# Analysis of Race and Gender Bias in Deep Age Estimation Models

Andraž Puc, Vitomir Štruc, Klemen Grm
University of Ljubljana, Faculty of Electrical Engineering
Tržaška cesta 25, SI-1000 Ljubljana, Slovenia
klemen.grm@fe.uni-lj.si

***Abstract*—Due to advances in deep learning and convolutional neural networks (CNNs) there has been significant progress in the field of visual age estimation from face images over recent years. While today's models are able to achieve considerable age estimation accuracy, their behaviour, especially with respect to specific demographic groups is still not well understood. In this paper, we take a deeper look at CNN-based age estimation models and analyze their performance across different race and gender groups. We use two publicly available off-the-shelf age estimation models, i.e., FaceNet and WideResNet, for our study and analyze their performance on the UTKFace and APPA-REAL datasets. We partition face images into sub-groups based on race, gender and combinations of race and gender. We then compare age estimation results and find that there are noticeable differences in performance across demographics. Specifically, our results show that age estimation accuracy is consistently higher for men than for women, while race does not appear to have consistent effects on the tested models across different test datasets.**

## I. Introduction

Age estimation from facial images (illustrated in Fig. 1) has seen increased interest from the machine learning and computer vision communities recently [3], [19], [22], [29]. The possibility of determining age from a face image automatically and with high accuracy can facilitate applications with considerable market potential, such as detecting minors for legal purposes or adjusting application interfaces based on users' age. However, it is paramount to understand the behaviour of existing age estimation models, especially with respect to their performance across different demographic groups, when deploying them in real-life applications .

While existing work has looked at the impact of demographics when evaluating new models, e.g., [4], this has mostly been a side result of the overall experimental evaluation. Studies focusing specifically on demographic model bias, on the other hand, are still limited in the literature. In this paper, we try to fill this gap and study the impact of race and gender on the accuracy of contemporary deep age estimation models. Specifically, we experiment with two pre-trained off-the-shelf age estimation models and evaluate their performance on two publicly available datasets. The main contribution of our work are important findings that help to better understand age estimation models and their performance on different sub-groups of subjects, such as:

- We report results that suggest that age estimation with the tested models is more accurate for male subjects than for female subjects. While we observe opposite settings for certain races, the age estimation models seem to favor men over women in term of estimation errors in general.
- We observe no consistent impact of race on age estimation accuracy. While different race groups produce consistent performance variations with all tested models, these appear to be inconsistent between different test datasets, suggesting that other nuisance factors affect results to a greater extent than race.

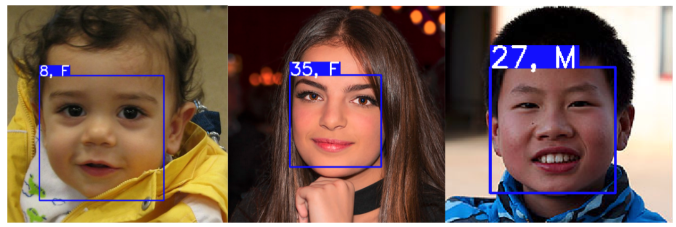


Fig. 1: Age estimation from face images has progressed considerably in recent years with state-of-the-art models producing highly accurate estimation results. In this paper we analyze and compare age estimation performance across different demographic groups in terms of both gender and race.

The rest of the paper is structured as follows: In Section II we briefly review existing work related to our study. Next, we elaborate on the methodology used in the paper in Section III and discuss experimental findings in Section IV. We conclude the paper with some final comments in Section V.

## II. Related work

In this section, we present prior work that relates to our study. We first discuss existing techniques for age estimation, then review existing studies on the impact of gender and race in various face-related tasks and finally elaborate on existing work on bias in age estimation models.

**Age estimation.** One of the early attempts at age estimation was presented by Kwon *et al.* in [17] and used an active contour snakelet model that focused on wrinkles and simplified the age estimation task into a binary classification problem. In [18], Lanitis *et al.* described an automatic age estimation approach relying on Active Appearance Models (AAMs) to jointly extract shape and texture information from an input face. Later Geng *et al.* [11] proposed a new approach by modeling aging patterns with representative sub-spaces. Guo *et al.* [14] used bio-inspired features (BIFs) and a multilayer

HMAX model for age estimation. Chang *et al*. [6] proposed replacing traditional multi-class labels with new ordinally arranged labels [10], [30] and developed a cost-sensitive ordinal ranking framework for age estimation. Chao *et al*. [7] proposed an age-oriented local regression algorithm that resulted in considerable performance, but already highlighted concerns regarding demographic imbalances in the training data.

Recent age estimation models are increasingly based on CNNs. Yan *et al*. [29], for example, reported impressive results by building a multi-layer CNN model for feature extraction and a Support Vector Machine (SVM) for classifying faces into different age groups. Levi *et al*. [19] presented a relatively simple CNN architecture for age estimation that can outperform previous methods if trained with sufficient training data. Niu *et al*. [22] presented a CNN model for joint feature learning and regression modeling, capable of making better use of large datasets. Overall, the use of CNNs greatly improved age estimation accuracy, however, problems with race and gender disparities in results are still present with these models.

**Gender and ethnicity covariates.** Covariates are variables that either increase intra-class variation or decrease inter-class variation [1] and, hence, affect the performance of machine learning models. While pose, occlusion or illumination can often be controlled, other covariates like race and gender cannot. Early work on face-related tasks, such as the Local Binary Pattern (LBP) model for demographics classification proposed by Yang et al. [31], already had trouble dealing with ethnicity and gender. The face recognition meta-analysis conducted by Lui *et al*. [20], examined 25 studies and showed that there is no general consensus on the influence of gender and race. However, the study found that the uneven distribution of subjects across age groups in datasets is a big problem. Datasets are dominated by younger subjects [15], since they rely on (typically student) volunteers [1]. Solutions were later proposed to make the trained models more robust to problematic covariates, such as partitioning the datasets into equal-sized sets with respect to gender and ethnicity [26] or using a framework for additional ethnicity and gender pre-classification, as proposed by Guo et al. [13]. Abdurrahim *et al*. [1] suggest that men and women have different local features, however, girls and boys have similar craniofacial features. Most results confirm that women are harder to recognize than men, however, with age, the difference diminishes [21]. The study in [20] was unable to determine which race or what qualities prove to be a problem for current models, which is a recurring issue among covariate analyses [12]. Drozdowski *et al*. [9] present a comprehensive survey of the challenges associated with algorithmic bias in biometric applications.

**Bias in age estimation models.** A handful of existing studies investigated the issue of bias in age estimation models. Xing *et al*. [28], for example, analyzed the performance of their model across gender and ethnicity sub-groups. Alvi *et al*. [4] suggested that training datasets that are not balanced in terms of gender can lead to age estimation models that are gender-biased. Clapes *et al*. [8] showed that there is some consistent bias across various demographic groups when relating the performance of apparent and real-age estimation tasks. In these paper we contribute to a better understanding of the bias of age estimation models with an analysis of the performance of two recent deep learning models with respect to race and gender.

## III. Methodology

We now present the methodology used in the evaluation. We discuss the age estimation models used, the experimental datasets and setup and finally present the performance measures used for our analysis.

**Age estimation models.** We use the following (pre-trained) off-the-shelf age estimation models for the experiments:

- *WideResNet:* Our first model[1] is based on the Wide Residual Network (WideResNet) architecture [24], [32], but has two classification layers for age and gender. WideResNet are similar in spirit to residual networks, but feature ResNet blocks with decreased depth and increased width. We use two variants of the WideResNet model for our analysis: the first is trained on the UTKFace dataset [33] and the second on the IMDB-WIKI dataset [23]. We denote these two models as WideResNet-UTK and WideResNet-IMDB, respectively. Both models are trained from scratch using a multi-task learning objective including both age estimation and gender recognition.
- *FaceNet:* Our second model[2] is based on the FaceNet architecture [27], which is one of the first deep CNNs optimized for face recognition. The model is initialized with weights of a FaceNet model trained for face recognition on the VGGFace2 dataset (featuring around 3.3 million faces and 9000 identities [5]). Similarly to the second WideResNet mode described above, the model then trained (or better said fine-tuned) for the tasks of age estimation and gender recognition on the IMDB-WIKI dataset.

The models above were selected for our analysis because of their state-of-the-art performance and the fact that two of the models have a different architecture, but were trained on the same dataset (WideResNet-IMDB and FaceNet), while two share the same architecture, but were trained on different datasets (WideResNet-IMDB and WideResNet-UTK).

**Experimental datasets.** We select two popular datasets for age estimation for the experiments:

- *The APPA-REAL dataset* [2] contains $7,591$ images with associated real and apparent age labels. The age range of subjects on the pictures is between 0 and 95 years. The dataset provides annotations with information about various covariates of the pictures, including gender and ethnicity [8]. The annotations partition the data into three race classes: Caucasian, Asian and Afro-American.
- *The UTKFace dataset* [33] is a relatively large face image dataset with subjects aged from 0 to 116 years. The

[1] Available from https://github.com/yu4u/age-gender-estimation
[2] Available from https://github.com/BoyuanJiang/Age-Gender-Estimate-TF

Fig. 2: Sample images from APPA-REAL (top) and UTKFace (bottom). Both datasets contain images of varying quality, different head poses, light settings, and facial expressions.

dataset consists of $23,708$ face images captured "in-the-wild" that cover a large variety of poses, illumination, occlusions, resolution, and facial expressions. The images are labelled by age, gender, and ethnicity and include five ethnicity categories: White, Black, Asian, Indian and Others. The ground truth of these labels was estimated by the Deep Expectation (DEX) algorithm [25] and then checked by human annotators.

Both datasets are widely used for age estimation and are free for non-commercial use. Since we use pre-trained off-the-shelf models for the analysis, no training data is required. We, therefore, use all available image data from the two datasets as the test data for experimentation. A few example images from the two datasets are shown in Fig. 2.

**Experimental setup.** To evaluate age estimation performance while also analysing race and gender bias, we partition the test data into different groups of interest, as also illustrated in Fig. 3. For analysing gender bias, we simply separate images into male (M group) and female (F group) categories based on the available image annotations. For analysing race bias, we similarly generate race categories. The number of race categories is defined by the available race labels in both datasets. As already indicated above, the UTKFace dataset provides 5 race labels, denoting White (W), Black (B), Asian (A), Indian (I), and Others (O) subjects. The APPA-REAL dataset only provides three separate race labels for Caucasian (W), Afro-American (B) and Asian (A). Furthermore, we generate combined gender-race sub-groups by additionally separating each race group by gender. In doing so, we intend to produce specific results for these sub-groups and provide an explicit comparison between them. This helps us determine if any sub-group stands out and has a specifically large impact on the performance of any particular group (e.g., whether white males affect the results for males the most). Since we can compare both genders within the same race group and vice versa, we can also investigate whether the models generate consistent deviations in performance throughout all groups, e.g., if one gender consistently over- or under-performs when compared to the opposite gender for a given race.

**Performance metrics.** In order to evaluate and compare the performance of the selected age estimation models on various demographic sub-group, we report Mean Absolute Error (MAE) scores, which serve as indicators of the average performance of the age estimators:

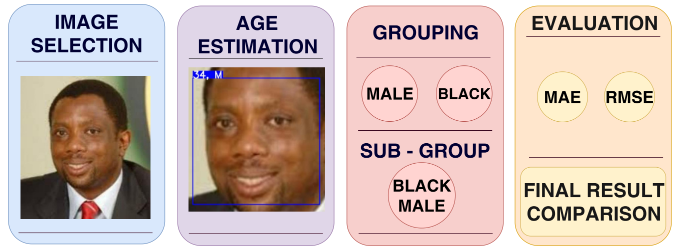


Fig. 3: Schematic demonstration of the methodology used to analyze gender and race bias of deep age estimation models.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|, \tag{1}$$

where $n$ denotes the number of all test images, and $\hat{y}_j$ and $y_j$ represent the predicted and the ground truth age, respectively.

Additionally, we also report the Root Mean Squared Error (RMSE), which emphasizes larger age estimation errors and penalizes them more:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|^2}, \tag{2}$$

where $n$, $\hat{y}_j$ and $y_j$ again represent the same variables as in Eq. (1). The two metrics represent the literature standard when evaluating age estimation models [29], [11], [3], [22], [16].

## IV. Experimental Results

In this section we present our experimental procedure in analysing gender and race bias of the considered CNN models. We start the section by describing our experimental setup and the used performance metrics and then present our final results.

**Baseline age estimation performance:** In the first series of experiments, we assess the performance of the three models over all available test data of the UTKFace and APPA-REAL. We test all models on the APPA-REAL dataset, but exclude the WideResNet-UTK models from the experiments on the UTKFace dataset, because this datasets was used to train the model.

The MAE and RMSE values reported in Tables I and II show estimation errors averaging between 6 and 10 years in terms of MAE and between 9 and 14 years in terms of RMSE. These results are a little above the current state-of-the-art, which we ascribe to the preprocessing procedure, where we do not explicitly align faces based on landmarks after the detection step. However, the absolute values of the performance metrics are not critical, as our focus in this study is on the relative comparison of the performance scores across different demographic groups and sub-groups.

Overall, we observe that the error scores for the APPA-REAL dataset are lower than for the UTKFace dataset, which points to a difference in the difficulty of the two datasets and a better fit of the off-the-shelf models for the type of data present

TABLE I: MAE and RMSE values (in years) for different race and gender groups. The groups are labelled with first letters for gender: Male (M) and Female (F), and race: White (W), Black (B), Asian (A), Indian (I) and Others (O). APPA-REAL does not have Indian (I) and Others (O) categories.

| Model | Test dataset | Subgroup Division & Performance Metrics | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gender | | | | Race | | | | | | | | | | |
| | | MAE (yrs.) | | RMSE (yrs.) | | MAE (yrs.) | | | | | RMSE (yrs.) | | | | | |
| | | M | F | M | F | W | B | A | I | O | W | B | A | I | O |
| **WideResNet-UTK** | UTKFace | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| | APPA-REAL | 6.45 | 7.72 | 8.67 | 10.20 | 7.03 | 7.69 | 7.36 | - | - | 9.41 | 9.94 | 9.65 | - | - |
| **WideResNet-IMDB** | UTKFace | 8.83 | 8.91 | 12.10 | 11.90 | 9.79 | 7.71 | 9.56 | 8.02 | 6.99 | 13.60 | 11.10 | 13.40 | 11.00 | 9.39 |
| | APPA-REAL | 6.89 | 8.01 | 8.99 | 10.40 | 7.38 | 8.65 | 7.70 | - | - | 9.63 | 10.60 | 10.10 | - | - |
| **FaceNet** | UTKFace | 8.22 | 8.20 | 11.20 | 11.50 | 8.22 | 7.25 | 10.4 | 7.66 | 7.68 | 11.20 | 10.10 | 14.30 | 10.20 | 10.70 |
| | APPA-REAL | 7.69 | 7.95 | 10.70 | 11.10 | 7.79 | 8.23 | 7.85 | - | - | 10.90 | 10.50 | 10.60 | - | - |
| **Average** | UTKFace | 8.52 | 8.55 | 11.6 | 11.7 | 9.01 | 7.48 | 9.98 | 7.84 | 7.34 | 12.40 | 10.60 | 13.9 | 10.60 | 10.10 |
| | APPA-REAL | 7.01 | 7.89 | 9.45 | 10.57 | 7.40 | 8.19 | 7.64 | - | - | 9.98 | 10.30 | 10.10 | - | - |

TABLE II: MAE scores (in years) for demographic sub-groups divided by both gender and race. The labels represent combined gender-race sub-groups with the first letter encoding the gender and the second letter the race - Male (M), Female (F) and White (W), Black (B), Asian (A) Indian (I) and Other (O), e.g., Male Asian - MA.

| Model | Test dataset | MAE of Race-Gender Divided Sub-Groups (yrs.) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MW | FW | MB | FB | MA | FA | MI | FI | MO | FO |
| **WideResNet-UTK** | UTKFace | - | - | - | - | - | - | - | - | - | - |
| | APPA-REAL | 6.33 | 7.75 | 7.47 | 7.89 | 7.36 | 7.37 | - | - | - | - |
| **WideResNet-IMDB** | UTKFace | 9.08 | 10.70 | 7.72 | 7.69 | 11.30 | 8.10 | 8.26 | 7.71 | 6.99 | 6.98 |
| | APPA-REAL | 6.80 | 7.98 | 8.13 | 9.12 | 7.47 | 7.89 | - | - | - | - |
| **FaceNet** | UTKFace | 7.89 | 8.61 | 7.13 | 7.36 | 11.90 | 9.06 | 7.84 | 7.42 | 7.42 | 7.90 |
| | APPA-REAL | 7.66 | 7.94 | 8.19 | 8.26 | 7.69 | 7.98 | - | - | - | - |
| **Average** | UTKFace | 8.49 | 9.63 | 7.43 | 7.53 | 11.60 | 8.58 | 8.05 | 7.57 | 7.21 | 7.44 |
| | APPA-REAL | 6.93 | 7.89 | 7.93 | 8.42 | 7.51 | 7.75 | - | - | - | - |

in APPA-REAL. When comparing models, we notice that the two WideResNet models perform similarly regardless of the training data. The performance difference between the models is minimal with a slight edge for the WideResNet-UTK model. The FaceNet model, on the other hand, outperforms both WideResNet models on the APPA-REAL data, but performs worse than WideResNet-IMDB on the UTKFace dataset.

**Gender group comparison:** Comparing the results reported in Table I for each gender, we observe noticeable differences in the calculated MAE and RMSE scores. Male subjects result in more accurate age predictions with both WideResNet models when tested on the APPA-REAL dataset regardless of the data used to train the models. Here, the MAE differences for the two genders are in the range of a $1.5$ years. The results for the FaceNet model show less divergence between genders, but still slightly favor male subjects over females. The performance difference is significantly smaller on the UTKFace dataset. On this dataset all models performs similarly for both genders with minimal differences in MAE and RMSE scores. Overall, we observe that that age estimation is more accurate (or at least comparable) for male subjects than for female ones, which may be related in part to the use of makeup, which affects female facial appearance and consequently age estimation. Given the fact that UTKFace is approximately gender balanced (with a ratio of $10 : 9$ in favor of males), while IMDB-WIKI is not (a ratio of $14 : 10$ in favor of males) the difference in the performance cannot be ascribed to the training data. Instead, it appears that the model architecture and training procedure (observe results for FaceNet) as well as the characteristics of the test images have a much larger impact on age estimation results in our experiments.

**Race group comparison:** When looking at the MAE and RMSE scores for different race groups in Table I, we observe clear differences in performances of individual groups. The results are similar for all three considered models, but vary greatly among the two test datasets. The main reason for this is the inconsistent race partitioning between datasets, where the race labels of the two datasets may not necessarily correspond to subjects from the same races. For example, APPA-REAL does not have an Indian label, which suggest that Indians are likely part of the White (W) label. We therefore discuss results separately for each of the two test datasets.

On the APPA-REAL dataset the performance is consistent for all three models. The estimation errors are comparable for the White (W) and Asian (A) groups followed by the Black (B) demographic group, where we observe between $0.4$ and $1.2$ years larger MAE scores compared to the best performing race group. Interestingly, on the UTKFace dataset, we observe a very different setting. Here, the Other (O) and Black (B) race groups result in the lowest age estimation errors, followed by the Indian (I) race group. Here, the largest errors are observed for the White (W) and Asian (A) groups. This observation is particularly interesting and points to the fact that other data characteristics have a much greater impact on age estimation performance than race. Our experiments did not identify a consistent trend with respect to race-related performance variations, but point to the need for establishing consistent quality criteria (e.g., with respect to pose, illumination, image quality, etc.) across different demographic groups to be able to compare age estimation performance across race groups irrespective of other image-quality factors.

**Gender-race sub-group comparison:** In Table II we report MAE scores for sub-groups of subjects partitioned with respect to race and gender. We do not report RMSE results for this experiments to keep the table uncluttered. While most results are consistent with our previous findings, we notice some exceptions. When comparing the gender-divided Asian (A) group results, we observe that male subjects performs a lot worse than female subjects on the UTKFace dataset for both tested models, which affects the overall Asian (A) group results discussed in the previous section. Other than this, male subjects produce better results than female subjects in all race categories (the Male Indian (MI) sub-group from UTKFace dataset being the only additional exception with minor differences). Interestingly, the MA sub-groups also appears to have been the deciding group in the gender-oriented experiments that balanced the performances of the FaceNet model on the UTKFace dataset, since we see that FaceNet performs better for males than for females on all other sub-groups. When examining results on the APPA-REAL dataset we observe a comparable results between genders across all races with the biggest gap between male and female subject being observed for the category of White (W) subjects.

## V. Conclusion

In this study, we systematically analysed the performance of two off-the-shelf deep age estimation models based on face images from two publicly available datasets, UTKFace and APPA-REAL. By performing age estimation on demographic sub-categories of interest, we took a deeper look into race and gender bias. Current datasets used when training and testing age estimation models do not represent all races and both genders equally. We observed a tendency in the tested models to perform better with male subjects than with female ones, but did not identify a clear and consistent bias towards any particular race. Test dataset characteristics (especially for uncontrolled face images), such as image-quality, pose, illumination, occlusion and the like appear to have a bigger impact on age estimation performance than race. Nevertheless, additional research is needed to better understand the factors affecting age estimation performance. A particular problem here seems to be the lack of consistent quality boundaries across different demographic groups that would allow to evaluate the performance of current models on equal footing.

## Acknowledgments

This research was supported by the ARRS Research Programs P2-0250 (B) “Metrology and Biometric Systems”.

## References

[1] S. H. Abdurrahim, S. A. Samad, and A. B. Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *The Visual Computer*, 34(11):1617–1630, 2018.

[2] E. Agustsson, R. Timofte, S. Escalera, X. Baro, I. Guyon, and R. Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *FG*, 2017.

[3] J. Alarifi, J. Fry, D. Dancey, and M. H. Yap. Understanding face age estimation: humans and machine. In *CITS*, pages 1–5, 2019.

[4] M. Alvi, A. Zisserman, and C. Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV*, 2018.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[6] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR*.

[7] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pat. Rec.*, 46(3):628 – 641, 2013.

[8] A. Clapés, O. Bilici, D. Temirova, E. Avots, G. Anbarjafari, and S. Escalera. From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *CVPR-W*, 2018.

[9] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 2020.

[10] Y. Fu, G. Guo, and T. S. Huang. Age synthesis and estimation via faces: A survey. *IEEE TPAMI*, 32(11):1955–1976, 2010.

[11] X. Geng, Z. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE TPAMI*, 29(12):2234–2240, 2007.

[12] K. Grm, W. Scheirer, and V. Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE TIP*, 29(1):2150–2165, 2020.

[13] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *CVPR-W*, pages 71–78, 2010.

[14] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *CVPR*, pages 112–119, 2009.

[15] W. H. Ho, P. Watters, and D. Verity. Are younger people more difficult to identify or just a peer-to-peer effect. In *CAIP*, pages 351–359, 2007.

[16] J. C. J. Junior, C. Ozcinar, M. Marjanovic, X. Baró, G. Anbarjafari, and S. Escalera. On the effect of age perception biases for real age regression. In *FG*, pages 1–8, 2019.

[17] Kwon, Y. H, and N. da Vitoria Lobo. Age classification from facial images. *CVIU*, 74(1):1–21, 1999.

[18] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE TSMC-B*, 34(1):621–628, 2004.

[19] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *CVPR-W*, pages 34–42, 2015.

[20] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *BTAS*, pages 1–8, 2009.

[21] M. Ngan, P. J. Grother, and M. Ngan. *Face recognition vendor test (FRVT) performance of automated gender classification algorithms*. US Department of Commerce, National Institute of Standards and Technology, 2015.

[22] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016.

[23] R. Rothe, R. Timofte, and L. V. Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV-W*, 2015.

[24] R. Rothe, R. Timofte, and L. V. Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 2016.

[25] R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCV-W*, pages 10–15, 2015.

[26] M. B. F. Saavedra and R. S. Reíllo. Evaluation methodologies for security testing of biometric systems beyond technological evaluation. *Universidad Carlos III de Madrid*, pages 43–57, 2013.

[27] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[28] J. Xing, K. Li, W. Hu, C. Yuan, and H. Ling. Diagnosing deep learning models for high accuracy age estimation from a single image. *Pat. Rec.*, 66:106–116, 2017.

[29] C. Yan, C. Lang, T. Wang, X. Du, and C. Zhang. Age estimation based on convolutional neural network. In *PR-CAMIP*, pages 211–220, 2014.

[30] S. Yan, H. Wang, X. Tang, and T. S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *ICCV*, pages 1–8, 2007.

[31] Z. Yang and H. Ai. Demographic classification with local binary patterns. In S.-W. Lee and S. Z. Li, editors, *ICB*, pages 464–473, 2007.

[32] S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[33] S. Y. Zhang, Zhifei and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *CVPR*, 2017.