Unraveling the Age Estimation Puzzle:
Comparative Analysis of Deep Learning Approaches for Facial Age Estimation

Jakub Paplhám
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
paplhjak@fel.cvut.cz

Vojtěch Franc
Department of Cybernetics
Faculty of Electrical Engineering
Czech Technical University in Prague
xfrancv@cmp.felk.cvut.cz

Abstract

Comparing different age estimation methods poses a challenge due to the unreliability of published results, stemming from inconsistencies in the benchmarking process. Previous studies have reported continuous performance improvements over the past decade using specialized methods; however, our findings challenge these claims. We argue that, for age estimation tasks outside of the low-data regime, designing specialized methods is unnecessary, and the standard approach of utilizing cross-entropy loss is sufficient. This paper aims to address the benchmark shortcomings by evaluating state-of-the-art age estimation methods in a unified and comparable setting. We systematically analyze the impact of various factors, including facial alignment, facial coverage, image resolution, image representation, model architecture, and the amount of data on age estimation results. Surprisingly, these factors often exert a more significant influence than the choice of the age estimation method itself. We assess the generalization capability of each method by evaluating the cross-dataset performance for publicly available age estimation datasets. The results emphasize the importance of using consistent data preprocessing practices and establishing standardized benchmarks to ensure reliable and meaningful comparisons. The source code is available at Facial-Age-Benchmark.

1. Introduction

Over the past decade, age estimation has gained considerable interest in biometrics, law enforcement, and child protection. Numerous papers employ deep learning methods for age estimation from facial images, utilizing diverse approaches. For both researchers and practitioners intending to utilize age estimation, it is crucial to know: (i) what is the current state-of-the-art method, (ii) what performance can be expected from the method, and (iii) in what ways it can be improved. However, due to differences in the data-splitting of datasets and the lack of detailed descriptions of all components of the prediction system, answering these questions using the published results is not possible.

This paper aims at answering these questions. To this end, we evaluate and compare recent age estimation approaches in a unified setting and analyze the impact of factors, such as facial alignment, facial coverage, image resolution, model architecture, or pre-training data, on the result. In our paper, we demonstrate that these factors frequently exert a more pronounced impact than the choice of the age estimation method itself.

Contributions

- We benchmark and fairly compare recent deep-learning methods for age estimation from facial images. We concentrate on state-of-the-art methods that adapt a generic architecture by changing its last layer or the loss function to suit the age estimation task. Besides the usual intra-class performance, we also evaluate their cross-dataset generalization.
- We analyze the influence of different components of the facial recognition pipeline. Namely, the influence of (i) facial alignment and preprocessing, (ii) model architecture, (iii) loss function, and (iv) pre-training and highlight the need to keep them fixed to ensure a fair comparison. Surprisingly, we find that the influence of the loss function and the last layer, usually the main component that distinguishes different methods, is negligible, see Fig. 1.
- We make the entirety of our code public, including the data preparation pipeline and implementation of the methods. The published code is valuable to both practitioners who are interested in a state-of-the-art implementation and researchers who are looking for a reliable benchmark to evaluate their novel methods.

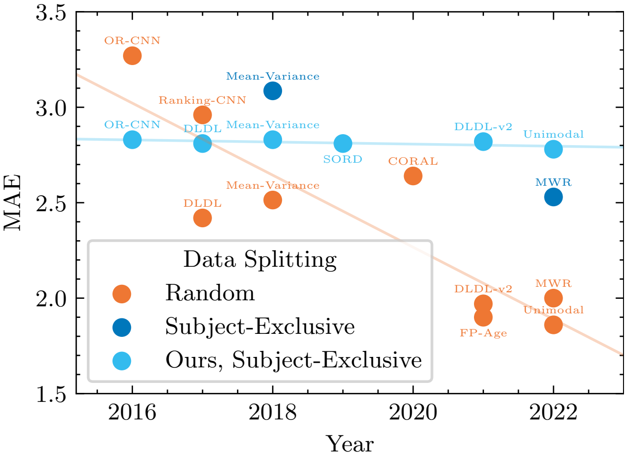


Figure 1. Mean Absolute Error (MAE) ↓ of age estimation methods on the MORPH dataset [23], as reported in existing literature. Random splitting remains the prevalent data splitting strategy and shows consistent performance improvements over time. Subject-exclusive data splitting, which offers more precise evaluations, is rarely employed. With unified subject-exclusive data splitting, all evaluated methods yield comparable results, failing to achieve the performance gains promised by the random splitting.

2. Related Work

This section explores the different age estimation techniques, that we compare in this paper, and examines the benchmarks currently utilized for evaluating them.

2.1. Age Estimation Methods

This paper compares various recent age estimation methods utilizing feedforward neural networks which receive an image $x \in \mathcal{X}$ and output an age estimate $y \in \mathcal{Y}$. We focus solely on age estimation methods that modify the standard classification approach by changing the last few layers of the neural network or the loss function. Although this may appear restrictive, it is essential to note that a majority of the methods proposed in the field fall into this category. By comparing methods that modify only a small part of the network, we aim to ensure a fair evaluation, as the remaining setup can be kept identical. Some recent methods, such as Moving Window Regression proposed by Shin *et al.* [26], were therefore omitted from this study.

Traditionally, age estimation relied on classification and regression-based approaches. However, these methods often overlook the inherent ordinal nature of age. In multi-class classification, misclassifications are treated equally, even though some age predictions may be more accurate than others. On the other hand, regression approaches can predict nonsensical and even negative age values. Ordinal regression has therefore emerged as a well-motivated approach to address these limitations. Unlike classification, where the labels merely represent categories, ordinal regression utilizes labels that provide sufficient information to order the objects. Below, we provide a concise overview of recent age estimation and ordinal regression approaches.

Classification The conventional classification approach still remains popular in the literature. For instance, Rothe *et al.* [25] achieved victory in the ChaLearn LAP 2015 challenge on apparent age estimation [10] with a model that employed cross-entropy to learn the posterior age distribution.

Extended binary classification Niu *et al.* [21] (OR-CNN) follow the approach proposed by Li and Lin [16] and transform the ordinal regression task into multiple binary classification sub-problems. For each age value $y_k \in \mathcal{Y}$, they construct a binary classifier to predict whether the true age $y \in \mathcal{Y}$ of a sample $x \in \mathcal{X}$ is larger than y_k. Cao *et al.* [3] (CORAL) modify this approach by restricting the hypothesis class such that the binary classifier predictions are consistent, i.e., the predicted probabilities satisfy $p(y > y_k|x) \geq p(y > y_{k+1}|x);\ \forall k$.

Fixed distribution learning Gao *et al.* [12] (DLDL) approach the task as multi-class classification. However, they encode the label distribution as a normal distribution centered at the true label. Díaz and Marathe [9] (SORD) approach the task similarly, but encode the label distribution as a double exponential distribution centered at the true label. In a follow up to their work [12], Gao *et al.* [13] (DLDL-v2) propose to also minimize the difference between (i) the true label $y \in \mathcal{Y}$, and (ii) the expectation $\mathbb{E}_{\hat{y} \sim f(x)}[\hat{y}]$ of the model output distribution $f(x)$.

Adaptive distribution learning An approach emerging in recent years is not to model a specific distribution, such as normal or double exponential distribution, but instead, to constrain the model by some statistical measure or a condition. Pan *et al.* [22] (Mean-Variance) approach the task as standard multi-class classification, but design a loss function that (i) minimizes the squared difference between the expectation $\mathbb{E}_{\hat{y} \sim f(x)}[\hat{y}]$ and the true label $y \in \mathcal{Y}$, and (ii) minimizes the variance $\mathbb{E}_{\bar{y} \sim f(x)}\left[(\bar{y} - \mathbb{E}_{\hat{y} \sim f(x)}[\hat{y}])^2\right]$ of the model output distribution $f(x)$. Similarly, Li *et al.* [17] (Unimodal) design a loss function (i) which constrains the model to output unimodal distributions, and (ii) concentrates the output distribution around the true label $y \in \mathcal{Y}$.

2.2. Age Estimation Benchmarks

The aforementioned publications evaluate their methods on a number of datasets, with the MORPH [23] dataset being the most commonly used. However, the evaluation procedures are not unified. For instance, OR-CNN [21] *randomly* divides the dataset into two parts: 80% for training and 20% for testing. No mention is made of a validation set for model selection. Identical *splitting protocols* are used in [3, 5, 12, 13, 19, 27], but the specific *data splits* differ between studies. Since the dataset contains multiple images per person; with many of them captured at the same age; there is a possibility for the same individual to be present in both the training and testing sets. This overlap introduces a potential bias, resulting in overly optimistic evaluation outcomes. The degree of data leakage can vary when using random splitting, resulting in certain data splits being more challenging than others. Consequently, comparing different methods and discerning which method truly stands out as the most effective becomes problematic.

Some publications [22,26] recognize this bias introduced by the splitting strategy and address it by implementing *subject-exclusive* splitting. This approach ensures that individuals are not included in both the training and testing sets. Standardized data splits are provided for two public age estimation datasets: the ChaLearn Looking at People Challenge 2016 (CLAP2016) dataset [1] and the Cross-Age Celebrity Dataset (CACD2000) [4]. However, CLAP2016 is relatively small, consisting of fewer than 8000 images. On the other hand, CACD2000 has training annotations that may contain noise, and its authors do not recommend its use for evaluating age estimation. Consequently, comparing methods using only these datasets is not satisfactory. Other popular datasets, AgeDB dataset [20], Asian Face Age Dataset (AFAD) [21], and FG-NET [15] also consist of multiple images per person, requiring subject-exclusive splitting. However, they lack any standardized data splits and as such suffer the same issues as MORPH [23].

Motivated by these findings, we aim to fairly compare the methods listed in Section 2.1. To achieve this, we evaluate their performance on standardized data splits, ensuring a unified setup. In an effort to facilitate future comparisons, we make our data splits publicly available.

3. Experimental Setup

For a fair comparison of multiple methods, the same experimental setup should be used for each of them. To accurately recreate an experiment, it is essential to replicate four key components: (i) preprocessing, (ii) model architecture, (iii) the decision layer and the loss function, and (iv) the data. To compare different age estimation approaches, we keep components (i), (ii), and (iv) constant while varying the component (iii). This allows us to isolate the impact of the selected method on performance. We compare the methods based on the achieved Mean Absolute Error (MAE).

In addition to comparing age estimation methods, we aim to assess the impact of the individual components. To accomplish this, we establish a baseline approach and systematically modify each component to gauge their respective influence on the results. In this section, we provide an overview of the components and describe our baseline approach. Where relevant, we also review how they are typically addressed in the literature.

3.1. Preprocessing

Unlike other deep learning tasks, where the entire image is often used as input, age estimation models should only be provided with a specific region of the image that corresponds to the face. A typical data preprocessing pipeline for facial age estimation consists of multiple steps. The initial step is face detection. Facial alignment is then performed to align facial landmarks, such as the eyes or corners of the mouth, rectifying variations in head pose, rotation, and scale. Finally, the region of interest is cropped and interpolated to a standardized size. We describe how we handle these steps in more detail below.

Facial Detection & Alignment For face detection and facial landmark detection we use the RetinaFace model developed by Deng *et al.* [8]. Our facial alignment procedure involves the following steps. Firstly, we determine a point that serves as the center of the final bounding box. This point is positioned between the center of the eyes and the center of the mouth corners. Secondly, the size of the bounding box is adjusted based on the distance between the eyes for frontal images and the distance between the eyes and mouth for profile images. Lastly, the bounding box is rotated to ensure horizontal alignment of the eyes. To provide a concise explanation, we refrain from presenting an extensive description of the alignment procedure. We encourage interested readers to refer to the accompanying implementation for more details.

While facial alignment defines the positioning, orientation, and scale of facial landmarks, the extent to which the face is visible in an image also needs to be specified. We refer to this notion as *facial coverage*. It measures how much of the face is shown in an image and can range from minimal coverage, where only the eyes and mouth are visible, to complete coverage, where the entire head is visible. Determining the optimal compromise between complete facial coverage and minimal coverage is not immediately clear. Complete facial coverage provides a comprehensive view of the face, allowing age estimation algorithms to consider a broader range of facial cues. On the other hand, partial coverage may help reduce overfitting by eliminating irrelevant facial cues and features with high variance. For a visual demonstration of various facial coverage levels, refer to Fig. 3. Surprisingly, the concept of facial coverage has received limited attention in age estimation literature. Consequently, the extent of facial coverage utilized in previous studies can only be inferred from the images presented in those works. For instance, Berg et al. [2] seemingly employ minimal coverage, showing slightly more than just the mouth and eyes. The majority of other works [3, 12, 13, 17, 21, 26] tend to adopt partial coverage, where a significant portion of the face, including the chin and forehead, is visible, but not the entire head. In the works of Pan *et al*. [22], Rothe *et al*. [25], and Zhang *et al*. [27], the entire head is shown. In our baseline setup, we have opted for complete facial coverage, where the images encompass the entire head. We further study the effects of facial coverage on model performance in Sec. 4.4.

Image Resolution After applying the facial detection and alignment pipeline, a suitable square bounding box is chosen to match the desired facial coverage. Subsequently, we extract the bounding box and resize the resulting image to the desired size. In our baseline setup, we use a model input size of 256×256 pixels. However, in practice, we extract a slightly larger bounding box (enlarged by an additional 10%) to facilitate data augmentation without introducing padded pixels. We further study the effects of input resolution on model performance in Sec. 4.4.

Representation Lastly, we normalize the pixel values of the images. To this end, we subtract the mean and divide by the standard deviation of the color channels computed over ImageNet [7]. Recently, Lin *et al*. [18] proposed to also transform the image to a different coordinate system and the approach has demonstrated impressive results in the semantic segmentation of faces. We do *not* utilize any such transformation in our baseline setup, however, we investigate the potential benefits of this transformation in Sec. 4.4.

3.2. Model Architecture

Multiple different backbone architectures can be found in the age estimation literature. Among these architectures, VGG16 [9, 13, 17, 22, 26, 27] and ResNet-50 [2, 3, 19] stand out as the most common choice. We have chosen to use the ResNet-50 [14] as baseline backbone architecture and experiment with different architectures in Sec. 4.3.

3.3. Method Details

Note that for all methods which model the posterior distribution $p(y|x)$, namely (i) cross-entropy, (ii) DLDL [12], (iii) DLDL-v2 [13], (iv) SORD [9], (v) Mean-Variance loss [22], and (vi) Unimodal loss [17], we use optimal plugin Bayes predictor for MAE loss, i.e., we predict $\arg\min_y \mathbb{E}_{\hat{y}\sim f(x)}[|y-\hat{y}|]$. For regression, we use the absolute error as the loss function.

Training Details We utilize the Adam optimizer with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. For pre-training on the IMDB-WIKI dataset, we set the learning rate to $\alpha = 10^{-3}$ and train the model for a total of 100 epochs. For fine-tuning on the remaining datasets we reduce the learning rate to $\alpha = 10^{-4}$ and train the model for 50 epochs. We use a batch size of 100. The best model is selected based on the MAE metric computed on the validation set. We utilize two data augmentations during training, (i) horizontal mirroring, and (ii) cropping out an 80% to 100% portion of the bounding box and resizing it to the model input shape.

3.4. Data

The main challenge in comparing different age estimation methods lies in the fact that published results often utilize different data splits. Even when using a consistent data-splitting protocol, such as random or subject-exclusive, the results can differ significantly (i.e., the difference is on the same level as the impact of a different loss function) if the specific data splits are not identical. To address this, we evaluate all methods on identical data splits. In this section, we discuss the used datasets and the data-splitting strategy.

Datasets We evaluate the methods using a total of 7 datasets: AgeDB [20], AFAD [21], CACD2000 [4], CLAP2016 [1], FG-NET [15], MORPH [23], and UTK-Face [28]. We also utilize the IMDB-WIKI dataset [24] for pre-training. However, it is important to note that the labels in the IMDB-WIKI dataset are known to be noisy. To mitigate this problem, Lin *et al*. [19], and Franc and Čech [11] attempted to clean the labels. To assess the quality of these labels, we trained ResNet-50 models on the dataset using the labels and evaluated the models' performance on the other datasets. The results are presented in Tab. 1. Both models achieved similar results, so the choice of labels between [11, 19] is arbitrary. Due to a slightly lower overall Mean Absolute Error (MAE), we have decided to use the labels from Franc and Čech [11] for pre-training in the rest of this paper. We further analyze the importance of pre-training in Sec. 4.2.

Data Splits For the CLAP2016 and CACD2000 datasets, we use the single data split provided by the dataset authors. However, for the remaining datasets, we create 5 subject-exclusive data splits, ensuring that each split has the same age distribution. In this setup, 60% of the dataset is used for training, 20% for model selection, and 20% for evaluating the model performance. Due to its small size, we only use FG-NET for evaluation.

4. Experiments

We conducted experiments to compare various deep-learning methods for age estimation. We also evaluated how different factors, such as data preparation, architecture, and pre-training, affect the performance of the models. In Sec. 4.1 we provide a comparative analysis of methods that were pre-trained on the IMDB-WIKI dataset and then fine-tuned on downstream datasets. The importance of the pre-training is assessed in Sec. 4.2. We also investigate how the choice of the backbone architecture impacts model performance in Sec. 4.3. Lastly, we evaluate different components of the data preparation pipeline in Sec. 4.4. Specifically, we examine the influence of the facial alignment, level of facial coverage, image resolution, and image representation.

4.1. Method Comparison

To fairly compare the performance of the methods, we start their training from the same initialization, specifically from weights pre-trained on ImageNet and then further pre-trained on IMDB-WIKI with cross-entropy. After the pre-training, the last layer of the model is replaced with a layer specific to the desired method. The models are then fine-tuned on the downstream dataset. It is important to note that for the baseline cross-entropy, we also replace the final layer prior to fine-tuning. This ensures that the experimental setup remains identical to that of the other methods. The performance of the models, evaluated using the Mean Absolute Error (MAE), is presented in Tab. 8.

Intra-Dataset Performance The intra-dataset results achieved with the IMDB-WIKI pre-training are highlighted with a grey background. To determine whether any method is consistently better than others, we employ the Friedman test and the Nemenyi critical difference test as described by Demšar [6]. Using a significance level (p-value) of $\alpha = 5\%$, we can conclude that there is *no* significant difference in performance between any of the methods [9, 12, 13, 17, 21, 22] and the cross-entropy method. In other words, we do not observe any systematic improvement by using any of the methods.

Cross-Dataset Generalization Cross-dataset results, Tab. 8, were obtained by assessing the performance of models on datasets that were not part of their training. For all of the methods, the cross-dataset performance is consistently and significantly worse than the intra-dataset performance. It is important to note that for computing the cross-dataset performance, the entire evaluation dataset is utilized as the test set. Using the Friedman test and the Nemenyi critical difference test with a significance level (p-value) of $\alpha = 5\%$, we can conclude that there is *no* significant difference in generalization capability between any of the methods [9, 12, 13, 17, 21, 22] and the cross-entropy.

4.2. Importance of Pre-training

We also investigate the impact of pre-training data on the results. We compare the following three scenarios: (i) models initialized with random weights, (ii) models initialized with weights pre-trained on ImageNet, and (iii) models initialized with weights pre-trained on ImageNet and further pre-trained on IMDB-WIKI. The results are presented in Tab. 8, where the initialization scenarios are labeled as follows: (i) Rand., (ii) Imag., and (iii) IMDB.

Random Initialization When starting from random initialization, we noticed that training with the Unimodal loss [17] tends to be unstable. To draw conclusions, we again use the Friedman test and the Nemenyi test with a significance level (p-value)of $\alpha = 5\%$, excluding the Unimodal loss [17]. The results indicate that with the random initialization, OR-CNN [21], DLDL [12], and the Mean-Variance loss [22] demonstrate a significant performance improvement over the cross-entropy.

In scenarios with limited data availability, where pre-training is not possible, it is, thus, advisable to utilize one of the aforementioned methods.

Pre-training Data Our findings reveal that the IMDB-WIKI pre-training method consistently outperforms other approaches in terms of performance. Nonetheless, in certain isolated cases, ImageNet initialization has shown superior results. To draw definitive conclusions, we use the Friedman test and the Nemenyi test, with a significance level (p-value) of $\alpha = 5\%$. Our analysis indicates that pre-training with IMDB-WIKI consistently outperforms pre-training solely on ImageNet, and both methods outperform random initialization.

Pre-training Loss Function It can be argued that the utilization of cross-entropy for the IMDB-WIKI pre-training puts the other methods at a disadvantage, as the extracted features become more suitable for cross-entropy rather than for the alternative methods. In order to investigate this possibility, we have chosen the two most recent methods, namely the Mean-Variance loss proposed by Pan *et al.* [22] and the Unimodal loss presented by Li *et al.* [17]. We pre-train the models on IMDB-WIKI using the corresponding loss functions and subsequently finetune them on the downstream datasets. The obtained results are displayed in Tab. 6. We do not find any benefit when using the same method for both pre-training and finetuning. On the contrary, we observe that the Unimodal loss [17] yields better results across all datasets when the model is pre-trained using cross-entropy. We, therefore, conclude that the pre-training with cross-entropy does not invalidate our results.

4.3. Model Architecture

In the age estimation literature, numerous backbone architectures can be found. We aim to evaluate the influence of the architecture choice on the obtained results and present our findings in Tab. 7. We observe that the selection of the model has a more pronounced impact on the performance than the choice of the age estimation method itself.

4.4. Data Preparation Pipeline

Age estimation models require only a specific region of an image, specifically the person's face, as input, rather than the entire image. However, the influence of this selection process on the model's performance is not apriori known. Additionally, facial images used for age estimation can differ in terms of scale and resolution since they originate from various sources and as such need to be resized to a uniform resolution. In this section, we examine the impact of the aforementioned data preparation pipeline on the performance of age estimation models.

Facial Alignment Numerous studies lack an explanation of their facial alignment procedure. Others merely mention the utilization of facial landmarks. To assess whether a standardized alignment is needed for a fair comparison of multiple methods, we adopt three distinct alignment procedures and evaluate their effect on model performance. Firstly, we (i) perform no alignment and employ the bounding box proposed by the facial detection model [8] as the simplest approach. Secondly, (ii) we utilize the proposed bounding box but rotate it to horizontally align the eyes. Lastly, (iii) we use our baseline alignment procedure, which normalizes the rotation, positioning, and scale, and is described in Sec. 3. A visual representation of these facial alignment methods is depicted in Fig. 2. The performance of models trained using the various alignment procedures is presented in Tab. 2. When working with pre-aligned datasets like AFAD, we observe that procedure (iii) does not yield significant improvements compared to the simpler variants (i) or (ii). Similar results are obtained on datasets collected under standardized conditions, such as the MORPH dataset. However, when dealing with in-the-wild datasets like AgeDB and CLAP2016, we find that alignment procedure (iii) leads to noticeable improvements over the simpler methods. Interestingly, on the UTKFace dataset, which also contains in-the-wild images, approach (ii) of solely rotating the proposed bounding boxes achieves superior outcomes compared to approach (iii). In summary, the disparities among the various alignment procedures are not substantial. Hence, it can be argued that any facial alignment technique that effectively normalizes the position, rotation, and scale of the faces would yield comparable results.

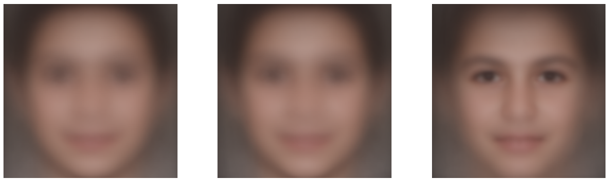
Facial Coverage We also investigate the impact of facial coverage on the performance of age estimation models. For a visual representation of different levels of facial coverage, refer to Fig. 3. The performance of models trained with the different coverage levels is presented in Tab. 3. Generally, complete facial coverage, which includes the entire head in the model input, yields the best results across the majority of datasets. However, in certain cases such as the AFAD dataset or the MORPH dataset, partial coverage performs better. It is important to note that the AFAD dataset contains preprocessed images that do not capture the entire head. Consequently, using complete facial coverage with this dataset results in the presence of black bars and a decrease in the effective pixel resolution of the face. Therefore, it is expected that increased facial coverage would yield inferior results compared to partial coverage on this dataset. The smallest coverage, limited to the facial region up to the eyes and mouth, consistently performs the worst. It can be concluded that with sufficient pixel resolution, the full facial coverage is superior.

Input Resolution To investigate the influence of input resolution on age estimation, we performed experiments using multiple resolutions on all datasets: specifically, 256×256, 128×128, and 64×64 pixels. The performance of the models trained with different input resolutions is presented in Tab. 4. Our findings indicate that an increase in image resolution consistently results in improved model performance across all datasets. Hence, the best performance was achieved with a resolution of 256×256 pixels. However, it remains uncertain whether further increases in image resolution would yield even more impressive results.

Input Transform We also examined the input transformation proposed by Lin *et al.* [18], which involves converting a face image into a tanh-polar representation. This approach has shown great performance in face semantic segmentation. Lin *et al.* then modified the network for age estimation, reporting impressive results [19]. We explored the potential benefits of applying this transformation for age estimation. However, our findings indicate that the transformation does not improve the results compared to the baseline, as shown in Tab. 5. Therefore, we conclude that the improved age estimation performance observed by Lin *et al.* [19] does not arise from the use of a different representation, but rather from pre-training on semantic segmentation or their specialized model architecture.

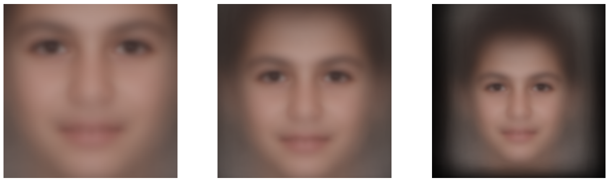
5. Discussion and Conclusions

In this paper, we aimed to establish a fair comparison framework for evaluating various approaches for age estimation. We conducted a comprehensive analysis on seven different datasets, namely AgeDB [20], AFAD [21], CACD2000 [4], CLAP2016 [1], FG-NET [15], MORPH [23], and UTKFace [28], comparing the models based on their Mean Absolute Error (MAE). To determine



(a) Crop. (b) Rotation. (c) Rot., Trans., Scale.

Figure 2. Comparison of different alignment methods using the average face from the FG-NET dataset.



(a) Eyes & Mouth. (b) Chin & Forehead. (c) Head.

Figure 3. Comparison of different facial coverage levels using the average face from the FG-NET dataset.

Evaluation Dataset	Annotations	
	EM-CNN [11]	FP-AGE [19]
AgeDB	6.44	**6.30**
AFAD	**6.86**	7.23
CACD2000	**5.81**	5.90
CLAP2016	6.24	**5.53**
FG-NET	10.32	**6.09**
MORPH	**4.94**	5.30
UTKFace	8.31	**6.26**
Overall	**6.28**	6.36
IMDB	4.90	5.15

Table 1. MAE ↓ of ResNet-50 trained on IMDB-WIKI with clean age labels from (i) EM-CNN [11], and (ii) FP-AGE [19]. Results on IMDB-WIKI are not included in the *Overall* result.

Dataset	Alignment		
	Crop	Rotation	Rot. + Trans. + Scale
AgeDB	5.93	5.92	**5.84**
AFAD	3.12	**3.11**	**3.11**
CACD2000	4.01	**4.00**	**4.00**
CLAP2016	4.68	4.57	**4.49**
MORPH	2.81	**2.78**	2.79
UTKFace	4.49	**4.42**	4.44

Table 2. MAE ↓ of ResNet-50 models with different facial alignment. The models were pre-trained on IMDB-WIKI.

Dataset	Facial Coverage		
	Eyes & Mouth	Chin & Forehead	Head
AgeDB	6.06	5.84	**5.81**
AFAD	3.17	**3.11**	3.14
CACD2000	4.02	4.00	**3.96**
CLAP2016	5.06	**4.49**	**4.49**
MORPH	2.88	**2.79**	2.81
UTKFace	4.63	4.44	**4.38**

Table 3. MAE ↓ of ResNet-50 models with different facial coverages. The models were pre-trained on IMDB-WIKI.

Dataset	Image Resolution		
	64 × 64	128 × 128	256 × 256
AgeDB	8.43	6.90	**5.81**
AFAD	3.36	3.25	**3.14**
CACD2000	5.01	4.55	**3.96**
CLAP2016	11.34	5.90	**4.49**
MORPH	3.33	3.07	**2.81**
UTKFace	5.83	4.81	**4.38**

Table 4. MAE ↓ of ResNet-50 models with different image resolutions. The models were pre-trained on IMDB-WIKI.

Dataset	Transform	
	No Transform	RoI Tanh-polar [18]
AgeDB	**5.81**	5.93
AFAD	**3.14**	3.15
CACD2000	**3.96**	4.07
CLAP2016	**4.49**	4.71
MORPH	2.81	**2.80**
UTKFace	**4.38**	4.39

Table 5. MAE ↓ of ResNet-50 models with different input transformations. The models were pre-trained on IMDB-WIKI.

Dataset	Transform		
	Cross-E.	Mean-Var. [22]	Unimod. [17]
AgeDB	5.81	**5.75**	6.17
AFAD	**3.14**	3.18	3.31
CACD2000	**3.96**	4.06	4.23
CLAP2016	4.49	**4.30**	5.25
MORPH	**2.81**	2.83	2.91
UTKFace	**4.38**	4.49	4.66

Table 6. Intra-dataset MAE ↓ of different methods. The models were pre-trained on IMDB-WIKI. Pre-training and finetuning were performed with the same method.

if any method outperformed the others, we employed the Friedman test and the Nemenyi critical difference test.

When incorporating pre-training on the IMDB-WIKI dataset, we observed no significant improvement in the age estimation results compared to the baseline cross-entropy method. It is widely recognized that as the amount of available data increases, the cross-entropy loss becomes asymptotically optimal for learning the true posterior. Hence, we contend that with the pre-training, we are approaching this asymptotic regime. Previously published results report continuous performance improvements over time (as depicted in Fig. 1), however, our findings challenge these claims. We argue that the reported improvements can be attributed to either the random data splitting strategy or hyperparameter tuning aimed at achieving the best test performance.

When employing random model initialization, we observed some improvement over the baseline cross-entropy on small datasets. Specifically, the Mean-Variance loss [22], OR-CNN [21], and DLDL [12] demonstrated significant improvements. These improvements can be attributed to the regularization provided by these methods.

Furthermore, our analysis of the data preparation pipeline revealed that factors such as the extent of facial coverage, input resolution, and the facial alignment procedure had a more significant impact on the achieved results compared to the choice of the age estimation method itself. Moreover, we suggest that for age estimation tasks outside of the low-data regime, designing specialized methods may not be necessary. The standard approach of utilizing cross-entropy loss is sufficient in such cases. Based on our results, we emphasize the necessity of employing standardized subject-exclusive data splitting when comparing different methods. To facilitate reproducibility and future comparisons, we have made the splits, as well as our implementation, publicly available.

Backbone	Dataset					
	AgeDB	AFAD	CACD2000	CLAP2016	MORPH	UTKFace
ResNet-50	5.81	**3.14**	3.96	4.49	**2.81**	4.38
EfficientNet-B4	**5.76**	3.20	4.00	**4.06**	2.87	**4.23**
ViT-B-16	9.07	4.04	6.22	8.55	4.35	6.88
VGG-16	6.02	3.22	**3.92**	4.65	2.88	4.64

Table 7. Intra-dataset MAE ↓ with different backbone architectures. The models were pre-trained on IMDB-WIKI [24].

		Evaluation Dataset																				
		AgeDB			AFAD			CACD2000			CLAP2016			FG-NET			MORPH			UTKFace		
	Method / Init.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.	IMDB	Imag.	Rand.
Training Dataset AgeDB	Cross-Entropy	5.81	7.20	7.65	7.83	12.61	14.70	5.90	8.10	8.73	6.83	10.86	12.41	10.82	16.27	18.87	4.83	6.74	6.93	8.45	11.88	11.82
	Regression	6.23	6.54	7.60	8.05	12.13	14.19	6.76	7.56	8.74	8.32	9.95	12.32	10.56	13.83	18.00	5.64	6.66	6.94	9.42	10.42	12.14
	OR-CNN [21]	5.78	6.51	7.52	7.47	11.86	13.80	6.05	7.71	8.55	6.64	10.18	11.74	9.74	13.82	17.40	4.73	6.61	7.05	8.20	10.75	11.41
	DLDL [12]	5.80	6.95	7.46	7.81	1.85	14.92	5.99	7.96	8.52	6.51	10.84	11.48	9.23	15.63	16.95	4.74	6.34	6.82	7.97	11.87	11.24
	DLDL-v2 [13]	5.80	6.87	7.58	7.61	12.98	16.19	5.91	8.15	8.61	6.48	10.88	12.50	9.91	15.36	19.01	4.92	7.53	7.20	7.97	11.61	12.30
	SORD [9]	5.81	6.93	7.58	7.81	12.80	15.42	5.96	7.90	8.77	6.61	10.37	12.22	9.72	14.76	17.85	4.76	6.58	7.14	8.12	11.53	11.60
	Mean-Var. [22]	5.85	6.69	7.33	7.26	12.40	14.35	6.00	7.89	8.35	6.70	10.50	11.90	10.55	14.32	17.43	4.99	6.87	7.33	8.25	10.77	11.64
	Unimodal [17]	5.90	7.11	15.49	8.37	13.11	20.87	6.22	8.24	16.11	6.73	11.14	21.31	10.15	16.13	32.77	4.84	6.78	17.42	8.23	11.86	23.09
AFAD	Cross-Entropy	15.70	17.31	18.05	3.14	3.17	3.32	9.54	11.18	11.21	8.96	10.23	10.32	10.92	11.38	11.96	6.80	6.83	8.19	12.10	13.07	13.29
	Regression	13.67	15.91	17.21	3.17	3.16	3.30	8.72	10.51	10.72	8.33	9.91	10.02	11.20	11.89	12.35	6.27	7.34	7.99	11.23	12.83	12.96
	OR-CNN [21]	12.08	15.65	16.72	3.16	3.17	3.28	8.87	11.05	10.89	7.85	9.73	9.92	10.63	11.94	12.58	6.68	6.85	7.81	10.50	12.43	12.74
	DLDL [12]	14.12	15.70	17.21	3.14	3.16	3.25	9.40	10.70	11.06	8.68	9.54	9.98	11.31	11.64	12.07	7.04	6.75	7.82	11.52	12.43	12.82
	DLDL-v2 [13]	13.90	16.33	17.78	3.15	3.17	3.28	9.46	10.68	11.02	8.60	9.76	10.32	10.83	11.81	12.64	6.92	6.79	7.94	11.29	12.61	13.18
	SORD [9]	14.30	16.08	17.49	3.14	3.15	3.24	9.45	10.70	11.09	8.64	9.79	10.10	11.21	11.63	12.19	6.87	6.82	7.93	11.59	12.79	13.10
	Mean-Var. [22]	12.54	15.07	16.68	3.16	3.16	3.26	8.98	10.33	10.75	7.93	9.33	9.78	10.96	12.24	12.43	6.61	6.76	7.88	10.57	12.00	12.62
	Unimodal [17]	13.99	15.89	20.97	3.20	3.24	9.30	9.23	10.68	14.56	8.64	9.79	14.51	11.31	11.83	18.29	7.07	7.32	12.53	11.26	12.33	17.47
CACD2000	Cross-Entropy	9.66	11.84	10.60	10.70	8.50	13.08	3.96	4.59	4.89	8.42	8.64	10.51	17.45	23.64	20.86	7.21	12.20	10.39	11.16	11.38	12.61
	Regression	10.91	10.44	10.76	10.23	7.23	11.66	4.06	4.52	4.83	8.84	7.75	9.98	17.55	19.50	19.60	8.61	8.81	11.79	11.34	10.38	11.78
	OR-CNN [21]	10.43	11.02	11.85	9.66	9.48	12.17	4.01	4.60	4.74	8.57	8.85	10.29	18.47	24.32	20.85	7.52	10.04	11.05	11.17	12.30	12.27
	DLDL [12]	9.84	10.79	11.28	10.09	9.30	13.20	3.96	4.42	4.76	8.39	8.49	9.99	18.38	18.99	21.52	7.27	9.16	11.01	11.19	11.94	12.27
	DLDL-v2 [13]	9.90	12.31	11.20	8.03	11.50	11.51	3.96	4.57	4.69	7.67	8.88	9.43	18.11	22.89	19.02	7.20	13.46	9.73	10.52	12.32	11.47
	SORD [9]	9.77	10.90	11.04	10.35	9.55	11.95	3.96	4.42	4.70	8.38	8.51	9.89	18.05	20.84	21.73	7.23	8.98	11.59	11.18	12.06	12.22
	Mean-Var. [22]	10.81	11.42	10.83	9.71	10.82	11.49	4.07	4.60	4.78	8.88	9.20	10.08	20.48	22.68	20.14	8.14	12.59	11.72	11.74	12.29	12.23
	Unimodal [17]	10.46	11.04	46.26	10.63	9.85	25.74	4.10	4.73	37.41	9.19	8.92	30.96	19.37	19.75	15.84	8.94	11.64	32.63	11.89	11.75	32.98
CLAP2016	Cross-Entropy	7.35	10.15	12.26	5.41	7.03	5.34	6.65	8.11	9.11	4.49	5.96	8.73	5.92	9.28	12.02	4.96	6.61	6.90	5.74	7.21	8.58
	Regression	7.51	8.52	11.74	6.07	5.19	5.95	6.86	7.24	9.45	4.65	4.77	7.89	4.85	6.31	10.14	5.09	5.49	8.83	6.02	5.93	8.66
	OR-CNN [21]	6.83	8.74	11.24	5.83	5.92	5.44	6.73	7.25	8.65	4.13	4.60	7.38	5.09	6.47	9.22	4.92	5.78	6.52	5.43	5.95	7.68
	DLDL [12]	7.20	9.33	11.39	5.57	6.90	5.85	6.85	7.64	9.26	4.18	5.10	7.39	5.26	7.44	9.18	4.89	5.92	6.52	5.51	6.37	7.87
	DLDL-v2 [13]	7.14	9.42	12.36	5.47	5.95	6.45	6.69	7.99	9.34	4.23	4.87	8.52	5.22	7.04	8.75	4.85	6.04	7.29	5.53	6.12	8.23
	SORD [9]	7.19	9.60	12.16	5.47	7.74	6.62	6.63	8.09	9.66	4.27	5.34	7.81	5.59	7.77	7.62	4.92	6.01	6.62	5.48	6.46	8.08
	Mean-Var. [22]	7.08	9.16	12.58	5.18	6.30	5.38	6.64	7.37	9.94	4.28	4.87	7.95	5.45	6.69	11.14	4.96	7.38	7.49	5.52	6.16	8.65
	Unimodal [17]	7.01	9.77	20.71	5.58	6.10	5.54	6.47	8.20	13.08	4.17	5.39	13.83	5.13	6.39	15.13	4.80	6.05	10.02	5.44	6.67	15.27
MORPH	Cross-Entropy	9.66	11.73	12.63	6.69	7.78	10.36	8.53	10.83	10.11	6.90	8.96	10.64	9.45	11.96	15.38	2.81	2.96	3.01	8.97	10.81	11.92
	Regression	10.48	12.99	12.56	6.60	6.65	10.66	9.82	11.47	9.68	7.83	9.27	10.67	9.24	10.13	16.69	2.83	2.74	2.97	9.40	10.97	12.06
	OR-CNN [21]	9.35	11.65	12.82	6.78	7.78	11.81	8.39	11.34	10.23	6.84	8.73	11.05	9.58	11.09	17.47	2.83	2.85	2.99	8.82	10.37	12.06
	DLDL [12]	9.41	12.00	12.66	6.58	7.78	11.76	8.58	11.92	10.10	6.85	9.26	11.15	9.44	11.43	16.94	2.81	2.92	2.98	8.80	10.81	12.46
	DLDL-v2 [13]	9.79	11.49	12.68	6.60	8.22	12.45	8.79	10.98	9.81	6.98	8.98	11.22	9.52	11.63	17.57	2.82	2.93	3.00	8.97	10.70	12.47
	SORD [9]	9.48	11.84	12.73	6.54	7.91	11.19	8.73	11.18	10.13	6.84	8.99	10.72	9.34	11.08	15.90	2.81	2.91	2.99	8.83	10.85	11.97
	Mean-Var. [22]	9.70	11.62	12.93	6.68	7.81	10.41	8.65	10.59	10.11	7.03	8.80	10.56	9.51	11.45	15.81	2.83	2.89	2.95	8.94	10.59	11.95
	Unimodal [17]	9.93	12.31	17.44	6.63	7.04	8.18	8.68	10.11	12.03	7.19	8.95	12.38	9.80	12.17	17.83	2.78	2.90	8.66	9.07	10.75	15.45
UTKFace	Cross-Entropy	6.61	8.88	9.58	5.51	6.42	6.75	6.56	9.10	8.98	4.82	7.34	7.50	4.78	6.62	7.64	5.09	6.61	7.35	4.38	4.75	5.32
	Regression	7.01	7.79	8.83	5.96	6.26	6.43	6.77	7.87	8.61	5.24	5.93	6.67	4.41	5.07	7.27	5.41	5.95	6.71	4.72	4.53	5.34
	OR-CNN [21]	6.71	8.29	8.75	5.56	6.74	6.52	6.61	8.89	8.37	4.95	6.79	6.70	4.54	5.71	6.55	5.26	6.07	6.76	4.40	4.43	5.15
	DLDL [12]	6.65	8.60	9.00	5.42	6.68	6.19	6.52	9.01	8.84	4.81	7.19	7.46	4.85	5.87	7.28	5.16	6.25	7.03	4.39	4.66	5.30
	DLDL-v2 [13]	6.79	8.43	8.91	5.45	6.65	5.82	6.61	9.39	8.50	4.98	7.27	6.85	4.79	5.93	7.44	5.46	6.24	6.79	4.42	4.60	5.19
	SORD [9]	6.61	8.96	9.11	5.42	7.18	6.32	6.52	9.42	8.69	4.82	7.87	7.18	4.83	6.15	7.55	5.14	6.36	7.28	4.36	4.68	5.25
	Mean-Var. [22]	6.79	8.36	8.53	5.41	6.54	6.32	6.55	8.55	8.32	5.04	6.81	6.32	5.05	6.30	6.90	5.37	6.15	6.39	4.42	4.57	5.05
	Unimodal [17]	6.68	8.66	22.42	5.35	7.68	16.64	6.58	9.28	17.17	4.86	7.60	18.83	4.55	6.25	22.98	5.22	5.96	16.44	4.47	4.78	21.01

Table 8. Intra-dataset and cross-dataset Mean Absolute Error (MAE) ↓ of ResNet-50 models. Results marked as *Initialization: IMDB* are of models that are initialized to ImageNet weights, then trained with Cross-Entropy on IMDB-WIKI [24] and then finetuned on the downstream dataset. *Imag.* signifies initialization to weights pre-trained on ImageNet. *Rand.* denotes random initialization.

References

[1] 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 *12th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2017*. IEEE, 2017. 3, 4, 6

[2] A. Berg, M. Oskarsson, and M. O'Connor. Deep ordinal regression with label diversity. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2740–2747, Los Alamitos, CA, USA, jan 2021. IEEE Computer Society. 3, 4

[3] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. 2, 3, 4

[4] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 768–783, Cham, 2014. Springer International Publishing. 3, 4, 6

[5] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 742–751, 2017. 3

[6] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006. 5

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[8] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 3, 6

[9] Raúl Díaz and Amit Marathe. Soft labels for ordinal regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4742, 2019. 2, 4, 5, 9

[10] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzàlez, Hugo Jair Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 243–251, 12 2015. 2

[11] Vojtech Franc and Jan Cech. Learning cnns from weakly annotated facial images. *Image and Vision Computing*, 2018. 4, 7

[12] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017. 2, 3, 4, 5, 8, 9

[13] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng. Age estimation using expectation of label distribution learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 712–718. International Joint Conferences on Artificial Intelligence Organization, 7 2018. 2, 3, 4, 5, 9

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[15] A. Lanitis, C.J. Taylor, and T.F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002. 3, 4, 6

[16] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems*, page 865 – 872, 2007. Cited by: 195. 2

[17] Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chunmao Wang, and Shiliang Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20513–20522, June 2022. 2, 4, 5, 7, 9

[18] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 112, 2021. 4, 6, 7

[19] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Transactions on Image Processing*, 2022. 3, 4, 6, 7

[20] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, 2017. 3, 4, 6

[21] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4920–4928, 2016. 2, 3, 4, 5, 6, 8, 9

[22] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018. 2, 3, 4, 5, 7, 8, 9

[23] 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, 2006. 2, 3, 4, 6

[24] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018. 4, 9

[25] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *2015*

IEEE International Conference on Computer Vision Workshop (ICCVW), pages 252–257, 2015. 2, 4

[26] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: A novel approach to ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18760–18769, June 2022. 2, 3, 4

[27] Yunxuan Zhang, Li Liu, Cheng Li, and Chen-Change Loy. Quantifying facial age by posterior of age comparisons. In Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 108.1–108.12. BMVA Press, September 2017. 3, 4

[28] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 4, 6