This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Rapid Test for Accuracy and Bias of Face Recognition Technology

Manuel Knott[1*] Ignacio Serna[1,2*] Ethan Mann[1*] Pietro Perona[1]

[1]California Institute of Technology
[2]Center for Humans and Machines, Max Planck Institute for Human Development
[*]Equal contribution

Abstract

Measuring the accuracy of face recognition (FR) systems is essential for improving performance and ensuring responsible use. Accuracy is typically estimated using large annotated datasets, which are costly and difficult to obtain. We propose a novel method for 1:1 face verification that benchmarks FR systems quickly and without manual annotation, starting from approximate labels (e.g., from web search results). Unlike previous methods for training set label cleaning, ours leverages the embedding representation of the models being evaluated, achieving high accuracy in smaller-sized test datasets. Our approach reliably estimates FR accuracy and ranking, significantly reducing the time and cost of manual labeling. We also introduce the first public benchmark of five FR cloud services, revealing demographic biases, particularly lower accuracy for Asian women. Our rapid test method can democratize FR testing, promoting scrutiny and responsible use of the technology. Our method is provided as a publicly accessible tool at https://github.com/caltechvisionlab/frt-rapid-test.

1 Introduction

Face recognition technology (FRT) is a convenient, no-contact, fast, accurate, and inexpensive way to interface securely people and machines. From logging into our smart devices to boarding a plane, crossing a border, and finding missing children, FRT can make our lives more convenient and safer. Conversely, FRT misuse is possible and may lead to loss of privacy and violation of civil rights [12, 46]. As applications increase, it is crucial to understand FRT's potential and implications. In particular, characterizing FRT systems' accuracy and bias is fundamental to informing developers, users, the public, and regulators about the merits and downsides of the technology and to improve it if necessary [12, 14, 21, 24, 26]. On the positive side: Accuracy in FRT systems has improved dramatically in the past five years. Today, systems achieve super-human accuracy [16, 32, 41, 50] and outperform even expert face analysts [38], with the potential to make our life more convenient and help reduce the harm that is currently caused by human error [9, 45]. Furthermore, measuring and mitigating bias in algorithms is both feasible and effective, while measuring and correcting biases in human operators is notoriously difficult and can take a long time [13, 35]. Additionally, applications of FRT to policing could work alongside DNA testing to help solve crimes quickly [20], reduce bias in the justice system [34], and reduce the rate of wrongful identification [9, 10] and imprisonment [45]. Thus, AI and FRT may become powerful agents of progress towards more fair, accountable, and transparent institutions [27, 35]. Amongst the potential downsides: willful or inadvertent misuse, inaccuracy, and bias in face recognition technology could inflict harm on individuals and lead to social inequities [12, 26, 46]. Responsible practice in developing and deploying the technology starts with measuring algorithmic accuracy and bias.

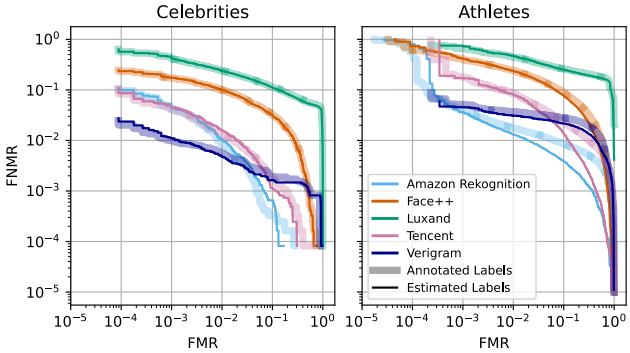

Figure 1. **Unsupervised accuracy estimates on five FR cloud services match supervised estimates.** The plots show the False Non-Match Rate (FNMR) (equivalently, the False Reject Rate) vs. the False Match Rate (FMR) (equivalently, the False Accept Rate) of five commercial cloud services on two collections of face images: Celebrities and Athletes. The thin dark lines indicate accuracy as estimated by our automated method. The thick pale lines indicate the ground-truth estimates through human labeling, which took more than a month of human labor to produce. See also Sec. 9.

Unfortunately, testing FRT is expensive and laborious, and thus, it is not within the reach of most organizations.

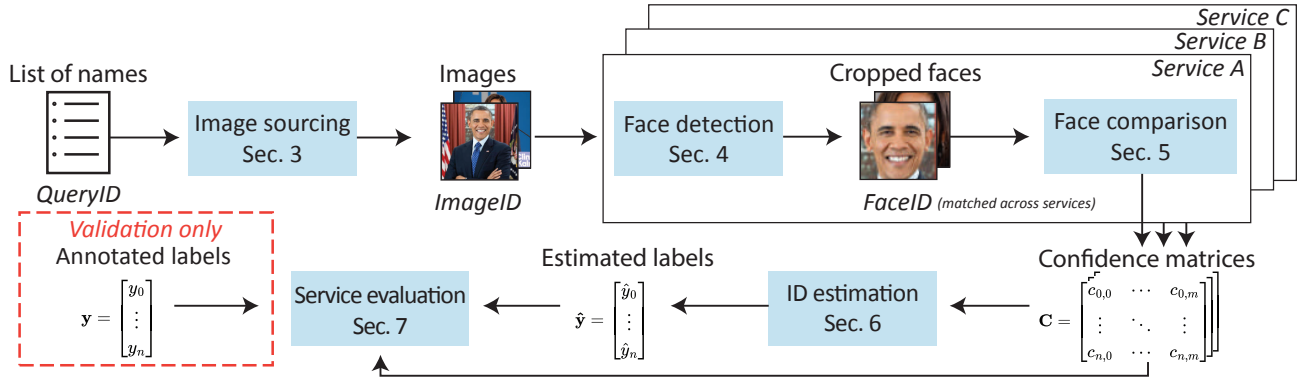

Figure 2. **Overview of our method.** An operator provides a list of people's names that are used as queries for image URL sourcing from the internet. The images are accessed through their URL and are not stored. Several face recognition services are evaluated simultaneously (five in this study). Each service detects faces in the selected images and assigns a same-identity (or "match") confidence value to pairs of faces. From this information, an estimate of which faces belong to which identity is computed. Using this estimate FNMR-vs-FMR curves and bias estimates may be produced (Figs. 1 and 5) to estimate the accuracy of each service. Our method does not require hand-annotation and estimates identity labels for each face image from the data. In this study hand-annotated labels were collected purely to validate our method and were not available to our method.

The best data on FRT accuracy is published by the U.S. National Institute of Standards and Technology (NIST). While the NIST team is reputable, experienced, and uses first-rate test sets (Sec. 2), the situation is not ideal. First, not all FRT vendors submit their software to NIST for testing. Second, NIST is not testing cloud services. Third, to prevent vendors from overfitting, the test sets are kept secret and cannot be checked by independent experts for correlations and confounds. As a result, acceptance of NIST's figures rests on the institution's reputation rather than peer review. A better state of affairs would be for any interested party to carry out those tests that it deems important independently. Today, this is not possible: the use of public datasets often leads to evaluation data leaking into the training process; new test sets collected by independent and academic teams are subject to privacy and copyright issues and are either too small or the ground truth is noisy at best, making testing results unreliable.

We propose a method to address these shortcomings and democratize the testing of FRT systems. It is based on two main ideas. First, use public data which is open to verification. To minimize the risk that the test data was used in model training, our method will only make use of images that have been made public recently. To minimize use risks images are analyzed on the fly and not stored. Second, to make the testing practical and affordable and to avoid human labeling errors, our method does not rely on hand-labeled images but rather infers the ground truth labels of face identity from the algorithms' confidence values. This is a delicate algorithmic step that sits at our method's core. Fig. 2 shows an overview of our proposed method. Our main contributions are: (a) An inexpensive and practical observational method for accurately benchmarking face recognition algorithms, (b) a thorough experimental validation of our method, and (c) the first public benchmark of accuracy and bias for face recognition services in the cloud, with a side-by-side comparison of five popular cloud services.

2 Previous work

Estimating FRT accuracy and bias requires large, accurately labeled datasets to ensure tight confidence intervals. Furthermore, one needs diverse attributes representative of the general population to explore effects on all demographics. A team with the U.S. National Institute of Standards and Technology (NIST) [18, 19] has, over the past 20 years, developed state-of-the-art test datasets and testing practices. They test algorithms on six large (∼10M images) datasets collected from visas, visa applications, border crossings, arrest mugshots, kiosk images, and images collected in the wild. Accurate identity annotations are achieved by combining trained government officials and identity documents. NIST publishes updated reports every few months on NIST's "Face Recognition Vendor Test" web page [36]. A number of academic teams are also engaged in testing FRT [6–8, 28]. They use public datasets that may have been included in the training sets of FRT vendors and whose identity labels are often not accurate. Thus, while valuable for science, academic tests may not be suitable for probing the accuracy of commercial systems.

Only governments and large tech companies have access to large, accurately labeled datasets. To democratize the testing of FRT algorithms, we need to reduce the cost of ground-truth identity annotation dramatically. This has long been considered unlikely since accurate face identification using human annotators is very difficult [38], and benchmarking without an independently annotated ground truth might seem impossible. Semi-automated clean-up meth-

ods have been proposed to reduce the cost of improving label quality in training sets. The state-of-the-art, WebFace260M [51], iteratively employs a face recognition (FR) model to clean the images and then re-trains that model with the new clean dataset to obtain an improved model. This dynamic has three shortcomings in our application: (i) Constructing a high-quality embedding requires millions of images, which is the case for training sets. We focus on test sets, which are a couple of orders of magnitudes smaller and yet require higher accuracy. (ii) The stable point of the method may suffer from partial mode collapse due to errors in the original data leaking into the model. This has been documented on small datasets [44] and is difficult to verify on large datasets. (iii) Legislation in many countries prohibits the storage of biometric data that enables the identification of individuals. To address the first two problems, our method takes advantage of the ensemble of the embeddings of the systems being tested, which is high-quality and stable. To address the latter, our method does not store any identity labels or images, relying solely on the scores of FR systems.

Recent studies propose semi- and unsupervised methods for benchmarking algorithm accuracy [15, 23, 47], which can estimate algorithms' accuracy even without access to an externally provided ground truth. These methods take advantage of statistical regularities of the confidence values of classification algorithms to estimate the underlying error statistics. We take inspiration from this work. We note that face recognition algorithms are highly accurate and thus will produce strongly bimodal distributions of confidence values when confronted with an (unlabeled) mix of same-ID and different-ID pairs of face images. We exploit this fact to estimate the ground truth image identity labels.

Our method addresses the practical case of image collections obtained through web image searches, where face identities are typically correct in the 20-80% range. Thus, our method is supervised in that identity (ID) labels are provided. However, crucially, it is designed to tolerate highly noisy face identity labels and does not require human supervision, such as hand-labeling, to correct such errors. In other words, we address the situation that lies in between the availability of exact ground truth (conventional benchmarks based on carefully annotated test sets) and the (quasi) complete absence of identity labels (previous unsupervised and semi-supervised methods [15, 47]).

Like other benchmarks, our method uses *observational* test sets with demographic annotations (typically age, gender, ethnicity) [18, 25, 33, 40]. Recent literature indicates that observational methods are susceptible to bias from unmodeled confound variables and propose *experimental* approaches based on synthetic images [11, 31]. We agree that there is merit in this concern. For the time being, we believe that both observational and experimental methods need to be used to assess the accuracy of FRT.

3 Image sourcing

The procedure we recommend is designed to involve the minimum amount of human curation during image sourcing and does not create a static test image dataset. Instead, it selects images on the web, feeds pairs of them to the FRT cloud services being benchmarked, and retains only the resulting confidence values for analysis.

The process starts with a human-generated list of names that serve as queries in an image retrieval system. For bias analysis, one can additionally provide demographic attributes (e.g. race, gender, age) for each name. For the evaluation part of this research, we generated two test sets with different image statistics. The first, *Celebrities*, starts from a list of names of famous people compiled by one of us. It contains 10 names in each of the eight demographic categories. The second, *Athletes*, is a subset of Wikipedia's list of 2020 Tokyo Olympic athletes. Our list comprises 2755 names and aims to be balanced across six demographics. Detailed statistics for both datasets can be found in Sec. A.

Only images published shortly before testing are considered to reduce the chance that test data was used to train the models being tested. In our experiments, URLs of the images were obtained from the Google Images API and the Google News API. Our script obtained a total of 5k images (an average of 67 per ID) for the Celebrities dataset and 223k images (an average of 81 per ID) for the Athletes dataset, leaving 2.2k and 58.6k, respectively, after face detection (see Sec. 4). The number of images found per identity varies significantly (see Fig. S.1). All images obtained using a given name string were assigned the same *QueryID* (abbreviated as q in the following). At this point, some, but not all, of the faces in the pictures obtained belong to the person whose identity corresponds with the search query. E.g., we expect that a search for "Barack Obama" will yield images of Barack Obama, as well as Michelle Obama, Joe Biden, and other world leaders. The algorithm described in Sec. 6 is designed to clean up these noisy labels.

To validate the results of this study, we manually added ground truth identity labels to each face image: one of the authors assigned label $y = 1$ if the identity matched the query name, $y = 0$ if not, and $y = -1$ for rare cases where the identity could not be confirmed even after meta information was consulted. The manual annotation process took a total of 200 hours, about 12 seconds per image on average. Additional details about the image sourcing and annotation process can be found in Sec. A.

4 Face detection

Face detection, i.e. computing bounding boxes around each visible face, was carried out in every image using each one of the cloud services we tested. Images were sent to each cloud service's face detection API in our benchmark. Since

bounding box sizes vary by service, we retained service-specific crops, assuming matching models were trained on these.

To establish a unified *FaceID* across providers, we grouped bounding boxes of detected faces from different services. For this, we computed the IoU (Intersection over Union) metric for all pairwise combinations of detected faces in an image, excluding pairs of faces detected by the same provider. Pairs of detected faces with sufficiently high IoU measures (we used IoU $>$ 0.2) were grouped using Kruskal's algorithm [29] to compute the minimum spanning tree. Thresholding the minimum spanning produced face groups across services. Finally, each group was assigned a *FaceID*.

We use the additional constraint only to include images where each provider found exactly one face. This allows us to include services that do not offer a face detection API and, therefore, can only include single-face images (Verigram in our case, see Sec. B for details on how to interrogate services without prior face detection). A positive side effect of using single-face images only is that lower-resolution background faces are less likely to occur, leading to more identifiable face crop images.

5 Face matching confidence scores

Face recognition cloud services assign a *confidence score* $C_{i,j}$ to each pair of faces (i, j). A high confidence score indicates that the pair of faces are likely to belong to the same person, while a low confidence score indicates they are likely to belong to different people. Estimating a service's accuracy requires computing the false non-match rate (FNMR, a.k.a. false reject rate) vs the false match rate (FMR, a.k.a. false accept rate) as a function of a minimum confidence threshold. Thus, for each cloud service provider and each face pair in the test set, we obtain pairwise confidence matches and evaluate the quality of such matches (Sec. 8) vis-a-vis the estimated labels (Sec. 6).

FRT providers we tested are Amazon Rekognition [1], Face++ [2], Luxand [3], Tencent [4], and Verigram [5]. We used paid services through regular subscriptions, except for Verigram, for which we received complimentary access for research purposes. Computing confidence scores for all pairs of faces in the dataset is very expensive and unnecessary. Our method requires same-query pairs to evaluate FNMR as well as a comparable number of cross-query pairs to evaluate FMR. The set of face pairs to be evaluated by cloud providers was selected as follows: 1. All pairs of faces with the same q were used both for face ID label estimation (Sec. 6) and model evaluation (Sec. 7). 2. A random sample of pairs of faces with different q and from the same demographic group was used for model evaluation (Sec. 7). We sampled as many different-query pairs as same-query pairs.

A bimodal distribution of confidence values is expected from the services, where one mode is associated with different-ID (or impostor) matches, and one mode is associated with same-ID (or genuine) matches. Since the distribution of the confidence values, and thus the modes of the distributions, are different for each service, for estimating the identity labels (Sec. 6), we normalized the confidence values to the $(0, 1)$ range by mapping the range of values linearly so that the two modes are mapped to 0.0 and 1.0, clipping all smaller and larger values, respectively. Modes can be specified manually or estimated by fitting a bimodal Gaussian Mixture model to the confidence value distribution. For service evaluation (Sec. 7), we report results using the services' original confidence values.

6 Identity label estimation

The next step is estimating the *identity label* $\hat{y}_i$ for each face image i. This requires two steps: deciding which identity (i.e., which physical person) corresponds to the query q and deciding whether image i corresponds to that person. Amongst the faces that were downloaded using the search string q, many will actually belong to different identities (Fig. S.6). Which person is the *correct identity* for a given name query q? Many people may be associated with the same name. How is this ambiguity resolved? For the hand-annotated labels, the "correct identity" is decided by the annotator. For our estimation method, the person/identity whose faces are prevalent in the set associated with q is defined as the *correct identity*. The two criteria coincide almost always.

We describe an algorithm that estimates which identity is prevalent, i.e., it decides which is *correct identity* and estimates the corresponding face images. The end result is an estimate of the identity label for every image in each name. Such identity labels will be used to estimate the error rates for each service (Sec. 7). We start with an intuitive description of the algorithm's steps, and we give a more formal description of the algorithm at the end of the section.

The intuition for our algorithm is simple: pairs of faces in query q corresponding to the same person will often, although not always, receive high pairwise confidence C^{qs}_{ij} from service s. If the confidence is low, chances are that a third image k of the same person will have high pairwise confidence with i and j. Thus, we may use C^{qs} as an *affinity* estimate to be used for grouping such faces using spectral factorization [42]. The largest group is most likely associated with the correct identity. Our algorithm is shown in Fig. 3. For the sake of simplicity, consider first the most common case: the collection of images associated with a name consists of images that belong to the *correct identity* and other images corresponding to a sprinkling of different identities (Fig. 3, first row). In this case, after reordering w.l.o.g. the image indices so that the correct identity is assigned contiguous indices, the confidence matrix C is block-

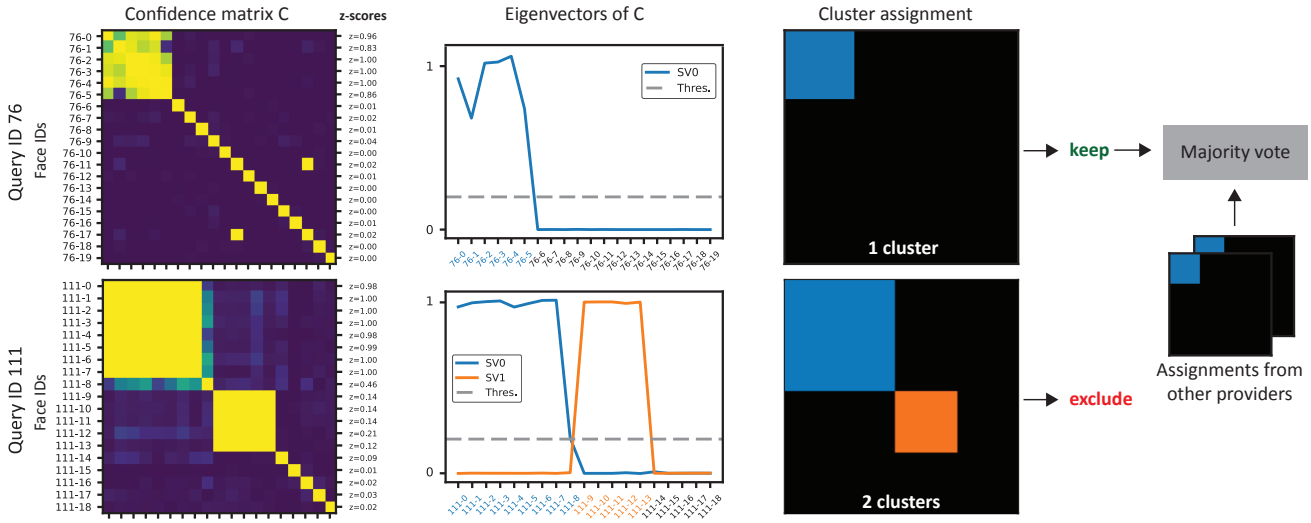

Figure 3. **ID Label Estimation Method (Sec. 6).** (Left column) Matrices showing the confidence values assigned by one of the FRT services to face pairs in queries $q = 76, 111$. Each row and each column corresponds to a face image, and each matrix entry indicates the service's confidence that the corresponding pair of face images belongs to the same person (the indices have been rearranged to make the block structure apparent). Yellow indicates high confidence, and blue indicates low confidence. The top matrix has a single block, while the bottom one has two blocks, suggesting that two different identities with a significant number of images are associated with the query. (Second column) The top eigenvectors (singular vectors whose singular value exceeds a threshold) of the matrices, where the x-axis indicates the image index, act as indicator functions of which images are associated with each identity. (Third column) By thresholding the eigenvectors, the algorithm discovers which images belong to which identity. (Right column) In the top row, information from the eigenvector is combined with corresponding eigenvectors from other services by majority vote. The bottom row does not meet the criteria for inclusion since it contains more than one identity and is discarded from further consideration.

diagonal, with one large block with $C_{ij} \simeq 1$ where both i and j correspond to the correct identity, and the rest of the entries are $C_{ij} \simeq 0$ (Fig. 3, left). Of course, the entries on the main diagonal are $C_{ii} = 1$ since they correspond to the confidence of an image matching itself.

The indices of the best estimate for the correct identity can thus be discovered automatically: it is well-known that the first eigenvector of such matrix is the vector z where $z_i \simeq 1$ for i corresponding to the correct identity and $z_i \simeq 0$ for i corresponding to the other identities [37, 42, 43, 49].

Two more challenging situations are possible. The first is where more than one identity is present in the collection and associated with multiple images (Fig. 3, 2nd row). In this case, multiple $z_i \simeq 1$ blocks are present in C. As discussed in the *spectral factorization* literature [37, 43, 49], each block corresponds to a distinct eigenvector of C that is associated with a large eigenvalue (the size of the eigenvalue is proportional to the size of each block) and may thus be identified automatically. The second insidious case occurs when no IDs are prevalent, i.e., when no identity is represented by at least a few images. In this case, no block-diagonal structure is detected, and all eigenvalues are small.

We may thus rely on two signals to automate the analysis of the confidence matrix C [37]. First, the magnitude of the eigenvalues of C indicates the size of the corresponding blocks. The largest eigenvalue indicates the block that is associated with the *correct identity*. If the largest eigenvalue is small, that indicates that no identity is prevalent. Second, the entries of the eigenvectors of C are non-negative for the eigenvectors corresponding to the unit blocks. Thus, eigenvectors with negative entries may be excluded from consideration. There is one rare exception: when two blocks have the same size (i.e., two distinct identities appear in equal numbers), the corresponding eigenvectors will be an arbitrary linear combination of the ideal non-negative eigenvectors and thus may contain negative entries. This case may be resolved automatically by rotating these *twin* (or triplets, or more) eigenvectors to obtain the canonical representation where all entries are non-negative.

We exclude test cases where a single identity cannot be reliably associated with a name. This includes the case where a single identity might not be represented by a single coherent cluster of face images. By discarding this challenging case, our algorithm is expected to slightly overestimate models' accuracies.

Algorithm. In sum, our algorithm to estimate face identity labels relies on simple linear algebra operations on C. First, for each q and s, the eigenvalue decomposition of the corresponding confidence matrix C^{qs} is computed (since C is symmetric, this is the same as computing the principal component analysis of the matrix). Second, the eigenvalues that exceed a threshold $T = 4$ are selected. This guarantees

that at least 4-5 faces are associated with the *correct identity*. If only one eigenvalue meets this criterion, call z^{qs} the corresponding eigenvector, provided that its entries are non-negative. If either multiple eigenvectors or none exceeds T or some of the eigenvector entries are (non-trivially) negative, we discard the name q and the corresponding image collection from further consideration.

At this point, we have computed vectors z^{qs} for each remaining name q and service s. Each such vector contains information on the face identity label, where $z_i \approx 1$ means that image i belongs to the correct identity, while $z_i \approx 0$ means that image i is a spurious identity. We now need to consolidate such estimates across services s. First, we exclude those names q where we do not find exactly one identity for all services s. Second, we exclude names where the number of faces associated with the prevalent identity is smaller than 5. Excluded faces are labeled $\hat{y}_i = -1$. Third, for all included faces, we determine the final estimated label $\hat{y}_i$ by majority voting across services, where we set a threshold $\tau = 0.2$ and, for a given i, if the majority of the $z_i^s > \tau$ over services s then we assume that image i is associated to the correct identity and set $\hat{y}_i = 1$, and otherwise we set $\hat{y}_i = 0$. We have empirically found that aligning the labels this way across services improves the alignment with the annotated face identities for all services (see Fig. S.7 for details).

Error types. Note that by excluding faces, we make our method's test set different from the test set that one would obtain from manual annotation. Thus, the accuracies obtained through the two methods may differ because of two distinct types of errors: **(A)** The error introduced by *excluding* certain queries from the estimation set and thus altering the underlying data distribution ($y_i \neq -1 \wedge \hat{y}_i = -1$); and **(B)** The error from misclassifying *included* faces ($y_i \neq \hat{y}_i \neq -1$). The impact of these two error types will be further studied in Sec. 8.

7 Service evaluation

To compute FNMR-vs-FMR curves, we need to divide each service's confidence values into two distinct sets: the "genuine" and "impostor" distributions. The genuine confidence values correspond to pairs of images that belong to the same identity. The impostor to pairs belonging to different identities. We do this twice: for our method's estimated identities and for the hand-annotated identities so that we may compare performance curves from our method with those from human annotation. We only generate impostor pairs within the same demographic group since same-demographics impostors are the main challenge for FR services. Lastly, we demand that both cross-query images belong to the correct identity to guarantee that the two identities are different – it is (remotely) possible that two images belonging to different queries but not to the correct identity actually belong to the same identity. Fig. 4 (left panel) shows these four distributions for one cloud service and one dataset.

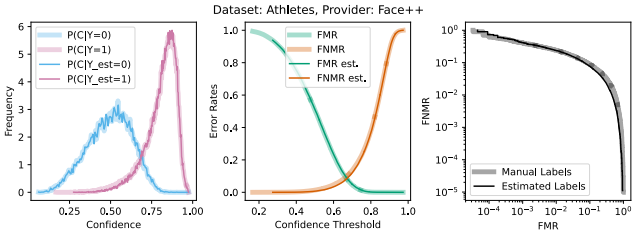

Figure 4. **Sample of service output and evaluation.** (Left) Distributions of confidence values for same-ID face pairs (pink) and different-ID pairs (blue) for the Face++ FRT service. (Mid) FMR and FNMR curves as a function of confidence thresholds. (Right) FMR-FNMR curves. Our method's estimate (thin dark lines) is close to the values obtained through hand-annotation (thick pale lines). These curves are incorporated in Fig. 1 (left). Plots showing the same statistics separately for all datasets and all services may be found in Figs. S.9, S.10.

To be clear, we name G and I the genuine and impostor sets from hand-labeled identities, and $\hat{G}$ and $\hat{I}$ those from our method's estimates. We define the sets as follows, where C_{ij} is the confidence value of an assessed face pair, q is the QueryID, d is the demographic group associated with q, y is the annotated label, and $\hat{y}$ the estimated label:

$$G = \{C_{ij} \mid (q_i = q_j) \ \wedge \ (y_i = 1) \ \wedge \ (y_j = 1)\}$$
$$\hat{G} = \{C_{ij} \mid (q_i = q_j) \ \wedge \ (\hat{y}_i = 1) \ \wedge \ (\hat{y}_j = 1)\}$$
$$I = \{C_{ij} \mid (q_i \neq q_j) \ \wedge \ (y_i = 1) \ \wedge \ (y_j = 1) \ \wedge \ (d_i = d_j)\}$$
$$\hat{I} = \{C_{ij} \mid (q_i \neq q_j) \ \wedge \ (\hat{y}_i = 1) \ \wedge \ (\hat{y}_j = 1) \ \wedge \ (d_i = d_j)\}$$

(the notation is simplified to avoid clutter). Based on these distributions we derive the False Match Rates (FMR) and False Non-Match Rates (FNMR) as functions of the confidence threshold values (Fig. 4, mid panel) as well as FMR-FNMR curves (Fig. 4, right panel).

Semi-supervised setting. While our approach is entirely unsupervised, in principle, it may be adapted to a semi-supervised setting by active sampling, i.e., by collecting human annotations just for a handful of the most ambiguous cases. These may be identified by looking at queries that were dropped by the estimation procedure and by picking samples with a high degree of ambiguity based on their z-scores (Fig. 3, left columns). One such analysis is shown in Fig. S.15.

8 Validation experiments

Does our method work? Does it correctly estimate the accuracy of face recognition systems? We validate our method by measuring the accuracy and bias of three face recognition services and comparing results to traditional hand-annotation. There are two main face recognition tasks: 1:1 matching and 1:n (one-to-many) matching. We focus on 1:1

matching, which is easier to analyze and thus the standard benchmark for accuracy [7, 18, 39].

Two test datasets of face images, "Celebrities" and "Athletes", were defined, face-detected, and manually annotated following the steps described in Sec. 3 and Sec. 4. The distribution of face image sizes is shown in Fig S.4. The athletes' face images are, on average, smaller than the celebrities', and about 10% are smaller than 64 pixels (harmonic mean of width and height), making recognition more challenging. The counts of *correct identity* face images are shown in Fig. S.6 (top row, x-axis) – most of the queries yielded 10-40 *correct identity* useful face images for Celebrities, and 4-40 useful face images for Athletes. The fraction of *correct identity* faces is mostly in the 60-90% for Celebrities and 10-90% for Athletes (Fig. S.6, bottom row, x-axis).

Five commercial cloud services (Amazon Rekognition, Face++, Luxand, Tencent, and Verigram) were used to compute confidence scores, as described in Sec. 5. Histograms of the confidence scores we obtained are shown in Fig. 4 (left) and Figs. S.9, S.10 (left). *While the plots show the confidence scores separately for the correct and false matches, our method has no access to this information and must estimate it.*

Identity labels $\hat{y}$ were estimated following the algorithm presented in Sec 6. Examples of label estimates for specific queries are shown in Fig. 3 and Fig. S.5. Discrepancies between the number of positive labels $y_i = 1$ in the hand annotations and estimated positive labels $\hat{y}_i = 1$ are shown in Fig. S.6. The estimated labels mostly agree with the human-annotated labels (∼99.5% agreement for Celebrities and ∼97.8% for Athletes, see also Tab. S.1). Disagreements may cause our method's estimate to differ from the estimate obtained from the hand-labeled faces (see below).

FNMR-vs-FMR curves are the end result of our process, as shown in Fig. 4 (left) (for Face++), and Fig. 1 (comparing the five services). These plots show the trade-off between false rejects and false accepts for each service. We compare curves obtained through hand-labeling of the dataset and through the estimation that is produced by our method. The fact that hand-labeling and our method produce similar curves shows that our fully automated method may be used to replace the manual (and very expensive) annotations. The discrepancies are caused by two factors: first, some of the queried names q were discarded by the estimation method (Sec. 6, *Type A error*), i.e., the sets of faces on which the services are tested are (slightly) different. Second, as discussed above, there are a few disagreements $\hat{y}_i \neq y_i$ in the label estimates (*Type B error*). Further evaluation shows that while the former is the main cause of discrepancy for the Celebrities dataset, the latter predominantly causes the error when evaluating the Athletes dataset (Fig. S.8). Despite these differences, our estimates closely match hand annotations, slightly underestimating error rates.

9 Accuracy and bias

The accuracy of the five services may be assessed from the FNMR-vs-FMR plots of Fig. 1 as well as Figs. S.9, S.10. The same conclusions on absolute and relative accuracy may be reached both from the hand-annotated test sets and from our method. First, Luxand and Face++ services are markedly less accurate than the other three. Verigram is the most accurate on Celebrities at relevant FMRs (low FMR), and is a tad less accurate than Amazon Rekognition on Athletes. Second, all services are more accurate on Celebrities than on Athletes—this is expected since Celebrities have many well-lit posed photographs and overall good resolution, while the Athletes dataset contains challenging photographs taken during athletic events, where the subjects are wearing sports equipment such as goggles, are grimacing, and the poses are more challenging.

Each identity in our datasets was annotated for gender and race. Therefore, we can estimate demographic biases in the services we test. The FNMR-vs-FMR curves are shown disaggregated by demographic groups in Fig. S.11. To make it easier to understand the biases, we show the equal error rate (FNMR=FMR) for each curve in Fig. 5 where each point corresponds to a demographic group, and the equal error rate estimated by our method is compared to the equal error rate that is computed using hand-labeling of the identities. It is clear from these plots that our method is able to estimate bias accurately when errors are large. When algorithmic errors and biases are small, estimate errors are proportionally larger, possibly due to smaller sample sizes.

Observational methods cannot resolve whether biases are in the algorithm or in the test data [11] (see also in Sec. 2). Since different bias patterns are revealed for Athletes and Celebrities (see Fig. 5, Fig. S.11), it is prudent to assume that biases in the test data are prevalent here.

10 Discussion and conclusions

We have presented a novel method to estimate the accuracy and bias of face recognition services. Our method eliminates the need for hand-annotating the identity of faces in a test set, which is slow, extremely expensive, and can be inaccurate. Dataset annotation is the main blocker for anyone wishing to test face recognition systems' accuracy and bias. An attractive feature of our method is speed since each step is entirely automated after an initial source of names has been chosen. A test, including forming a test set, obtaining confidence ratings from the services to be tested, and analyzing the data to estimate performance, will be completed in about one day (∼2k photos) to four weeks (∼60k photos). We estimate that the alternative, which includes collecting images not used to train face recognition models, as well as hand-labeling and hand-curation of the test set, may take many months. Thus, our method democratizes access to testing face recognition

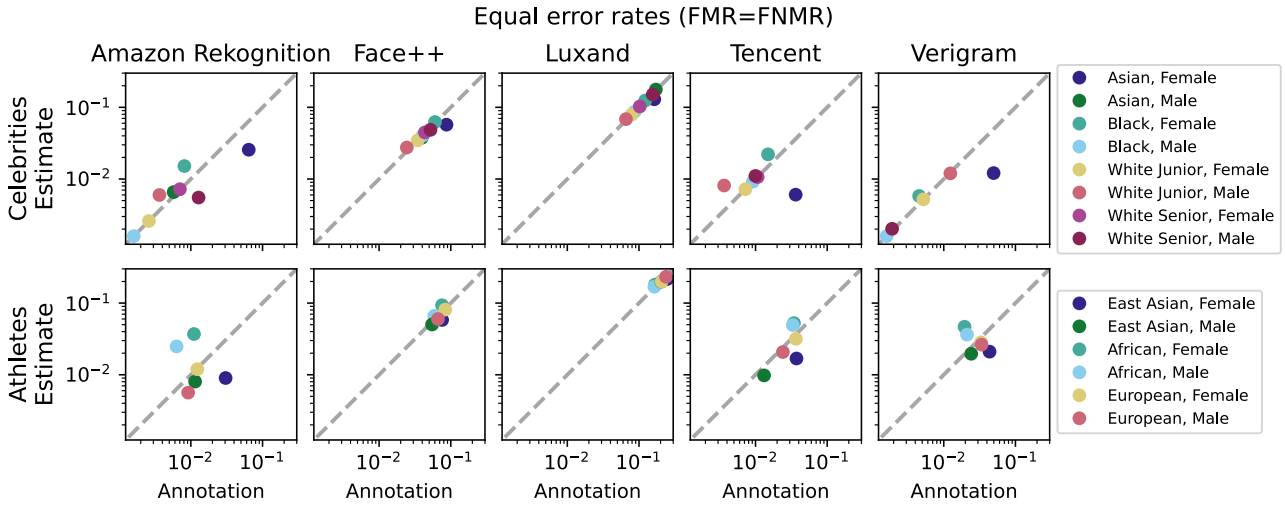

Figure 5. **Measuring bias vis-a-vis gender and race or geographical area**. Our method's (Estimate) vs hand-annotated (Annotation) equal error rate (FMR=FNMR) of the five services computed for each intersectional group defined by gender and race (Celebrities) or geographical area of country (Athletes). Our method correctly detects large biases (i.e., differences in accuracy across demographic groups): see the markedly higher error rates for Asian female celebrities in Amazon Recognition and Verigram, the two more accurate services. Detailed FMR-FNMR curves per demographic group are depicted in Fig. S.11. Confidence intervals, computed using Wilson's method [17], are about 3x larger than the markers (not shown to preserve visual clarity).

systems, a crucial activity in responsible AI. Our method tests services at a certain point in time. Thus, one can discover when and whether a service has changed (see Fig. S.14 for such an analysis). Our system uses face images as transient data and does not require persistent storage of images. In addition, the method can easily be used by a trained operator to select and hand-annotate a fraction of images where the identity label is ambiguous, thus maximizing accuracy while minimizing the additional cost for annotation.

Using our method, we could estimate the error rates and biases of five cloud-based face recognition services quickly and accurately. To the best of our knowledge, this is the first published assessment of the accuracy of cloud-based commercial face recognition systems. The only other available measurements of commercial systems come from the National Institute of Standards and Technology (NIST), which does not directly test cloud-based services but rather relies on standalone code that is submitted to NIST by the vendors.

We validated our method by comparing its estimates with those provided by hand-annotation and found a very close agreement for the Celebrities dataset and good-enough agreement for the more challenging Athletes dataset, where good-enough means that the same conclusions on absolute accuracy, relative accuracy, and bias may be reached.

Some steps in our method could be further refined. First, queries that yield multiple identities may be used rather than discarded since our label estimation scheme can handle multiple identities. Second, simultaneous testing of more than five services ought to improve the majority vote we use to estimate identity labels and thus further improve accuracy.

Our method has limitations. First, in some contexts, using images published online may be undesirable. Second, identity estimation relies on the fact that face recognition services are accurate, and thus, the confidence values they produce are strongly bimodal. If this were not true, then our method would not work. Thus, our method is unsuitable for testing face recognition services in extreme conditions where face recognition models will struggle, e.g., images with very low pixel resolution, resolving the identity of identical twins, faces wearing surgical masks, and extreme age differences. In these conditions, it will always be best to resort to hand-annotated datasets. Third, while our method is designed to reduce potential image overlaps between training and testing data, it does not address the issue of potential identity overlaps [48].

Ethical considerations. Our goal is to enable greater transparency and public awareness of the accuracy of cloud-based face recognition systems (risks and benefits of face recognition technology are discussed in Sec. 1). The images processed by our method are used to evaluate the accuracy and bias of face recognition systems, not to train face recognition models. Our method is designed with privacy in mind [30], and no face image datasets are created or stored. Instead, images are processed on the fly, and any identifiable data is used only for image search and discarded afterward. Our tool will produce aggregated accuracy and bias metrics, minimizing privacy risks.

Acknowledgments. Ignacio Serna's research was supported by the UAM-Santander Young Researcher Program 2022. We thank Verigram for providing access to their face recognition API for the purpose of this research.

References

[1] Amazon Rekognition. https://aws.amazon.com/rekognition. Accessed: 2022-10-15, 2023-02-11 and 2024-06. 4

[2] Face++. https://www.faceplusplus.com/face-comparing. Accessed: 2022-10-15 and 2023-02-11. 4

[3] Lxuand. https://www.luxand.com. Accessed between 2024-04 and 2024-07. 4

[4] Tencent. https://www.tencentcloud.com/products/facerecognition. Accessed between 2024-04 and 2024-07. 4

[5] Verigram. https://verigram.ai/. Accessed between 2024-04 and 2024-07. 4

[6] Vítor Albiero, Kevin Bowyer, Kushal Vangara, and Michael King. Does face recognition accuracy get better with age? deep face matchers say no. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 261–269, 2020. 2

[7] 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. 2, 7

[8] 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. 2

[9] Thomas D Albright. Why eyewitnesses fail. *Proceedings of the National Academy of Sciences*, 114(30):7758–7764, 2017. 1

[10] Thomas D Albright. The U.S. Department of Justice stumbles on visual perception. *Proceedings of the National Academy of Sciences*, 118(24):e2102702118, 2021. 1

[11] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021. 3, 7

[12] Davide Castelvecchi. Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–350, 2020. 1

[13] Edward H Chang, Katherine L Milkman, Dena M Gromet, Robert W Rebele, Cade Massey, Angela L Duckworth, and Adam M Grant. The mixed effects of online diversity training. *Proceedings of the National Academy of Sciences*, 116(16):7778–7783, 2019. 1

[14] Patrick Chiroro and Tim Valentine. An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 48(4):879–894, 1995. 1

[15] Alexandra Chouldechova, Siqi Deng, Yongxin Wang, Wei Xia, and Pietro Perona. Unsupervised and semi-supervised bias benchmarking in face recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 289–306. Springer, 2022. 3

[16] 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*, pages 4690–4699, 2019. 1

[17] Riccardo Fogliato, Pratik Patil, and Pietro Perona. Confidence intervals for error rates in 1:1 matching tasks: Critical statistical analysis and recommendations. *International Journal of Computer Vision*, pages 1–26, 2024. (See also Fig C6 in preprint arXiv:2306.01198). 8, 21

[18] Patrick Grother, Mei Ngan, and Kayee Hanaoka. *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019. 2, 3, 7

[19] Patrick J Grother, Patrick J Grother, and Mei Ngan. *Face recognition vendor test (frvt)*. US Department of Commerce, National Institute of Standards and Technology, 2014. 2

[20] Mark Harris. How facial recognition technology is helping identify the U.S. capitol attackers. *IEEE Spectrum*, January 2021. 1

[21] Peter J Hills and J Michael Pake. Eye-tracking the own-race bias in face recognition: Revealing the perceptual and socio-cognitive mechanisms. *Cognition*, 129(3):586–597, 2013. 1

[22] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 12

[23] Disi Ji, Padhraic Smyth, and Mark Steyvers. Can I trust my fairness metric? Assessing fairness with unlabeled data and bayesian inference. *Advances in Neural Information Processing Systems*, 33:18600–18612, 2020. 3

[24] Kareem J Johnson and Barbara L Fredrickson. "we all look the same to me" positive emotions eliminate the own-race bias in face recognition. *Psychological science*, 16(11):875–881, 2005. 1

[25] 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, 2021. 3

[26] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, 2019. 1

[27] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018. 1

[28] 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. 2

[29] Joseph B Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical society*, 7(1):48–50, 1956. 4

[30] Fabrizio Lala. Data collection via web scraping: privacy and facial recognition after clearview. *i-lex*, 16(2):34–45, 2023. 8

[31] Hao Liang, Pietro Perona, and Guha Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4977–4987, 2023. 3

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 1

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 3

[34] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3, 2001. 1

[35] Sendhil Mullainathan. Biased algorithms are easier to fix than biased people. *The New York Times*, 12/6/2019. 1

[36] National Institute of Standards and Technology (NIST). Face recognition vendor test (FRVT). https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt. Accessed: 2024-12-05. 2

[37] Pietro Perona and William Freeman. A factorization approach to grouping. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June, 2–6, 1998 Proceedings, Volume I 5*, pages 655–670. Springer, 1998. 5

[38] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Castillo Chen, Carlos D. Chellappa, Rama White, David O'Toole, and Alice J. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 1, 2

[39] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018. 7

[40] 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. 3

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1

[42] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 731–737. IEEE, 1997. 4, 5

[43] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 5

[44] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. 3

[45] Jennifer Thompson. I was certain, but I was wrong. *The New York Times, Opinion*, June 18, 2000. 1

[46] Richard Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–359, 2020. 1

[47] Peter Welinder, Max Welling, and Pietro Perona. A lazy man's approach to benchmarking: semisupervised classifier evaluation and recalibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3269, 2013. 3

[48] Haiyu Wu, Sicong Tian, Jacob Gutierrez, Aman Bhatta, Kağan Öztürk, and Kevin W Bowyer. Identity overlap between face recognition train/test data: Causing optimistic bias in accuracy measurement. *arXiv preprint arXiv:2405.09403*, 2024. 8

[49] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. *Advances in neural information processing systems*, 17, 2004. 5

[50] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 1

[51] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Dalong Du, Jiwen Lu, et al. Webface260m: A benchmark for million-scale deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2627–2644, 2023. 3