LCW: Labeled Children in the Wild - FG 2021 Submission

Mats Richter, Leila Malihi, Patricia Windler and Ulf Krumnack
Institute of Cognitive Science, University of Osnabrück, Osnabrück, Germany ****

***Abstract*— Over the years, interest in an age-related dataset has grown significantly. There are several databases according to age, but the main drawback is the noisy labels because most of them are collected semi-automatically. Among this dataset, the dataset related to the age group of children is very few. So in this paper, we propose a novel dataset ”Labeled Children in the Wild” (LCW) for facial recognition tasks, that features a more challenging (high variance) visual domain for modern face recognition solutions. The LCW dataset includes not only the age group of children but also teenagers, minors, and the elderly. The 28,943 images feature high variance on the age of the 1,921 unique identities, image quality, camera technology as well as recording scenario. The dataset further features a similar structure to the LFW dataset, allowing for easy evaluation with preexisting experimental setups.**

Fig. 1. Some images of Labeld Children in the Wild (LCW) face dataset

I. Introduction

Face identification and verification are some of the oldest problems in computer vision, with many different solutions being proposed over the years. Face identification can be described as assigning the correct identity from a list of known identities to a provided image. Face verification, on the other hand, is a binary classification task, where a pair of images is fed into the verification model and classified as matching, if both depict the same identities and non-matching if they depict different identities.

Measuring the quality of identification and verification models adequately and quantifying improvements is crucial for further developing these models. A popular dataset for evaluation is the LFW (Labeled Faces in the Wild) dataset [10]. While this dataset is widely accepted in the community as a benchmark, it could be argued that the dataset is also saturated, with recent models achieving over 99% accuracy [29]. This can be considered problematic, for improvements up to this point have to be very small increments, making the effects of changes in the recognizers hard to quantify. Therefore, more challenging and domain-specific datasets are needed that allow for a more precise assessment of predictive performance, which in turn allows to expand the potential range of applications of such technologies with domain-specific adaption.

Most attention in the field of face identification and verification is devoted to adults, while these tasks across age progression have received very little attention. Studies so far have shown that adult faces are more accurately verified than young children [20], [8]. The tasks of identification and verification become more challenging, the larger the age gap between a reference picture and the image to be identified. Nevertheless, there are several applications that could profit from reliable recognition of children. Starting from image retrieval that could help to organize private photo collections, to support system that can help to find missing children. The United Nations Office on Drugs and Crime reports that by 2018, almost a third of person trafficking is children (from that 19 percent for girls and 15 percent for boys). Child trafficking for forced labor in low-income countries like sub-Saharan Africa for farm work, mining and selling, and in South Asia, to work in brick kilns, and hotels has been reported [27].

The development and evaluation of face recognition systems for children is currently impeded by a lack of suitable datasets. While face datasets for adults abound, especially depicting celebrities and people of public interest, datasets with a relevant fraction of young faces are still rare and limited in size. In this paper, we present an approach to close this gap, by providing a dataset that focuses on young faces, together with evaluation protocols that allow to compare performance of different face recognition systems.

Our proposed dataset has a structure similar to LFW to serve as a drop-in-replacement for LFW. The data is selected to be more difficult, featuring additional challenges that can occur in recognition scenarios. These involve such challenges as strong age differences, ranging from early childhood to very high ages, strong use of costumes and makeup, as well as a strong variety of source media. The dataset features photographs as well as frames from films and movies. The images were recorded under heterogeneous conditions between 1870 and 2019, featuring a wide range of scenarios, camera technology and image quality. LCW has four different age groups, namely child, teenager, minor and adult. This splitting of the dataset results in an easy way to evaluate them. In figure 1, some example images of the LCW dataset with their identities are shown.

II. State of the Art

During the past decades and before the start of deep learning, relatively numerous datasets were defined. The characteristic feature of these datasets was that they were created under completely controlled conditions such as light, background, etc. The justification for these restrictions are the limitations of the models at the time, which were not invariant towards these factors. Some generally utilized datasets were XM2VTS [16], the Multi-PIE database [7], the AR face database [15], the Caltech faces database, the FERET [19], and the Yale face database. Since its introduction in [10], the *Labeled Faces in the Wild* dataset (LFW) has developed as a standard evaluation dataset for face verification, with most current systems still reporting results based on this dataset. This development was supported by the definition of clearly stated evaluation protocols that allow for an objective comparison of different systems. The LFW dataset is a set of face images that are defined for face verification. The dataset contains more than 13,000 images of 5,749 people that have been collected from the internet. The picture labels are the names of the people, and a total of 1,680 people have at least 2 or more pictures. While originally intended to provide training and test data, nowadays, training data is massively extended or completely replaced by data from other sources and only the evaluation part of LFW is used. While originally introduced as a "difficult" dataset, from today's perspective this no longer holds, especially in the "Unrestricted, Labeled Outside Data" protocol, according to which most modern face recognizers achieve more than 99% accuracy.[1]

In recent years, especially when deep learning techniques started to revolutionize the field of face recognition, the demand for larger and more demanding datasets became more pressing and several "in-the-wild" datasets were created at this time. These datasets contain data in uncontrolled conditions with various backgrounds and distinctive lighting conditions and cameras with different resolutions. The PubFig dataset appeared in 2009. It contains a total number of 58,797 images of 200 people, which were collected from the internet [11]. The YouTube Faces (YTF) dataset was defined in 2011. This collection contains 3,425 videos from 1,595 people, all of which have been collected from YouTube. The YTF dataset is intended for studying unconstrained video face recognition [12]. The WDRef database was collected in 2012. This dataset contains more than 99,700 images from 2,995 people. Each person has at least 15 images [1]. CelebFaces dataset [13] contains more than 202,000 images from 10,177 individuals introduced in 2014. The CASIAWebFace database [30] came into being in 2014. This dataset was collected for face verification and face identification tasks and consists of 494,414 images of 10575 different people. The VGG Face dataset [14] consists of 2.6M images of 2,622 identities. The IARPA Janus Benchmark-A (IJB-A), which was introduced in 2015, contains 5,712 images and 2,085 videos of 500 individuals [28]. The MegaFace dataset [17] was introduced in 2015 and 2016, containing 4.7M images from 672K people.

An especially challenging problem is the recognition of faces over a large age gap, for example projecting children faces to their grown-up versions or vice versa, but also face recognition within specific age groups entered the focus of research. Datasets that focus on the facial aging process are AgeDB [1], FG-Net [18], Cross-Age Celebrity Dataset (CACD) [3], Large Age-Gap Face Verification (LAG) [2], Cross-Age LFW (CALFW) [31], MORPH [22], VADANA [25], IMDB-Wiki [23], and AdienceFaces [6]. Most of these datasets contain only a small number of images of subjects below the age of 18. FG-Net and VADANA are rather small in general, with only 82 and 43 subjects, respectively. The MORPH dataset, which was released in 2006, contains 1,724 images of 464 people, and CACD, which was released in 2014, contains 163,446 images of 2,000 subjects. However, both contain no images of subjects younger than 15 years old. LAG contains 3,828 images of celebrities, with at least one young and one old picture of each celebrity, but it does not provide age labels that are more precise than that. AdienceFaces consists of images of non-famous persons with a wide variety of age groups, including young children. However, it is not longitudinal. IMDB-Wiki is a large dataset with 523,051 celebrity images, also depicting many celebrities at different ages. However, the labels in this semi-automatically generated database contain much noise and could, for our purposes, only be used after further manual cleaning.

Longitudinal face datasets that focus specifically on children's aging are very rare. Datasets that fit into this category are In-the-Wild Child Celebrity (ITWCC) by Ricanek et al. [21], ITWCC-D1 by Srinivas et al. [26] and Children Longitudinal Face (CLF) by Deb et al. [4]. ITWCC consists of 1,705 images of 304 subjects whose ages range from 5 months to 32 years. Each subject has at least two images below the age of 16. The images were gathered from the internet by searching for famous child celebrities. The dataset has later been extended by the authors, resulting in the ITWCC-D1 dataset [26], which consists of 7,990 images of 745 subjects, and shows each subjects at at least three different ages, with at most three images per age. These constraints have been introduced in order to keep a balance in the number of images per subject, and across ages. The CLF dataset [4] consists of 3682 images of 919 subjects who are within the age range of 2 to 18. The images of each child were collected over an average time span of 4.2 years. As opposed to ITWCC, the subjects of CLF are non-famous children. All of these three datasets have the major disadvantage that they are not publicly available.

III. Our Dataset

In this paper we introduce a database aiming at the evaluation of face verification and identification systems in different age groups. The base of our dataset consists of established datasets that feature age information in their data for a large number of identities like IMDB-Wiki by [23], AgeDB by [1] and FGNet by [18], as well as additional data

[1] http://vis-www.cs.umass.edu/lfw/results.html

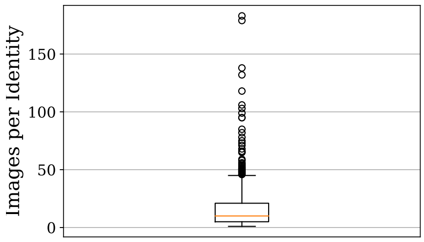

Fig. 2. Distribution of number of images per identity over all age groups.

that were collected from the internet. These datasets were chosen since they provide the most identities and are publicly and freely available at the time of writing of this paper. In this section, we will briefly discuss how the dataset was created from the sources and which cleaning and structuring steps were undertaken to produce the final dataset. The dataset can be found here: https://tinyurl.com/55rzs3t3.

A. *Structure of the dataset*

The dataset consists of 28,943 images belonging to 1,921 unique identities. The number of images per person is not evenly distributed. On average, every identity has 15.06 images associated with it. However, the number of images deviates strongly from person to person, depending on their age and number of publicly available images ($\sigma = 15.06$), as depicted in figure 2. The maximum number of images for one identity is 183 and the median is 10 images. Each image contains only a single person (more exactly, their head).

B. *Creation of the dataset*

First, all images from the source datasets were grouped by their identities. These identities are then manually cleaned, which is necessary since especially the IMDB-Wiki dataset contains inconsistent and very noisy data. The first cleaning step involves removing identities that contain only a single image or do not feature a sufficient age range. We define a sufficient age range for an identity with at least one image depicting them as a teen or younger (below 14 years of age) and as an adult (25 or older). The age information is taken from the metadata of the respective dataset. We then proceed to clean the dataset by removing noisy images. We observe 3 distinct types of noise: The first type are images with wrong identities, the second type are images with wrong age label, and the third type are images with incorrectly drawn bounding boxes. We remove the first and third type of noisy images entirely, since the depicted person cannot be identified in most cases. Images with the second noise type are relabeled into one of four distinct age groups: Baby (0-3 years of age), Toddler (3-6 years of age), Child (7-12 years of age), Teen (13-20 years of age) and Adult (25 and older). Note that there is an age gap between the Teen and the Adult age group, which is intentionally neglected, since the apparent age in this region is very difficult to determine for a large quantity of images. Images that belong into this age gap

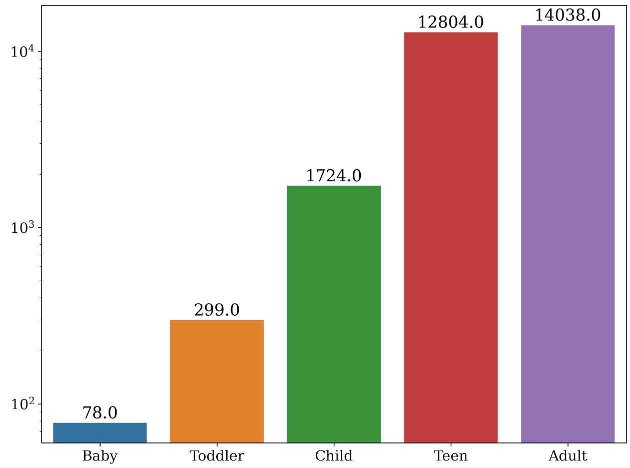

Fig. 3. Number of images per age group. As *Baby* and *Toddler* group are too small for most practical tasks, they are combined with *Children* to from the subset *LCW-Young*. The more encompassing set *LCW-Minor* is formed by also adding *Teenagers*.

were removed. Finally, 6,069 additional images were added manually by image search through the internet to add more data points to identities with insufficient number of images. The faces on these images were cropped in a first step automatically using MTCNN [29]. The resulting bounding boxes were adjusted with a 33% margin to also include hair, ears and neck on the image (which makes manual identification easier). The bounding boxes were manually corrected and cropped. The cropped images serve as data points for the task specific datasets. In an additional cleaning step, noisy bounding boxes depicting false positives or the wrong person (if more than one person is present on the source image) were removed manually. An age label was assigned to these images from the aforementioned categories of Baby, Toddler, Child, Teen and Adult. The final distribution of age labels in the dataset is depicted in figure 3. The imbalance from younger to older age classes is a product of the availability of publicly available data from the existing identities that can be legally used. We make sure that the size of the bounding box relative to the size of the face is similar for all images in the dataset. The cleaning of the bounding boxes involved minor corrections to the coordinates, removing false positives and unwanted true positives (in case another person was on the image), as well as removing face crops of poor quality, on which facial feature were no longer visible or strongly obscured by poor image quality, large sun-glasses or other visual barriers. In a final step, we created task-specific subsets of this dataset based on the images, age and identity labels, which we will discuss in greater detail in the next section.

IV. Task-specific Datasets

In this section, we will briefly discuss benchmark datasets that are derived from the data discussed in the previous section. We will first introduce the structure of the identification datasets and present some basic statistics regarding the data points, as well as the evaluation methodology. Finally, we will introduce the verification datasets along with their evaluation methodology.

A. Identification datasets

For the identification task, we provide four distinct datasets belonging to a specific age group each. The datasets contain only images which depict the person in the respective age range. We provide a single base-line group containing 1,202 identities and consisting only of images of the Adult-Group (LCW-Adult). The public availability of images from specific age groups is a problem that we faced during the development of this dataset (figure 3). Since some previously categorized age groups like Baby, Toddler and Child have significantly less data than other groups, we need to summarize these groups to provide enough data for training and testing. The four datasets are:

1) LCW-Young, containing images of Children, Toddlers and Babies
2) LCW-Teen, containing only Teenagers
3) LCW-Minor, containing Teenagers, Children, Toddlers and Babies
4) LCW-Adult, containing only people belonging to the Adult-Category

All datasets are split with a 10-fold cross-validation split, with an 80% training and a 20% test sets. The split is conducted per identities, meaning that the split is conducted for every identity in the respective dataset. For example, if a Person A has 10 images in the respective dataset, 8 of these would go in the training set and 2 in the test set for an 80% split. This is repeated for all identities. If the identity contains too few images to conduct the split, it is not used for these datasets.

TABLE I

NUMBER OF IMAGES AND IDENTITIES IN THE IDENTIFICATION DATASETS (10 OR MORE IMAGES PER IDENTITY)

Dataset	Number of all Images	Number of identities
LCW-Young	551	38
LCW-Teen	8399	387
LCW-Minor	10077	483
LCW-Adult	11478	483
LFW	4333	161

For evaluation purposes, we treat the identity of a person depicted on an image as the class label of the respective image. This allows us to phrase the identification task as a classification problem. We use the accuracy score for evaluation. The final score on the identification benchmark is computed by averaging the accuracy scores of each test fold. We refer to this score as $\hat{\mu}$. It is computed using the following formula:

$$\hat{\mu} = \frac{1}{10} \sum_{i=1}^{10} P_i \qquad (1)$$

where P_i is the accuracy of the model in the i-th test set of the 10-fold cross-validation.

B. Verification datasets

The aim of the face verification task is to decide whether two given images show the same individual or not. We reuse the same dataset categories as in the previous section, that is Adult, Teen, Minor and Young. The training test split for the images is 80%. However, for verification the split is conducted over the identities and not within in the identities. This means, roughly 80% of the images are contained in the training set, while 20% of the images are in the test set, without an overlap in identities between the images used for generating the training and testing pairs. In order to allow for comparisons of different algorithms on a face verification task, we follow the original LFW protocol and release four lists of image pairs (one for each age group): For the age groups Teen, Minor and Adult, that list contains 6,000 pairs, like in the original LFW protocol, of which 3,000 are positive pairs, depicting the same person in both images of the pair, and 3,000 are negative pairs, depicting different persons. These 6,000 pairs are divided into 10 subsets, each containing 300 positive and 300 negative pairs, and each person appears in only one of the 10 subsets. This ensures that for each training-test-split of each fold, the persons in the test set do not appear in the training set, and persons in the training set do not appear in the test set. For the age group Young, the list contains only 2,340 pairs, with 117 positive and 117 negative pairs in each of the 10 folds. This is due to the lower number of images that were available in this age group. The performance is computed for each of these ten folds via cross-validation, as described earlier, to obtain the mean accuracy $\hat{\mu}$.

V. EVALUATION

In this section, we report the performance of some pre-trained models on these datasets, in order to provide some baseline performances that can be easily reached with pre-trained models.

A. Verification Task

We report the performance of a state-of-the-art face recognition model, namely ArcFace [5] on our face verification paradigm, and compare it to the performance of the model on LFW. Please note that we did not further train the model on our data, but took a pretrained, unofficial implementation of the ArcFace model[2] which achieves a performance on LFW that is similar to the performance reported for the official ArcFace model. The model was pretrained on MS-Celeb-1M [9], a dataset with 10M face images of 1M celebrities. The vast majority of these images show adult faces. Before applying the ArcFace model to LCW and LFW for performance evaluation, we aligned the face images using MTCNN to detect the face and rotating it such that the eyes lie on a horizontal line.

The accuracy of the unofficial ArcFace ResNet50 model on our face verification task can be seen in table II. The table shows the accuracy P_i for 6 different cross-validation

[2]https://github.com/peteryuX/arcface-tf2 (30.04.2021)

splits, as well as the mean accuracy $\hat{\mu}$ over all 10 splits. One can see that the accuracies on LCW are lower than on LFW, which indicates that LCW is more challenging in general, due to higher variance in pose, illumination, make-up and glasses. When comparing the accuracies on the different age groups of LCW, one can see that the performance on LCW-Adult is the highest, as adult faces are less challenging than child faces. LCW-Young and LCW-Teen are roughly equally challenging, while LCW-Minor is most challening. This might be due to the fact that LCW-Minor includes larger age gaps than LCW-Young and LCW-Teen. To summarize, it can be seen that face recognition on children and teenagers, especially with age gaps, poses a challenge to existing face recognizers, which shows that further research and development on this topic is needed.

TABLE II

COMPARISON OF THE ACCURACIES P_i FOR THE VERIFICATION TASK WITH ARCFACE ON THE DIFFERENT AGE GROUPS OF LCW AND ON LFW

Dataset	i=1	i=3	i=5	i=7	i=9	i=10	$\hat{\mu}$
LCW-Teen	92.7	90	89	92.7	91.7	91.7	91.37
LCW-Young	89.7	88.9	96.6	86.3	89.7	94.4	91.11
LCW-Minor	88.2	87.2	90.3	91.2	92.8	90.5	90.4
LCW-Adult	94.3	94.3	92.8	95.7	93.5	94.5	93.42
LFW	99.5	99.2	99.2	98.8	99.7	99.8	99.47

B. Identification task

In this section we report on the application of VGGFace [14] developed for providing baseline accuracies for the different subsets of the LCW dataset. The idea is to apply a VGG16 model [24] on all identification sets. Due to the small size of the data, it is unlikely that a robust model can be fully trained on this set alone. For this reason, we use a model that was pre-trained on VGGFace [14] and fine-tune its weights by training on our dataset. VGGFace consists of 2.6M images of 2,622 identities. To put the results of the LCW identification tasks into a frame of reference, we evaluate the model on LFW as well. Table III shows the number of all images and the number of all identities in each group for this subset.

TABLE III

DATA POINTS IN TRAINING AND TEST FOLDS OF THE IDENTIFICATION DATASETS

Dataset	Train	Test	Number of identities
LCW-Young	1369	732	552
LCW-Teen	9570	3234	1497
LCW-Minor	11234	3671	1595
LCW-Adult	10672	3363	1202

The ROC curves for both LFW and LCW face identification protocol are shown in figure 4. The mean accuracy of correct classification, as well as 6 exemplary splits of both datasets is shown in table IV. The results show that LCW is in all scenarios significantly more difficult than LFW,

(a)

(b)

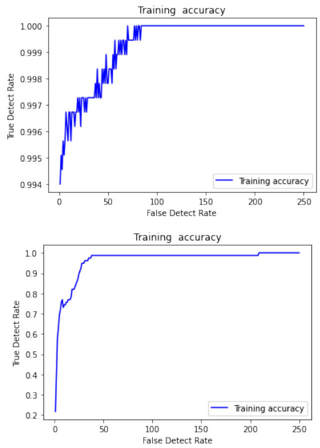

Fig. 4. Face identification results on both dataset, (a) LCW (b) LFW

when considering $\hat{\mu}$. We attribute this to the high visual heterogeneity of our dataset, which contains photographs taken in the last 100 years, resulting in a high variance in the visual domain.

TABLE IV

COMPARISON OF THE ACCURACIES P_i FOR THE IDENTIFICATION TASK ON THE DIFFERENT AGE GROUPS OF LCW AND ON LFW

Dataset	i=1	i=3	i=5	i=7	i=9	i=10	$\hat{\mu}$
LCW-Teen	77	80	83	84	82	85	82
LCW-Young	84	82	83	83	81	85	84
LCW-Minor	76	82	85	83	86	84	83
LCW-Adult	80	75	82	83	87	84	82
LFW	92	94	93	95	92	95	93.5

VI. CONCLUSION

We introduced a novel dataset ”Labeled Children in the Wild” (LCW) as a drop-in replacement for the classical LFW dataset for evaluating facial recognition tasks. Our dataset exhibits a higher variance and is supposed to be a more serious challenge for modern face recognition solutions. Our LCW dataset consists of 28,943 images of 1,921 unique identities, featuring high variance in age, image quality, camera technology as well as recording scenario. We provide task-specific subsets for identification and verification tasks grouped by the apparent age of the people depicted on the respective images. This allows the practitioner to investigate potential age-related biases in their models and (in case of the identification task) fine-tune them. The dataset is distributed in a similar structure as the original LFW dataset, allowing for easy evaluation with preexisting experimental setups.

REFERENCES

[1] J. D. Athanasios Papaioannou, Christos Sagonas and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[2] S. Bianco. Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90:36–42, 2017.

[3] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

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

[5] J. Deng, J. Guo, N. Xue, and S. 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.

[6] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2008.

[8] P. Grother and M. Ngan. Face recognition vendor test: Performance of face identification algorithms. *NIST Interagency report*, 8009(5), 2014.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[10] 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.

[11] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2009.

[12] T. H. Lior Wolf and Itay. Agedb: The first manually collected, in-the-wild age database. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] R. C. Malli. Kkeras-VGGFace: VGGFace implementation with keras framework. Technical report, University of Oxford, 19-Oct-2017.

[15] A. Martínez and R. Benavente. The ar face database. Technical Report 24, Computer Vision Center, Bellatera, Jun 1998.

[16] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second International Conference on Audio and Video-based Biometric Person Authentication*, March 1999.

[17] D. Miller, I. Kemelmacher-Shlizerman, and S. M. Seitz. Megaface: A million faces for recognition at scale. *CoRR*, abs/1505.02108, 2015.

[18] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the fg-net ageing database. *Iet Biometrics*, 5(2):37–46, 2016.

[19] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image Vis. Comput.*, 16(5):295–306, 1998.

[20] N. Ramanathan and R. Chellappa. Face verification across age progression. *IEEE Transactions on Image Processing*, 15(11):3349—-3361, 2006.

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

[22] K. Ricanek and T. 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.

[23] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations (ICLR)*, 2015.

[25] G. Somanath, M. Rohith, and C. Kambhamettu. Vadana: A dense dataset for facial image analysis. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 2175–2182. IEEE, 2011.

[26] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King. Face recognition algorithm bias: Performance differences on images of children and adults. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[27] United Nations Office on Drugs and Crime. Global report on trafficking in persons, 2020.

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

[29] Z. Yang and R. Nevatia. A multi-scale cascade fully convolutional network face detector. In *23rd International Conference on Pattern Recognition (ICPR)*, pages 633–638, 2016.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch, 2014.

[31] 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.