the same performance as the ResNet100 model on the medium and high quartile ranges, but the ResNet100 models perform significantly better on the low quartile range. This suggests that bigger models may only make improvements on challenging samples. When comparing between training sets, Table 3 shows that WebFace4M models drop the least in performance between medium to low, while models trained on MS1Mv3 have the greatest drop in performance.

5.4. Probabilistic Inter-Category Evaluation Results

Figure 7 shows results on demographic splits for both standard disjoint sets and for probabilistic inter-category evaluation (introduced in Section 4). For this experiment, the datasets from Table 2 were combined for male and female. For probabilistic inter-category evaluation, the same evaluation sets from Table 2 are made it to one large 1,400,000 pair evaluation set. From Figure 7, it can be seen that the evaluation methods produce similar overall results. While it can not be ascertained from Figure 7 which is a more accurate

Model			SDD-FIQA Quartile Ranges		
Data	Model	Loss	**0-10%**	**45-55%**	**90-100%**
WF4	R50	CosFace	89.81±0.88	99.59±0.15	99.59±0.16
WF4	R50	ArcFace	88.77±0.50	99.61±0.18	99.64±0.10
WF4	R50	MagFace	88.58±0.64	99.62±0.17	99.64±0.13
WF4	R100	CosFace	91.50±0.67	99.75±0.11	99.71±0.07
WF4	R100	ArcFace	90.01±0.93	99.71±0.15	99.73±0.08
WF4	R100	MagFace	90.91±0.66	99.72±0.17	99.70±0.11
G-Pre	R34	CosFace	84.81±0.70	99.68±0.14	99.61±0.12
G-Pre	R50	CosFace	86.54±0.81	99.77±0.12	99.73±0.11
G-Pre	R100	CosFace	88.84±0.54	99.82±0.11	99.81±0.10
M-Pre	R34	ArcFace	80.71±0.85	99.50±0.18	99.49±0.20
M-Pre	R50	ArcFace	82.79±0.73	99.72±0.16	99.63±0.13
M-Pre	R100	ArcFace	85.15±0.92	99.81±0.09	99.76±0.11

WF4=WebFace4M;MS-Pre=MS1Mv3[19]; G-Pre=Glint360k[19]

Table 3. Results on 10 datasets sampled from each of three quartile ranges of the SDD-FIQA attribute. It can be seen that the models pretrained on MS1Mv3 perform worse on the lower quality images and models trained on WebFace4M perform better on the lower quality sets..

evaluation method, the similarity between the two indicates that either may be suitable.

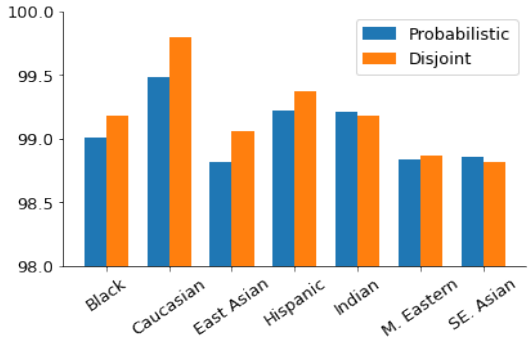

Figure 7. Results on demographic splits using probabilistic inter-category evaluation and standard disjoint sets. The test sets from Table 2 are combined for male and female for disjoint sets in this experiment. For probabilistic evaluation, all sets are combined into one.

5.5. C11 Results

We test each of our twelve evaluation models discussed in Section 5.1 on the C11 benchmark (introduced in Section 3.4). The results for each model on each of the 11 sub-benchmarks can be found in Table 4. From the bottom right entry in the table it can be seen that the average score between all models is 82.07, or 17.93 percent error. This is significantly different than the error of previous benchmarks such as those shown in Table 2, which range from 0.17%-4.09% error. The increased difficulty on C11 is due to the procedure of excluding easy pairs. For the sub-benchmarks, Low paq2piq has the lowest average score of 80.61 and Caucasian has the highest average score of 85.24. Of the three loss functions, CosFace outperforms the others with all other settings held constant on C11. Out of the three

CAST-11 (C11) Benchmark

Model			Test Sets					
Data	Backbone	Loss	Black	Caucasian	E. Asian	Latinx	M. E.	Young
WebFace4M	R50	CosFace	81.72±1.22	83.70±0.91	81.83±1.09	81.76±0.98	82.06±1.47	82.45±0.77
WebFace4M	R50	ArcFace	81.00±1.37	83.47±0.91	82.11±1.08	81.14±1.22	81.49±1.29	81.88±0.74
WebFace4M	R50	MagFace	80.89±1.17	83.42±0.96	81.27±0.88	80.58±1.35	80.78±1.38	82.23±1.02
WebFace4M	R100	CosFace	85.83±1.01	87.16±1.23	84.96±0.97	86.11±0.83	85.93±0.90	86.32±1.56
WebFace4M	R100	ArcFace	82.87±1.19	85.75±1.33	82.58±1.16	83.37±1.55	83.74±1.20	83.67±0.47
WebFace4M	R100	MagFace	84.49±0.85	87.10±0.92	84.45±1.33	85.28±0.97	85.25±1.28	85.40±1.59
G-Pre[19]	R34	CosFace	77.51±1.40	84.15±1.19	79.68±1.21	79.23±1.05	79.11±1.04	80.36±1.31
G-Pre[19]	R50	CosFace	83.07±1.19	87.48±0.98	83.38±1.24	83.74±0.97	83.39±1.20	84.33±1.09
G-Pre[19]	R100	CosFace	86.51±1.17	91.16±0.81	87.09±1.23	87.79±0.95	86.85±0.94	88.01±0.93
M-Pre[19]	R34	ArcFace	70.47±0.94	80.07±1.13	72.74±1.13	71.21±1.34	71.55±1.27	73.47±1.74
M-Pre[19]	R50	ArcFace	74.81±0.66	82.71±1.20	76.39±0.87	75.14±1.27	75.96±1.06	76.78±1.37
M-Pre[19]	R100	ArcFace	78.92±1.35	86.80±1.12	80.23±1.57	79.80±1.38	80.52±0.86	81.03±1.00
Average			80.67	85.24	81.39	81.23	82.16	82.63
Data	Backbone	Loss	Female	Male	G&FH	L-p2p	Random	**Overall**
WebFace4M	R50	CosFace	83.37±1.31	83.62±1.33	81.09±1.16	81.21±1.02	83.69±1.14	82.41±1.50
WebFace4M	R50	ArcFace	83.09±1.28	83.04±1.02	80.11±1.21	80.86±1.01	82.95±0.84	81.92±1.52
WebFace4M	R50	MagFace	82.59±0.92	82.64±0.88	80.29±0.74	80.14±0.68	83.47±0.86	81.66±1.56
WebFace4M	R100	CosFace	87.50±0.97	87.54±1.04	85.42±1.22	85.44±1.45	87.54±0.92	86.34±1.44
WebFace4M	R100	ArcFace	85.44±0.98	86.06±0.91	83.16±0.88	82.33±1.16	85.84±0.73	84.07±1.73
WebFace4M	R100	MagFace	86.79±0.90	87.08±0.98	84.82±0.93	84.78±0.86	87.09±1.12	85.68±1.51
G-Pre[19]	R34	CosFace	79.70±1.15	80.65±1.28	78.60±1.62	78.24±1.02	80.41±1.51	79.79±2.09
G-Pre[19]	R50	CosFace	84.69±1.21	85.19±0.89	82.84±0.89	82.26±1.10	85.35±1.25	84.16±1.78
G-Pre[19]	R100	CosFace	88.44±0.73	89.10±0.90	86.66±0.99	85.96±1.12	88.71±1.14	87.84±1.73
M-Pre[19]	R34	ArcFace	73.12±1.03	72.81±1.07	71.76±1.04	71.29±0.87	72.50±1.03	72.82±2.72
M-Pre[19]	R50	ArcFace	77.67±0.87	77.59±1.51	76.12±0.89	74.95±0.67	77.30±1.14	76.86±2.35
M-Pre[19]	R100	ArcFace	82.05±0.74	82.19±0.61	81.15±0.79	79.83±1.17	82.10±1.20	81.33±2.29
Average			82.87	83.13	81.00	80.61	83.08	**82.07**

R=ResNet; M.E.=Middle Eastern; G&FH=Glasses & Facial Hair; L-p2p=Low paq2piq; G-Pre: Glint360k; M-Pre: MS1M.

Table 4. Results on the CAST-11 (C11) benchmark. C11 includes 11 sub-benchmarks on demographics, face attributes (e.g., beard and glasses), and image characteristics. Each sub-benchmark has 10 1,000 pair folds, for a total 110,000 pairs in the benchmark. All genuine and imposter pairs are chosen close to the cosine similarity decision threshold, and thus the benchmark emphasizes performance on challenging (but not impossible) samples. Of the 12 models used in our experiments, the Resnet100-CosFace model pretrained on Glint360k scored the highest with an average score of 87.84 across the sub-benchmarks.

training sets, MS1Mv3 models perform the worst on C11, and Glint360k and WebFace4M perform similarly. The models trained on the WebFace4M have more fair performance across demographic groups (e.g., Male vs. Female). There is a surprisingly large difference between performance on Caucasian for Glint-R100-CosFace (91.16) and WebFace4M-R100-CosFace (87.16) of 4.0%.

6. Conclusion

This work examines fine-grained testing of deep learning models for face recognition. We build the Conditional Attribute Subsampling Toolkit for easy creation and evaluation of data subsets based on indexed metadata. Using CAST, face recognition models are evaluated on several attributes such as demographics and image-quality. We find that there are statistically significant differences in performances between demographic groups, which, like prior works, suggest further work is needed to develop less biased models. Additionally, based on observations that most sampled pairs are easily classified by deep learning models, we create a new benchmark (C11), which is designed to only contain challenging pairs. Using CAST, the C11 benchmark can be easily evaluated on new models. Paths of future work include using sampled training sets for creating fewer performance discrepancies across demographics, and using CAST attributes as a method for managing training distributions.

References

[1] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34(11):1617–1630, 2018.

[2] Alejandro Acien, Aythami Morales, Ruben Vera-Rodriguez, Ivan Bartolome, and Julian Fierrez. Measuring the gender and ethnicity bias in deep models for face recognition. In *Iberoamerican Congress on Pattern Recognition*, pages 584–593. Springer, 2018.

[3] Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. *arXiv preprint arXiv:2008.06989*, 2020.

[4] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 7617–7627, June 2021.

[5] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Proceedings of the ieee/cvf winter conference on applications of computer vision workshops*, pages 81–89, 2020.

[6] Vítor Albiero, Kai Zhang, and Kevin W. Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2020.

[7] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021.

[8] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4042–4051, June 2022.

[9] J Ross Beveridge, Geof H Givens, P Jonathon Phillips, and Bruce A Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762, 2009.

[10] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. *arXiv preprint arXiv:2206.04867*, 2022.

[11] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. *arXiv preprint arXiv:2112.06592*, 2021.

[12] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[13] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European conference on computer vision*, pages 768–783. Springer, 2014.

[14] Jingchun Cheng, Yali Li, Jilong Wang, Le Yu, and Shengjin Wang. Exploiting effective facial patches for robust gender recognition. *Tsinghua Science and Technology*, 24(3):333–345, 2019.

[15] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, and Jerry L Tipton. Fixed and varying effects of demographic factors on the performance of eleven commercial facial recognition systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1):2, 2019.

[16] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the european conference on computer vision (eccv) workshops*, pages 0–0, 2018.

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

[18] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*, pages 741–757. Springer, 2020.

[19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] John Denker and Yann LeCun. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3, 1990.

[21] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.

[22] Geof H Givens, J Ross Beveridge, P Jonathon Phillips, Bruce Draper, Yui Man Lui, and David Bolme. Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics & Data Analysis*, 67:236–247, 2013.

[23] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (frvt) part 3: Demographic effects. *Nat. Inst. Stand. Technol., Gaithersburg, MA, USA, Rep. NISTIR*, 8280, 2019.

[24] Patrick J Grother, Patrick J Grother, P Jonathon Phillips, and George W Quinn. *Report on the evaluation of 2D still-image face recognition algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2011.

[25] Manuel Günther, Andras Rozsa, and Terranee E. Boult. Affact: Alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99, 2017.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[27] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

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

[29] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

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

[31] KS Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C King, and Kevin W Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[32] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[34] Boyu Lu, Jun-Cheng Chen, Carlos D Castillo, and Rama Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):42–55, 2019.

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

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

[37] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, June 2021.

[38] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[39] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations*, 2018.

[40] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[41] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[42] Avanika Narayan, Piero Molino, Karan Goel, Willie Neiswanger, and Christopher Ré. Personalized benchmarking with the ludwig benchmarking toolkit. *arXiv preprint arXiv:2111.04260*, 2021.

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

[44] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, June 2021.

[45] Karl Ricanek, Shivani Bhardwaj, and Michael Sodomsky. A review of face recognition against longitudinal child faces. *BIOSIG 2015*, 2015.

[46] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 341–345. IEEE, 2006.

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

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

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

[50] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

[51] Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, JC Karure, Raja Raju, B Rajan, et al. Indian movie face database: a benchmark for face recognition under wide variations. In *2013 fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG)*, pages 1–5. IEEE, 2013.

[52] Philip Smith and Karl Ricanek. Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 90–97, 2020.

[53] Nisha Srinivas, Matthew Hivner, Kevin Gay, Harleen Atwal, Michael King, and Karl Ricanek. Exploring automatic face recognition on match performance and gender bias for children. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 107–115. IEEE, 2019.

[54] Nisha Srinivas, Karl Ricanek, Dana Michalski, David S Bolme, and Michael King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[55] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[56] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score

normalization. *Pattern Recognition Letters*, 140:332–338, 2020.

[57] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th international workshop on biometrics and forensics (iwbf)*, pages 1–6. IEEE, 2020.

[58] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[59] Ruben Vera-Rodriguez, Marta Blazquez, Aythami Morales, Ester Gonzalez-Sosa, Joao C Neves, and Hugo Proença. Facegenderid: Exploiting gender information in dcnns face recognition systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[60] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[61] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3779–3782, 2021.

[62] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019.

[63] Qingzhong Wang, Pengfei Zhang, Haoyi Xiong, and Jian Zhao. Face. evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*, 2021.

[64] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. *arXiv preprint arXiv:2206.01881*, 2022.

[65] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[66] Yunxuan Zhang, Li Liu, Cheng Li, et al. Quantifying facial age by posterior of age comparisons. *arXiv preprint arXiv:1708.09687*, 2017.

[67] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[68] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018.

[69] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[70] Manli Zhu and Aleix M Martínez. Optimal subclass discovery for discriminant analysis. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 97–97. IEEE, 2004.

[71] Manli Zhu and Aleix M Martinez. Subclass discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1274–1286, 2006.

[72] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502, June 2021.