2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Young Labeled Faces in the Wild (YLFW): A Dataset for Children Faces Recognition

Iurii Medvedev[1] and Farhad Shadmand[1] and Nuno Gonçalves[1]
[1] University of Coimbra, Institute of Systems and Robotics, 3030-194, Coimbra, Portugal

2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG) | 979-8-3503-9494-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/FG59268.2024.10582021

Abstract— **Face recognition has achieved outstanding performance in the last decade with the development of deep learning techniques.**

Nowadays, the challenges in face recognition are related to specific scenarios, for instance, the performance under diverse image quality, the robustness for aging and edge cases of person age (children and elders), distinguishing of related identities.

In this set of problems, recognizing children's faces is one of the most sensitive and important. One of the reasons for this problem is the existing bias towards adults in existing face datasets.

In this work, we present a benchmark dataset for children's face recognition, which is compiled similarly to the famous face recognition benchmarks LFW, CALFW, CPLFW, XQLFW and AgeDB. We also present a development dataset (separated into train and test parts) for adapting face recognition models for face images of children. The proposed data is balanced for African, Asian, Caucasian, and Indian races. To the best of our knowledge, this is the first standardized data tool set for benchmarking and the largest collection for development for children's face recognition. Several face recognition experiments are presented to demonstrate the performance of the proposed data tool set.

I. Introduction

Due to the significant progress of deep learning techniques in the last decade face biometrics has become one of the most accurate biometrics modalities. However, there are still many problems that address modern face recognition. For instance, the most challenging ones are the efficient distinguishing of relatives or twins, the impact of diverse image quality, and age bias.

Here the efficient recognition of children takes an important place. Its necessity is usually motivated by the problem of finding missing or abducted children and combating the children's exploitation. The children's involvement in criminal activity, both as a victim and as an offender is also a problem, where face recognition can be useful. The recognition of newborns is important for protecting from swapping of newborns in hospitals and maternity homes as an alternative biometric solution to accompany currently existing techniques based on RFID technology.

For instance, the annual reports in many countries present a large number of missing children. United Kingdom police forces recorded 50k missing individuals in 2020/21 years [2]. In the United States, an estimated 340k children are reported missing in 2021 [22]. In the same year, the Government of Canada reported an estimated 28k children missing [9].

Another motivation to study face recognition for children's faces is children's exploitation. According to UNICEF in 2021 across 129 countries 4.4 million children had experienced violence (of the 2.3 million for whom disaggregated data are available, 53 percent are girls) with health, social work, or justice/law enforcement services. This number increased by 80% compared to 2017. This significant growth indeed is also related to better availability of support but indeed it also discovers the large latent scale of this problem.

These evidences prove the importance of taking special measures for recognizing children's faces. Indeed the character of aging of children is different compared to adults. The maturation of children involves the nonlinear structure and shape changes of the skull and propositions of parts of a face. For instance, eyes grow rapidly right after birth. Then in several months their growth becomes more linear and undergoes an extra growth spurt during puberty [17], [5]. At the same time, the aging of adults has less extensive character and is usually limited to soft tissue changes like skin texture, hair colour, and wrinkles appearance.

In academia, a significant amount of face recognition benchmarks for adults exist [20], [56], [12], [55], [27], [26], [50], [31], [34]. For the children's age group such testbeds (usually private) were also proposed [15], [7], [6], [36]. However, the standard public tool is still not present in the academic community.

In this work, we address the above problem and present a novel face data toolset, which is specifically focused on children's face recognition. The collected source data consists of wild children's face images of diverse ages, thus we choose the abbreviation YLFW (Young Labeled Faces in the Wild) for referencing our toolset and its components. Our toolset consists of two parts YLFW-Benchmark, and YLFW-Dev for different aspects of face recognition research for the children's age group.

To address the purpose of estimating the performance of face recognition algorithms against children's face images, we propose a YLFW-Benchmark dataset. The dataset consists of $\sim$ 10k images of $\sim$ 3k identities and it is accompanied by a 1-1 verification protocol, which includes 3k match and 3k non-match pairs.

To address the purpose of the development of face recognition systems, which are adapted to children's faces we propose the YLFW-Dev dataset which is split into YLFW-Dev-Train (2k identities, $\sim$ 76k images) and YLFW-Dev-Test ($\sim$ 1k identities, $\sim$ 2k images) with disjoint identities. YLFW-Dev-Test is built similarly to YLFW-Benchmark but

This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020 and PhD grant number 2022.11941.BD.

979-8-3503-9494-8/24/$31.00 ©2024 European Union

includes fewer images and comparison pairs. Its purpose is to support benchmarking in a case when YLFW-Dev-Train is used in the training data (YLFW-Benchmark should not be used in this case since it shares identities and images with YLFW-Dev-Train).

To the best of our knowledge, our data toolset provides the first public benchmark that is directed at estimating the performance against the children's age group faces and the largest training dataset for the respective age group. The important advantage of the proposed data is that it is not collected by the list of identities of celebrities. That is why the face recognition algorithms, which are trained on famous academic datasets of celebrities (like CASIA-Webface[53], VGGFace2[11], MS-Celeb-1M[18]) can be tested on YLFW components following the correct open-set testing scenario (when images in the training and testing parts include images of disjoint identities). Standard benchmarks (like LFW[20], IJB[26], [50], [31] etc.) are usually based on the images of celebrities and can share the same identities with the training data. That is why we also argue that YLFW-Dev-Train indeed can be concatenated with the common academic face datasets for training face recognition networks.

Several features of the proposed data in YLFW should be noted. The proposed YLFW data is race-balanced in different means. Namely, YLFW-Benchmark and YLFW-Dev-Test are race-balanced by the number of pairs across the protocol. The original YLFW-Dev-Train is balanced only by the number of identities per race. That is why we also provide and employ the YLFW-Dev-Train-Balanced, which is obtained by additional augmentation of images of fewer represented races in the original YLFW-Dev-Train.

The gender differences between children are weaker and several works demonstrate that it is harder to estimate the gender of a child compared with an adult [30], [13]. For newborns, the gender property is usually hardly identified by the face image. That is why we do not prioritize gender balance in our work.

It is important to discuss the correlation of the proposed datasets to the cross-age recognition problem. Due to the semi-automatic nature of data collection, our data toolset has only collateral age longitudinal property. YLFW benchmark indeed contains pairs with the age difference, however, we do not explicitly control this effect since the main purpose of the dataset is to estimate the performance biased to children's faces. Namely, the proposed benchmark is not explicitly focused on a cross-age recognition problem. The YLFW-Dev set contains classes, where face images were collected with a large age gap between the sessions (and thus with a perceptible maturation of facial features), however, this is also an uncontrolled effect of the data collecting process. We do not combat this effect since it better approximates the introduced dataset to the real application scenarios.

II. Related Work

To introduce our methodology and results, we first need to discuss recent advances in face recognition, its benchmarking, and specificity when dealing with children's faces.

A. Face Recognition

The ability to learn highly discriminative features from unconstrained images, which is provided by deep learning tools, facilitated the significant development of race recognition technologies in the last decade.

Convolutional neural networks (CNN) have become a standard tool for face recognition due to their high efficiency in solving pattern recognition problems [39]. The training approach of such deep networks can vary, but the target is usually the same - to learn low-dimensional feature domain, where the sample discrimination may be performed with trivial similarity metrics.

Most commonly deep learning face recognition is approached by solving the classification task on the identity-labeled training dataset. The required feature domain (carried by hidden layers) can be learned during optimizing the class probabilities of training samples. This is usually achieved by utilizing Softmax loss and its modifications for classification [44], [43], [45].

The performance of this technique can be significantly improved by increasing intra-class compactness and maximizing inter-class discrepancy from different perspectives. For example, by applying additional regularization for pushing intra-class features to their center [49], or by introducing several kinds of marginal restrictions for inter-class variance [29], [47], [16], [42].

Modern approaches usually consider sample-specific strategies, which allow better control of a feature domain for achieving higher intra-class compactness and inter-class separation. For example, sample wise supervision may be performed by its hardness [54], [21], additional data augmentation applied [41] or even by treating its deep features in a distributional manner (by specifying sample *uncertainty*) [40].

Quality-based loss function adaptations have been intensively studied in recent works. Here the MagFace[33], QualFace[46], [32] and AdaFace[25] share conceptual similarities of the approaches. All these losses indeed modify the marginal-based softmax in a sample-specific way.

Several studies in the field of face recognition have been directed toward investigating compact face recognition models to make them more suitable for practical applications and embed them in portable devices [3], [8].

B. Face Recognition for Children's Age Group

The problem of bias in face recognition is diverse and includes many factors. Age bias is one of the most natural problems due to evident face feature changes during the process of maturation and further aging.

The number of the pioneer works on age-invariant face recognition were performed using the FG-NET database [35], which is composed of a total of 1k images of 82 people with an age range from 0 to 69 and the largest age gap of 45 years. Cross-Age Celebrity Dataset (CACD) [12], is a larger scale age-invariant dataset, which contains around 160k images of 2k celebrities with ages ranging from 16 to 62. However, the representation of children across this dataset is rather limited.

Several works specifically focus on children's face recognition. The general trend in most of the works is also directed at performing a longitudinal study. Many works contributed with manually collected or web-scraped datasets of children's face collections. Usually, the provided datasets remain private. The list of currently existing datasets is presented in Table I.

In one of such works, Best-Rowden et al. proposed a NITL (Newborns, Infants, and Toddlers Longitudinal) face image dataset [7], which was collected by the authors during several sessions. The database contains 314 subjects in total in the age range of 0 to 4 years old. The dataset is race-biased to Indian faces. The experiments of this work proved the extreme complexity of recognizing newborns.

Another dataset Children Longitudinal Face (CLF) [15] was developed to cover a different children's age group for face recognition. CLF contains 3.5 k face images of 1k children in the age group of 2 to 18 years. Within this group, authors observed better performance than in [7] for the age group 0 to 4 years. Another interesting result is that girls in the CLF dataset have higher overall genuine scores and appear to be easier to recognize than boys.

Bahmani and Schuckers proposed the Young Face Aging (YFA) dataset for analyzing the performance of face recognition systems over short age gaps in children [6]. The dataset was collected in the controlled acquisition and longitudinal time conditions and intended to be public via the BEAT research platform [4] (but not available at the moment of publication of this work). The authors demonstrated the positive correlation between face recognition performance decay and the age gap between the gallery and probe images in children, even at the short age gap of 6 months. At the same time, the authors did not observe a significant relationship between gender and match scores in their dataset.

In another work, Jin et al. developed a system for finding missing children without the exposure of photos on the web [23]. The deep face representation, which was trained on face images of adults was fine-tuned on the LCFW dataset, which was collected by the authors. The LCFW dataset is collected by scraping the professional photo album website and contains 60K images with 6K unique identities.

Ricanek et al. proposed the In-the-Wild Child Celebrity (ITWCC) database, which is a collection of longitudinal wild face images of celebrities [36]. It contains 304 subjects and 1705 images. The ages of the subjects within this dataset range from 5 months to 32 years. The authors also reviewed several face recognition algorithms on this dataset and showed that aging in non-adults is a challenging task for face recognition algorithms.

Another research with an emphasis on variations in facial expressions, pose, and illumination conditions was conducted by Dalrymple et al.[14]. Their result Dartmouth Database of Children's Faces contains a set of photographs of forty male and forty female Caucasian children between six and sixteen years of age.

The face recognition in the domain children's age group is partially addressed within the works in the adjacent field of kinship face recognition [1], [37]. These works approach the problem from the perspective of recognizing patterns in faces and familial relationships, which implies the collection of datasets from family photo albums. Those works also contain some representation of children's face image data.

In contrast to several of the above works, we do not explicitly focus on the longitudinal study and mainly consider the problem of age bias. From that perspective, we intend to propose a standardized benchmark for the testing and development of face recognition systems for children's faces.

C. Face Recognition Benchmarks

Modern face recognition deep networks are trained on large labeled collections of face images and the resulting performance is usually estimated on separate datasets with disjoint identities.

The most generic scenario for the benchmarking of face recognition systems is 1-1 verification. The respective benchmarks include the collection of face images with a pairing list (protocol), where each pair is given a match/non-match (by identity) label. In this work, we focus only on considering 1-1 verification.

The most used benchmark data toolsets are usually freely distributed on the web. One of the first and most popular face recognition benchmarks is Labeled Faces in the Wild (LFW), which is a combination of 3k match and 3k non-match pairs [20]. The data in the LFW have natural variability of wild face image characteristics (like pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, and background). However, LFW does not cover all the aspects of the 1-1 face verification performance and also includes unwanted biases (for instance, the average age difference between the matched and non-matched pairs).

That is why several revisited variants appeared. They usually tend to enhance the hardness of correct verification (both inter-class and intra-class), which results in a significant performance decrease. For instance, CALFW (Cross-Age Labeled Faces in the Wild) [56] is designed to reduce the age difference between match and non-match pairs and test the robustness of face recognition algorithms to face aging.

The head pose difference is emphasized in CPLFW (Cross-Pose Labeled Faces in the Wild). This dataset follows the LFW and provides a more realistic consideration of intra-class head pose variation and fosters the research on cross-pose face verification in unconstrained conditions.

Some benchmarks are distributed by the institutional request with a license agreement.

AgeDB dataset [34] was collected to investigate the aging problem in face recognition. It also provides protocols with pairs that include images with different age gaps to test the robustness of face recognition algorithms against aging.

IJB set of benchmarks (IJB-A [26], IJB-B [50], IJB-C [31]) provide a large-scale test (both by the number of images and the length of pair list) on several face recognition tasks, challenges, and scenarios. Face images are collected

TABLE I: Face datasets for children's face recognition

Dataset	Number of Identities	Number of Images	Age, years	Acquisition Type	Availability	Race Balanced	Longi-tudinal
FG-NET [35]	82	1002	6-18	In the Wild	Public	No	Yes
DDCF [14]	80	3200	6-16	Controlled	By Request	No	No
AgeDB [34]	568	16k	1 - 101	In the Wild	By Request	No	Yes
YFA [6]	231	2293	3 - 14	Controlled	To be published	No	Yes
ITWCC [36]	304	1705	3+	In the Wild	Private	No	Yes
NITL [7]	314	3144	3-5	Controlled	Private	No	Yes
CLF [15]	919	3682	2-6	Controlled	Private	No	Yes
LCFW [23]	6k	60k	1.5-9	In the Wild	Private	No	No
YLFW-Benchmark	3069	9810	0 - ∼ 18	In the Wild	Public	Yes	No
YLFW-Dev-Train	2000	75k	0 - ∼ 18	In the Wild	Public	Yes	No
YLFW-Dev-Test	1016	1887	0 - ∼ 18	In the Wild	Public	Yes	No
YLFW-Dev-Train-B.	2000	120k	0 - ∼ 18	In the Wild	Public	Yes	No

with wide variations in head pose, illumination, expression, resolution, and occlusion.

III. Methodology

A. Data collecting

The raw data acquisition follows the generic pipeline of collecting face datasets in computer vision. The face images are web scraped by iterating the list of identity references. In the most straightforward scenario such a list of identities is combined with the names of celebrities [53], [11]. In our work instead of proceeding with the collection by the list of known celebrities, the identities references are web scraped by a specific set of keywords, which refer to the different age and race labels. With the help of anonymous identity references, which are provided by the search engine, we collect the raw and noisy identity labelled set of images. The raw data is then filtered by hierarchical clustering to extract the main cluster for each identity label. The collected face images are subsequently subjected to manual verification to ensure their correspondence to the young age group.

For achieving a race balance the data collection proceeds with a race separation of the requests [38], [51]. We follow Wang et al. [48] in the definition of the race list and consider the following ones: African, Asian, Caucasian, and Indian. In our collecting process, we also observe that available data diversity for races decreases in the following order: Caucasian, Asian, African, and Indian.

Indeed the real race separation is not discrete but smooth across the globe. In our data collecting process, we observe that for the selected list of races the most sensitive gap is between the Asian and Indian races. This observation may be caused by geographical closeness of Asian and Indian ethnic groups and also pitfalls of racial categories definition [24].

B. Pairing methodology

In order to construct the 1-1 verification benchmark the image pair list (protocol) should be defined. In our work, we follow the semiautomatic process, where the proposed pairs are selected randomly. However, the proposed pairs are exposed to a human user verifier, who can accept or reject the pair. The process is repeated until the required number of approved pairs is achieved. All pairs indeed are also manually controlled with the intention to avoid extreme cases of bad quality, extremely easy match pairs, and rare but possible cases of mislabeling in the result data. It is important to note, that this pipeline is not similar to the estimation of the human performance since the cross-label indication is initially available to the person, who performs the verification. The task here is only to confirm the identity match, eliminating occasional labelling errors and image quality outliers.

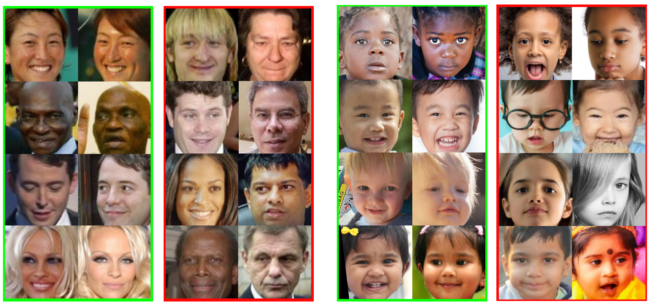
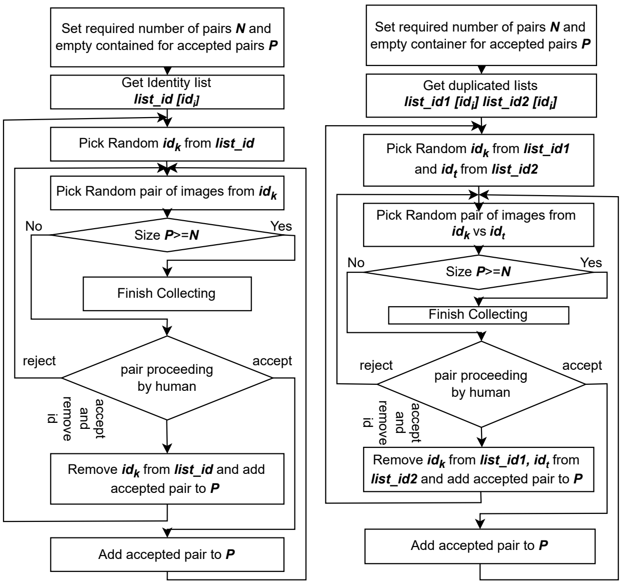

(a)
(b)

Fig. 1: Examples of pairs in LFW(a) and YLFW-Benchmark(b). Match pairs are in the green rectangle. Non-match pairs are in the red rectangle

The exact selection algorithm is slightly different for match and non-match pairs (see Fig 2). In the case of match pairs first, the list of identities is combined. Then a single identity is randomly picked. Then two different images are randomly selected from this identity and exposed to the human user. The user selects one of three options: "accept"; "accept and remove"; "reject". In the case of the "accept" decision, the image pair is appended to the protocol. In the case of the "accept and remove" decision, the respective identity is also removed from the identity list. The "reject" decision results in simple skipping of the proposed pair. After proceeding with the decision the cycle repeats.

In the case of collecting non-match pairs, two equal lists of identities are generated. Then from each list, a random identity is picked. If the selected identities are not equal and this specific combination of identities is not present in the protocol, then for both identities a random image is then


Set required number of pairs N and empty contained for accepted pairs P
Get Identity list
list_id [id_i]
Pick Random id_k from list_id
Pick Random pair of images from id_k
No
Size P>=N
Yes
Finish Collecting
reject
pair proceeding by human
accept
accept and remove id
Remove id_k from list_id and add accepted pair to P
Add accepted pair to P
Set required number of pairs N and empty container for accepted pairs P
Get duplicated lists
list_id1 [id_i] list_id2 [id_i]
Pick Random id_k from list_id1 and id_t from list_id2
Pick Random pair of images from id_k vs id_t
No
Size P>=N
Yes
Finish Collecting
reject
pair proceeding by human
accept
accept and remove id
Remove id_k from list_id1, id_t from list_id2 and add accepted pair to P
Add accepted pair to P

(a) (b)

Fig. 2: Schematics of the pairing strategy for match pairs(a) and non-match pairs(b)

selected to be exposed to the human user as a proposed pair. The user selects from the same decision list as for match-pairs, however here in the case of the "accept and remove" decision the selected identities are removed from their respective lists. After proceeding with the decision the cycle repeats.

C. YLFW-Benchmark database

Following the above pairing procedure for the full collected data, we assemble the YLFW-Benchmark database. To control the race balance this process of pairing is performed separately for different races and their cross combinations (for non-match pairs). Namely, the resulting protocol includes 750 match pairs for each of the selected races and 300 non-match pairs for each cross-race combination. In total, the resulting protocol includes 3000 match and 3000 non-match pairs, which are based on 9810 images of 3069 identities. Several examples of the pairs are presented in Fig. 1.

D. YLFW-Dev database

The goal of the development YLFW-Dev database is to provide separate parts for training and testing. The efficient training of deep networks for face recognition usually requires a large number of images per class (in the ideal scenario 10-50 images per identity [10]).

That is why basing on the collected data we select identities with the largest "samples per class" value to collect 500 identities per race. Thus the resulting YLFW-Dev-Train dataset is then only balanced by the "identities per race" parameter. That is we also design YLFW-Dev-Train-Balanced, which is obtained by random additional augmentations of images of fewer represented races in the original YLFW-Dev-Train. The commonly used types of augmentations were employed to obtain YLFW-Dev-Train-Balanced (horizontal flip, brightness and contrast control, slight image rotation, noise injection).

The remaining part ("tail") of our collected data is used to combine a small benchmark with identities disjoint from the training part. The resulting YLFW-Dev-Test is then assembled similarly to YLFW-Benchmark and includes 1887 images of 1016 identities. Similarly to YLFW-Benchmark the resulting protocol is race balanced and contains 150 match pairs per race and 60 non-match pairs per each cross-race combination (1200 pairs in total).

IV. Experiments and Results

We have performed several experiments with our data toolset. First, we evaluated several recent face recognition models on YLFW-Benchmark. Next, we trained several face recognition models on popular academic face recognition datasets, and their copies concatenated with YLFW-Dev-Train to demonstrate the benefits of such adaptation to the recognition of children's face images.

We report the performance by FNMR@FMR = α and also include additional metrics such as the Equal Error Rate (EER) of Detection Error Trade-off (DET), Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) and Accuracy (in %).

We estimate the performance on our generated tests YLFW-Benchmark and YLFW-Dev-Test in a combination with several similar compact face recognition benchmarks: LFW, CALFW, CPLFW, and ADEDB-30.

The experiments on both YLFW-Benchmark and YLFW-Dev-Test in Section IV-A are performed indeed to compare those two tests and demonstrate compromises, which are made in the development of YLFW-Dev-Test, compared to its larger companion.

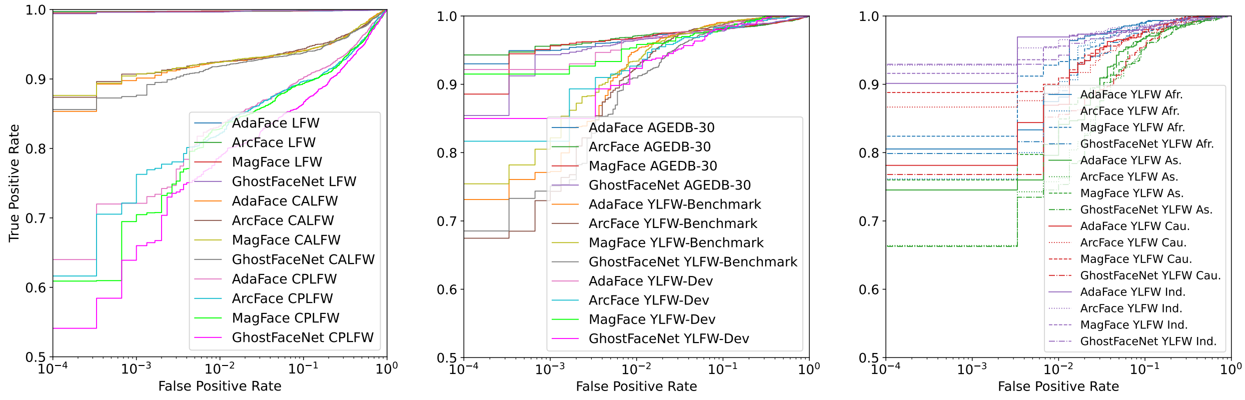
At the same time in Section IV-B, the identities from YLFW-Benchmark can match with the identities of the training data from YLFW-Dev-Train. Thus the results on YLFW-Benchmark are given under the disclaimer that they indeed do not correctly define the performance in the open-set scenario for the cases where YLFW-Dev-Train is used during training.

A. Testing on YLFW-Benchmark

To investigate the properties of the proposed benchmarks and demonstrate the performance of the SOTA face recognition approaches for children's faces we stress them against the YLFW data toolset.

We choose the following set of public deep networks: ArcFace[16], MagFace[33], AdaFace[25], GhostFaceNet[3], which are trained on MS1MV2 dataset [18], [16]. The selected versions of ArcFace, AdaFace and MagFace are based on the ResNet-50 [19] architecture. The alignment settings of the benchmark data correspond to the training data during our tests.

The results for these networks are presented in Fig. 3 and Table II. We observe the imperfectness of the SOTA

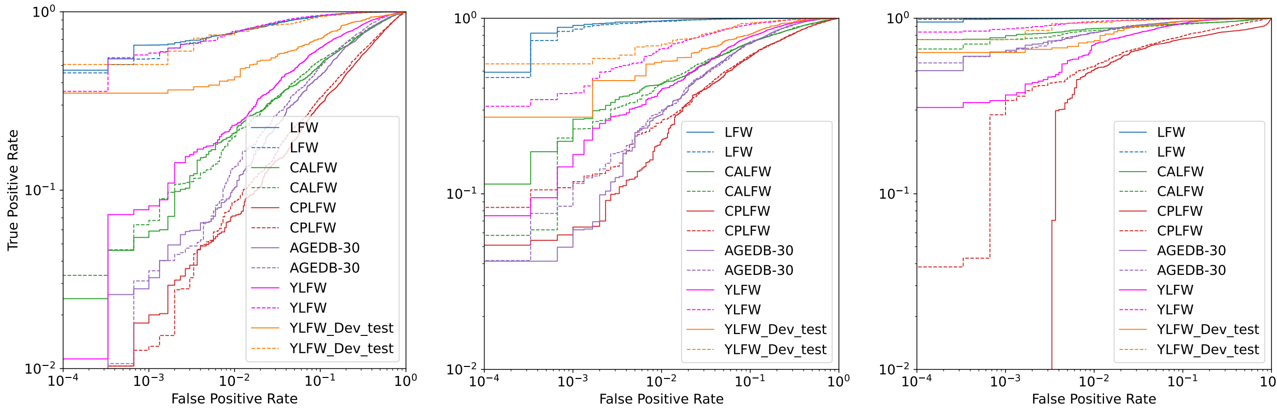


(a) (b) (c)

Fig. 3: ROC curves of different public face recognition models trained on MS1MV2 (AdaFace, ArcFace, MagFace, GhostFaceNet). a) - (LFW, CALFW, CPLFW), b) - (AGEDB-30, YLFW-Benchmark, YLFW-Dev-Test) and c) - (race separated parts of YLFW-Benchmark, where Afr. - African, As. - Asian, Cau. - Caucasian, Ind. - Indian).

TABLE II: Performance metrics for LFW, CALFW, CPLFW, AGEDB-30, YLFW-Benchmark and YLFW-Dev-Test of different public face recognition models trained on MS1MV2 (AdaFace, ArcFace, MagFace, GhostFaceNet).

Train Dataset	FNMR@FMR = α, EER, AUC of ROC and Accuracy (in %)														
	LFW					CALFW					CPLFW				
	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy
AdaFace	0.0020	0.0023	0.0026	0.9988	99.83	0.0590	0.0746	0.0633	0.9725	95.75	0.0989	0.1676	0.0989	0.9515	91.85
ArcFace	0.0013	0.0023	0.0030	0.9990	99.83	0.0566	0.0756	0.0616	0.9722	95.80	0.1029	0.1780	0.1029	0.9478	91.73
MagFace	0.0020	0.0023	0.0026	0.9992	99.80	0.0596	0.0759	0.0639	0.9724	95.80	0.1066	0.1713	0.1049	0.9460	91.40
GhostFaceNet	0.0016	0.0036	0.0036	0.9989	99.76	0.0620	0.0813	0.0659	0.9702	95.46	0.1343	0.2119	0.1246	0.9367	89.81
	AGEDB-30					YLFW-Benchmark					YLFW-Dev-Test				
AdaFace	0.0196	0.0320	0.0256	0.9906	97.89	0.0093	0.0506	0.0276	0.9955	97.41	0.0150	0.0333	0.0266	0.9948	97.83
ArcFace	0.0176	0.0290	0.0250	0.9903	98.08	0.0126	0.0726	0.0300	0.9945	97.13	0.0166	0.0733	0.0333	0.9929	96.75
MagFace	0.0190	0.0320	0.0246	0.9906	98.01	0.0103	0.0659	0.0276	0.9956	97.28	0.0183	0.0416	0.0316	0.9950	97.33
GhostFaceNet	0.0220	0.0356	0.0280	0.9910	97.78	0.0156	0.0906	0.0380	0.9927	96.30	0.0200	0.0766	0.0433	0.9931	96.08



(a) (b) (c)

Fig. 4: ROC curves of ResNet-50 trained on various baseline datasets: a - CASIA-Webface[53], b - VF2 - VGGFace2[11], c - MS1M - MS1MV2[16], [18]. Dashed lines correspond to the networks trained on the baseline + YLFW-Dev-Train data. Networks are tested on LFW, CALFW, CPLFW, AGEDB-30, YLFW-Benchmark and YLFW-Dev-Test benchmarks

TABLE III: Performance metrics for LFW, CALFW, CPLFW, AGEDB-30, YLFW-Benchmark and YLFW-Dev-Test of ResNet-50 network trained on various configurations training datasets (CW - CASIA-Webface [53], VF2 - VGGFace2[11], MS1M - MS1MV2[16], [18] , YDTR - YLFW-Dev-Train-Balanced).

Train Dataset	FNMR@FMR = α, EER, AUC of ROC														
	LFW					CALFW					CPLFW				
	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy	α=10^{-1}	α=10^{-2}	EER	AUC	Accuracy
CW	0.0520	0.2314	0.0721	0.9798	92.83	0.4981	0.7902	0.2753	0.8010	72.65	0.6970	0.9283	0.3721	0.6835	63.23
CW+YDTR	0.0573	0.2351	0.0763	0.9802	92.76	0.4671	0.7993	0.2710	0.8075	73.55	0.6712	0.9131	0.3593	0.6975	64.91
VF2	0.0056	0.0316	0.0193	0.9975	98.10	0.2756	0.5740	0.1896	0.8893	81.58	0.3976	0.7969	0.2376	0.8440	77.13
VF2+YDTR	0.0060	0.0416	0.0216	0.9971	97.83	0.2670	0.5770	0.1889	0.8886	81.91	0.3853	0.7370	0.2339	0.8449	76.85
MS1M	0.0016	0.0036	0.0046	0.9994	99.60	0.0689	0.1256	0.0763	0.9668	93.65	0.2380	0.4880	0.1946	0.8479	83.35
MS1M+YDTR	0.0010	0.0040	0.0050	0.9996	99.60	0.0763	0.1480	0.0816	0.9634	93.01	0.2053	0.4683	0.1636	0.9094	84.73
	AGEDB-30					YLFW-Benchmark					YLFW-Dev-Test				
CW	0.5611	0.8962	0.2820	0.7911	71.95	0.3722	0.7683	0.2231	0.8590	78.18	0.2931	0.5682	0.1781	0.8971	83.00
CW+YDTR	0.5071	0.8642	0.2591	0.8189	74.23	0.0433	0.2221	0.0651	0.9832	93.58	0.0600	0.2261	0.0812	0.9779	92.16
VF2	0.2830	0.7173	0.1736	0.9072	82.80	0.2136	0.6040	0.1506	0.9256	85.10	0.1883	0.4300	0.1383	0.9431	86.58
VF2+YDTR	0.2700	0.6996	0.1783	0.9060	82.51	0.0779	0.3286	0.0879	0.9698	91.31	0.0633	0.2866	0.0799	0.9741	92.66
MS1M	0.0386	0.1706	0.0600	0.9826	94.30	0.0629	0.2900	0.0819	0.9736	91.90	0.0383	0.2533	0.0699	0.9793	93.66
MS1M+YDTR	0.0373	0.1680	0.0606	0.9841	94.23	0.0070	0.0420	0.0243	0.9973	97.61	0.0083	0.0533	0.0300	0.9955	97.41

methods to the proposed in this work benchmarks. YLFW-Benchmark indeed challenges the SOTA methods on par with CALFW and CPLFW. YLFW-Dev-Test has a similar hardness to YLFW-Benchmark at high FMRs ($\sim\!> 10^-2$) and imposes easier challenges at lower FMRs.

According to Fig. 3c, we also observe a significant racial bias in the performance curves. However, we believe that this is related to the negative effect of a difference in the data diversity per race in the original data for YLFW-Benchmark (see Section III-A). Namely, the lower data diversity of a particular race subset indeed leads to the higher similarity of the images in the match pairs.

Low values of FNMR for CALFW and CPLFW are hardly achieved since these benchmarks explicitly decrease the similarity in match pairs by the intrusion of age and pose gap (at the same time this intrusion does not increase the similarity in non-match pairs). The YLFW-Benchmark curve is more similar to LFW and does not contain the above intra-class feature. In our benchmark FNMR start increasing at lower FMR, which means that different children's identities are harder to discriminate. In a general perspective, this conforms to the intuition that children are more similar between each other rather than adults.

B. Experiments on YLFW-Dev

The YLFW-Dev can support the development of face recognition algorithms for children's faces by providing the data for both training and testing. Since the collected images do not belong to celebrities, we argue that our data can be concatenated to popular academic face datasets with low risks of introducing label conflicts.

We perform a set of experiments to demonstrate the efficiency of such a proposal. We train a set of deep CNNs of several data configurations. As a backbone network, we use ResNet-50 [19], which is followed by pooling, dropout, and a dense feature layer with 512 nodes (features). Input images (RGB 3-channel) are aligned as in [16] and resized to 112×112 resolution. The MTCNN detector [52] for used for detecting face and landmarks.

As a training driver, we employ ArcFace, which allows us to learn highly discriminative feature embedding and is robust due to its simple implementation. ArcFace is a marginal modification of the Softmax loss, which is usually formulated as follows:

$$L_{Softmax} = \frac{1}{N}\sum_{i} -\log(\frac{e^{f_{y_i}}}{\sum_{j}^{C} e^{f_{y_j}}}). \tag{1}$$

Here the C is the number of classes (identities), N is a batch size, y_i is the numerical index of the class of the $i-th$ sample, and f_{y_j} is the $y_j - th$ element of the logits vector **f** in the last layer.

The feature layer is usually normalized by L2 constraining the feature embeddings on a hypersphere in $\mathbb{R}^d$ space. Then f_{y_j} can be represented as: $f_{y_j} = w_j^T x_i = \cos(\theta_j)$. ArcFace is then obtained by adding an angular marginal penalization parameter m to the positive logit:

$$L_{ArcFace} = \frac{1}{N}\sum_{i} -\log(\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}) \tag{2}$$

In our experiments in this Section, we use the ArcFace and maintain its margin $m = 0.5$, and the scaling constant $s = 30$.

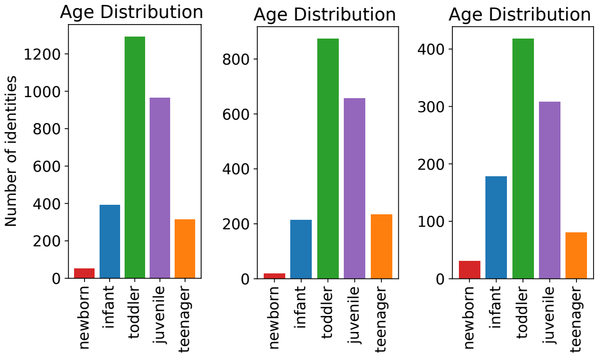
We train the ResNet-50 on three datasets: CASIA-Webface[53], VGGFace2[11], MS1MV2[18], [16] and then repeat the training after concatenating them with YLFW-Dev-Train-Balanced. We set the following hyperparameters: SGD optimizer with linear learning rate scheduling from 0.01 to 0.00001; momentum 0.9; 15 ep. for VGGFace2 and MS1MV2 and 30 ep. for CASIA-Webface; batch size 256.

The results in Fig. 4 and Table III demonstrate the evident improvement in the performance in cases of augmenting the original training data with YLFW-Dev-Train-Balanced. The most significant advance is observed for YLFW-Benchmark and YLFW-Dev-Test as expected. YLFW-Dev-Test better indicate the performance for the training datasets augmented with YLFW-Dev-Train-Balanced since it contains the data with completely disjoint identities from the training dataset in all configurations. As expected the beneficial effect of concatenating with YLFW-Dev-Test decreases with the increase of the number of classes in the original dataset, thus the most effective usage will require proper weighting of the classes from the YLFW-Dev-Train-Balanced.

C. *Age Distribution of YLFW*

To better understand the collected data, we performed age analysis by manually associating each identity in the dataset with an age label. However, it is rather complicated to accurately annotate the age of a child's face by its images without access to biographical data; therefore, we restrict the range of possible labels to several age groups that can be identified by a human: *newborn* (0-3 months), *infant* (3-12 months), *toddler* (1-5 years), *juvenile* (5-13 years), and *teenager* (13-18 years).

The resulting age distribution demonstrates a predominance of the toddler age group across all components of our dataset.



(a) (b) (c)

Fig. 5: Age distribution of YLFW dataset components. YLFW-Benchmark - *(a)*, YLFW-Dev-Train - *(b)*, YLFW-Dev-Test - *(c)*

D. *Human Performance on YLFW-Benchmark*

YLFW-Benchmark protocol was exposed for human performance evaluation. This experiment was performed by five student-volunteers, who manually verified each pair in the benchmark. We estimate the average accuracy across all participants as a measure of human performance evaluation. The results are presented in a form of confusion matrices (see Table IV (a, c, e)). We compare this result with the performance of the 1-1 verification by ArcFace features (same model as in Section III.C) (see Table IV (b, d, f)). The binarized verification scores are generated for the threshold, which corresponds to the point of equal error rate.

For both matchers (human and deep network) the aligned images are exposed to equalize the conditions and remove the impact of image context for human.

To better demonstrate the details of verification performance we also separate the confusion matrices for race matched pairs (Table IV (c,d)) and race non matched pairs (Table IV (e,f)).

TABLE IV: Confusion matrices for matching by human (left column - a, c, e) and ArcFace 1-1 verification (right column - b, d, f) on YLFW-Benchmark. a and b - full set (6000 pairs); c and d - subset of race matched pairs (4200); e and f - subset of race non-matched pairs (1800). (GT P. - ground truth positive, GT N. - ground truth negative, Pred. P. - predicted positive, Pred. N.- predicted negative).

	GT P.	GT N.
Pred. P	0.820	0.076
Pred. N	0.180	0.924

(a)

	GT P.	GT N.
Pred. P	0.97	0.0296
Pred. N	0.03	0.9704

(b)

	GT P.	GT N.
Pred. P	0.820	0.143
Pred. N	0.180	0.857

(c)

	GT P.	GT N.
Pred. P	0.97	0.065
Pred. N	0.03	0.934

(d)

	GT P.	GT N.
Pred. P	0	0.031
Pred. N	0	0.969

(e)

	GT P.	GT N.
Pred. P	0	0.0055
Pred. N	0	0.9945

(f)

Our results demonstrate the extreme hardness for humans to discriminate the unfamiliar children faces. In comparison to recent tests, which demonstrate excellent human performance on face verification [28] several notes should be given. First, we demonstrate face images in the aligned form to the human matcher, which reduce the ability to make decisions with the use of context. Second, children faces in our dataset does not belong to celebrities and were not known by the volunteer that took part in the experiment.

The deep network significantly outperform human in children face image verification task on our dataset.

The results of the verification by human in "match-race" scenario demonstrate that distinguishing between the identities of various children is more challenging then matching the face images of the same children individual (both by human and deep network).

V. Conclusion

In this work, we present a novel face data toolset, which is specifically focused on children's face recognition. Our toolset consists of two parts YLFW-Benchmark, and YLFW-Dev for different aspects of face recognition research for the young age group. To the best of our knowledge, this data proposes the first public standardized benchmark of children's faces in the wild and the largest training dataset of children's faces. The performed experiments demonstrate the imperfectness of modern deep face recognition approaches to children's faces. Also, we show that the use of the presented data can facilitate accurate face recognition in the young age group. We hope that the results of this work will stimulate and boost the research in the area of face recognition.

In future work, we aim to perform the analysis of our data on the gender aspect, develop protocols for 1-N identification and also cover the elder age group with a similar analysis.

VI. Acknowledgements

This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDP/00048/2020 and PhD grant number 2022.11941.BD.

References

[1] The 5th recognizing families in the wild data challenge: Predicting kinship from faces. In V. Struc and M. Ivanovska, editors, *Proceedings - 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021*, Proceedings - 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021. Institute of Electrical and Electronics Engineers Inc., 2021.

[2] N. C. AGENCY". UK Missing Persons Bureau – Missing Persons Data Report 2014/2015 15. https://www.missingpersons.police.uk/en-gb/resources/downloads/missing-persons-statistical-bulletins (accessed November 20, 2022), 2022.

[3] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023.

[4] A. Anjos, L. El-Shafey, and S. Marcel. Beat: An open-science web platform. In *International Conference on Machine Learning (ICML)*, Aug. 2017.

[5] N. Armstrong, W. van Mechelen, and A. D. Baxter-Jones. Growth and maturation. 04 2017.

[6] K. Bahmani and S. Schuckers. Face recognition in children: A longitudinal study. In *10th IEEE International Workshop on Biometrics and Forensics - IWBF 2022*, pages 1–8, 2022.

[7] L. Best-Rowden, Y. Hoole, and A. Jain. Automatic face recognition of newborns, infants, and toddlers: A longitudinal evaluation. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–8, 2016.

[8] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[9] G. O. CANADA". Canada's Missing – 2015 Fast Fact Sheet – MC/PUR Missing child subjects by province, sex and probable cause. http://www.canadasmissing.ca/pubs/2021/index-eng.htm (accessed November 20, 2022), 2016.

[10] J. Cao, Y. Li, and Z. Zhang. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410, 2018.

[11] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[12] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[13] Y. D. Cheng, A. J. O'Toole, and H. Abdi. Classifying adults' and children's faces by sex: computational investigations of subcategorical feature encoding. *Cognitive Science*, 25(5):819–838, 2001.

[14] K. A. Dalrymple, J. Gomez, and B. Duchaine. The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set. *PLoS ONE*, 8, 2013.

[15] D. Deb, N. Nain, and A. K. Jain. Longitudinal study of child face recognition. *2018 International Conference on Biometrics (ICB)*, pages 225–232, 2018.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[17] J. Eccles. The development of children ages 6 to 14. *The Future of Children*, 9:30 – 44, 09 1999.

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Proceedings of ECCV*, volume 9907, pages 87–102, 10 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[21] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[22] F. B. O. INVESTIGATION". NCIC Active/Expired Missing and Unidentified Analysis Reports. https://www.fbi.gov/file-repository/2021-ncic-missing-person-and-unidentified-person-statistics.pdf/view (accessed November 20, 2022), 2021.

[23] X. Jin, J. Lei, S. Ge, C. Song, H. Yu, and C. Wu. Double-blinded finder: A two-side secure children face recognition system. *Wirel. Netw.*, 28(2):927–936, feb 2022.

[24] Z. Khan and Y. Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 587–597, New York, NY, USA, 2021. Association for Computing Machinery.

[25] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18729–18738, 2022.

[26] B. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. C. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[27] M. Knoche, S. Hormann, and G. Rigoll. Cross-quality lfw: A database for analyzing cross- resolution image face recognition in unconstrained environments. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372, 2009.

[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[30] Y. Liu, R. He, X. Lv, W. Wang, X. Sun, and S. Zhang. Is it easy to recognize baby's age and gender? 36(3):508–519, jun 2021.

[31] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[32] I. Medvedev, J. Tremoço, B. Mano, L. E. Santo, and N. Gonçalves. Towards understanding the character of quality sampling in deep learning face recognition. *IET Biometrics*, 11(5):498–511, 2022.

[33] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A universal representation for face recognition and quality assessment. 2021.

[34] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.

[35] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. Cootes. Overview of research on facial ageing using the fg-net ageing database. *IET Biom.*, 5:37–46, 2016.

[36] K. Ricanek, S. Bhardwaj, and M. Sodomsky. A review of face recognition against longitudinal child faces. In *BIOSIG*, 2015.

[37] J. P. Robinson, Z. Khan, Y. Yin, M. Shao, and Y. Fu. Families in wild multimedia: A multimodal database for recognizing kinship. *IEEE Transactions on Multimedia*, 24:3582–3594, 2022.

[38] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. R. Fu, and S. Timoner. Face recognition: Too bias, or not too bias? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2020.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[40] Y. Shi and A. Jain. Probabilistic Face Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[41] Y. Shi, X. Yu, K. Sohn, M. Chandraker, and A. Jain. Towards Universal Representation Learning for Deep Face Recognition. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6816–6825, 06 2020.

[42] J. Sun, W. Yang, J. Xue, and Q. Liao. An Equalized Margin Loss for Face Recognition. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[43] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep Learning Face Representation by Joint Identification-Verification. In *NIPS*, 2014.

[44] Y. Sun, X. Wang, and X. Tang. Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[45] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900, 2015.

[46] J. Tremoço, I. Medvedev, and N. Gonçalves. QualFace: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6, 2021.

[47] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018.

[48] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[49] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing.

[50] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. C. Adams, T. Miller, N. D. Kalka, A. K. Jain, J. A. Duncan, K. E. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017.

[51] H. Wu and K. W. Bowyer. What should be balanced in a ”balanced” face recognition dataset?, 2023.

[52] J. Xiang and G. Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427, 2017.

[53] D. Yi, Z. Lei, S. Liao, and S. Li. Learning face representation from scratch. *ArXiv*, abs/1411.7923, 2014.

[54] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, and T. Mei. NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition. *CoRR*, abs/2007.10172, 2020.

[55] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[56] T. Zheng, W. Deng, and J. Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017.