IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 December 2022, accepted 28 January 2023, date of publication 10 February 2023, date of current version 15 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3244059

RESEARCH ARTICLE

Facial Age Estimation Models for Embedded Systems: A Comparative Study

ZORANA DOŽDOR, TOMISLAV HRKAĆ, (Member, IEEE), KARLA BRKIĆ, AND ZORAN KALAFATIĆ, (Member, IEEE)
Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
Corresponding author: Zoran Kalafatić (zoran.kalafatic@fer.hr)

This work was supported by the "Development of an Advanced Electric Bicycles Charging Station for a Smart City" Project through the European Structural and Investment Funds under the Operational Program under Grant KK.01.1.1.07.0066.

ABSTRACT Automated age estimation from face images is the process of assigning either an exact age or a specific age range to a facial image. In this paper a comparative study of the current techniques suitable for this task is performed, with an emphasis on lightweight models suitable for embedded implementation. We investigate both the suitable modern deep learning architectures for feature extraction and the variants of framing the problem itself as either classification, regression or soft label classification. The models are evaluated on Audience dataset for age group classification and FG-NET dataset for exact age estimation. To gather in-depth insights into automated age estimation and in contrast to existing studies, we additionally compare the performance of both classification and regression on the same dataset. We propose a novel loss function that combines regression and classification approaches and show that it outperforms other considered approaches. At the same time, with a lightweight backbone, such an architecture is suitable for implementation on embedded devices.

INDEX TERMS Age estimation, computer vision, deep learning, face detection.

I. INTRODUCTION

Automated age estimation (AAE) from face images can be defined as the process of assigning either an exact age or a specific age range to a facial image. AAE has a wide scope of applications in human-computer interaction, security systems, biometric systems, advertising industry etc. Therefore, age estimation has become a topic of interest for both industry and academic community. In spite of a large body of work dealing with facial age estimation, it is still a challenging problem, as the aging process significantly differs from one person to another. This is caused by internal factors such as genes, changes in the shape and the size of the face, but also by external factors like lifestyle and living conditions of an individual [1]. It has been shown [2] that in some cases it is very difficult to accurately infer the age of a person visually even for a human. While automated age estimation methods that approach or even surpass human performance have been proposed, there is still significant room for improvements, especially in unconstrained conditions [3], [4]. Several examples that have proven difficult to correctly classify in this study are shown in Fig. 1, along with the closest model predictions and ground-truth labels.

A typical pipeline of a state-of-the-art age estimation system consists of three steps: (i) pre-processing, including face detection and normalization, (ii) feature extraction, and (iii) applying the age estimation algorithm (Fig. 2). Regarding feature extraction, AAE systems can be divided into two groups: (i) systems that use hand-crafted features and (ii) systems based on deep learning. The systems that use hand-crafted features work reasonably well on images taken in constrained conditions (i.e. single face, frontally aligned, simple background etc.) [1], [5]. However, with recent development of in-the-wild datasets, hand-crafted methods have increasingly been surpassed by deep learning models for feature extraction. Deep learning models, especially convolutional neural networks (CNNs), have proven themselves to be more robust to noise, variations in appearance, pose

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

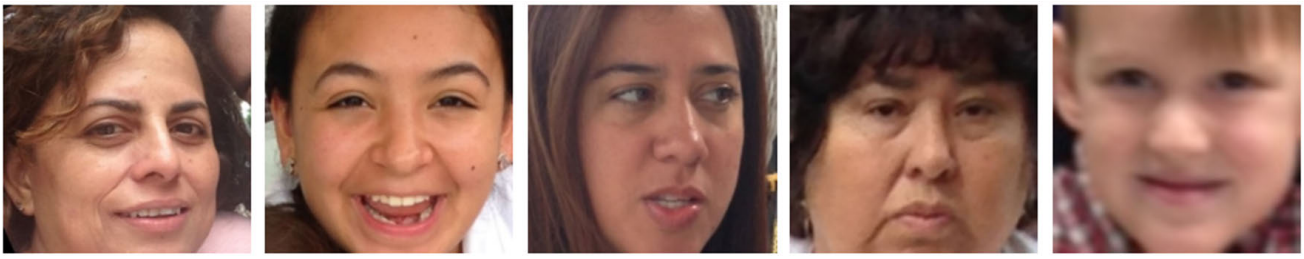

FIGURE 1. Examples of erroneous age estimation.

and lighting present in unconstrained datasets [6]. The problem of automated age estimation can be broadly framed either as a classification problem or as a regression problem [1], [7]. When framing age estimation as a classification problem, the classifier predicts an age group, e.g. ''35 to 39 years old''. Classification with soft labels is another possibility, in which class assignments are not binary. When framing age estimation as a regression problem, the goal is to predict the exact age as a number, e.g. ''29 years old''. Various hybrid approaches of classification and regression have also been proposed [7].

In this paper, we perform a comparative study of the current techniques suitable for the automated age estimation task, with an emphasis on running age estimation on embedded devices. Moreover, the intended application is adapting multimedia content to the age of a viewer, which does not require high age estimation accuracy. We investigate both the suitable modern deep learning architectures for feature extraction and the variants of framing the problem itself as either classification, regression or soft label classification. To gather in-depth insights into automated age estimation and in contrast to existing studies, we additionally compare the performance of both classification and regression on the same dataset. As the main contribution, we propose a novel loss function defined as a linear combination of regression and classification losses, where instead of manually selecting the linear coefficients, they are learned from data as trainable parameters of the model. This is achieved by incorporating constrains related to the coefficients into the loss function using a modification of the augmented Lagrangian formulation. We show that this formulation of loss outperforms all other considered approaches and achieves competitive results on FG-NET dataset. At the same time, with a lightweight backbone, our system is suitable for implementation on embedded devices. The new formulation of loss function does not affect the inference complexity, as it is used only for training.

II. RELATED WORK

Automated age estimation has been an actively researched topic in recent years, as detailed in a number of comparative surveys [3], [4], [6], [7], [8], [9], [10], [11]. While earlier works [1], [5] predominately focused on explicitly modeling the aging process using various computer vision techniques and hand-crafted features, current age estimation methods typically apply some form of deep learning. For example, in an early application of CNNs, Levi et al. [12] show how they outperform contemporary state-of-the-art by utilizing a neural network with only three convolutional and two fully connected layers. Wang et al. [13] propose using a combination of feature maps obtained at different layers of a CNN and manifold learning. A comparative analysis of several recent CNN architectures is presented in [8]. The authors also discuss the benefits of transfer learning, as well as the influence of noise in the images, variations in facial expressions and pose, ethnicity and other factors on the performance of the age estimation.

The performance of current automated age estimation methods has been shown to be equal to or better than human performance. For example, Han et al. [2] provide an experiment that measures the human accuracy of estimating the age from an image. On the FG-NET dataset [14], [15], the reported mean absolute error (MAE) achieved by humans is 4.7 years, while on the PCSO dataset [16] the MAE is 7.2. The authors show that their proposed hierarchical approach is capable of performing equally or better than humans on the age estimation task.

A seminal paper that uses deep learning for age estimation is the work of Rothe et al. [17] that introduces the Deep EXpectation (DEX) approach. In the first stage of their system, faces are detected and rotated to a normalized position, then the bounding box around them is extended by 40% to include more information, and the normalized face images are scaled to a standard size. The pre-processed face images are then fed to a CNN for age prediction. The DEX system uses the VGG-16 architecture [18] pre-trained on ImageNet as the feature extraction backbone. Using the same feature extractor, the authors explore both the regression and the classification approach to age estimation. During the training for regression, the authors noted instabilities and re-phrased the regression problem as classification into one year wide age ranges. In this re-phrased approach the network is trained for classification, but inference produces a single

value obtained by computing the expectation over all classes, taking into account the mean age values of the classes and the corresponding probabilities computed by the network. In the case of classification, the predicted age range corresponds to the output neuron with the highest probability. In all the experiments, the network is initialized with ImageNet weights, then further pre-trained on the IMDB-WIKI dataset introduced in the same paper, and finally trained on the target dataset. The authors test their system on multiple datasets, both for apparent and for real age estimation, and report mean average error of 3.09 years on the FG-NET dataset (exact age estimation) and the accuracy of 64% on the Adience dataset (classification into 8 predefined age groups). The authors also report the so called 1-off accuracy of 96.6%, which treats the neighboring age ranges as correct prediction.

To increase the accuracy of age estimation, various complex and non-standard models have been proposed. For example, Zhang et al. [19] propose several variants of models called Residual network of Residual networks (RoR) and achieve the state-of-the-art classification results on the Adience dataset with the classification accuracy of 67.34% for the best performing model. A follow-up work [20] further increases the model complexity by adding a visual attention mechanism based on LSTM units. Rodriguez et al. [21] also use attention mechanism to find the most informative image regions for the task of age estimation. Their attention module is based on the VGG-16 model trained to predict attention grid, which is then used to weight feature maps obtained from the corresponding image patches. Garain et al. [22] propose gated residual attention network, a deep learning model for gender and age estimation. They approach the age estimation problem as a combination of classification and regression. Guehairia et al. [23] propose a complex pipeline consisting of feature extraction using pretrained models for facial age estimation followed by a series of transformations in feature space for dimensionality reduction. They use deep random forest classifier to obtain final age estimation.

When age estimation is framed as standard classification, the classes are considered independent. However, some authors utilize the fact that aging is smooth and gradual and the age labels are ordered. Chen et al. [24] propose ranking-CNN, a CNN-based framework consisting of a series of basis binary CNNs, each trained to distinguish whether the age is less than or greater than a given threshold. The outputs of these basis CNNs are aggregated to obtain the final age estimation. Cao et al. [25] propose the Consistent Rank Logits (CORAL) framework, an architecture-agnostic framework with theoretical guarantees for rank monotonicity and confidence score consistency. They demonstrate that their method outperforms the standard cross-entropy loss that does not utilize the rank ordering. Shin et al. [26] propose an approach called Moving Window Regression (MWR). In MWR, training and test images are represented in feature space, obtained by VGG-16-based encoder network, followed by several additional fully connected layers. Each test image is then compared to two reference images in feature space, producing a so-called relative rank (or ρ-rank) – a numerical value indicating the relative position of the test image between the two reference images according to considered criterion (e.g. estimated age). The process is repeated iteratively, by selecting new test image pair in each iteration within the search window centered around the previous rank estimate, to refine the estimated age rank, until the convergence, i.e. until the estimated rank is at the center of the search window.

Geng et al. [27] propose modeling face images with a label distribution, and introduce two algorithms for learning automated age estimation from label distributions. Diaz et al. [28] propose a general framework for converting data labels into soft probability distributions suitable for ordinal regression, which addresses problems where class labels follow some inherent ordering, such as in age estimation. In an experiment with age estimation on the Adience dataset, they demonstrate that using soft labels improves the accuracy for about 2% over baseline. Antipov et al. [29] analyze optimal choices for CNN training for age estimation, including target age encoding, loss function, CNN depth, whether pre-training is performed or not and whether the problem is framed as mono-task or multi-task. They also show that label distribution encoding seems to be a better choice for automated age estimation CNNs than single number encoding. Pan et al. [30] propose adaptive label distribution learning by complementing the softmax loss with two additional loss functions: the mean loss which penalizes the deviation of the predicted distribution mean from the ground-truth age, and the variance loss which promotes narrow distributions. This approach is known as Mean-Variance loss. A similar approach is proposed recently by Zhao et al. [31]. They also combine softmax and mean loss, while the variance loss is substituted with the residue loss that penalizes the residue errors in the tails of estimated age distribution after the top-K pooling operation. Li et al. [32] observe that the Mean-Variance loss does not enforce unimodality of the learned distribution and propose a new variant called Unimodal-Concentrated loss. The unimodal loss is used instead of softmax loss and requires that the learned distribution values get smaller at larger distances from the ground truth age, and thus promotes unimodality. The concentrated loss simultaneously penalizes the deviation of the estimated age from the ground truth and promotes small variance of the predicted distribution.

While most of the above discussed approaches aim primarily to improve age estimation performance, some other authors focus on lightweight compact models suitable for use on more modest hardware resources, while trying not to sacrifice too much of the age estimation accuracy. Zhang et al. [33] propose an extremely compact and efficient regression model named Cascade Context-based Age Estimation (C3AE). The C3AE model has 1/9 parameters compared to MobileNets/ShuffleNets and 1/2000 of the parameters of VGG-16, while achieving performance similar to the DEX

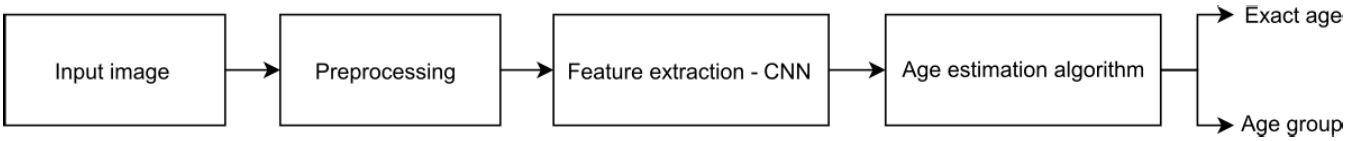

FIGURE 2. Age estimation pipeline.

model on the FG-NET dataset. Deng et al. [34] propose fusing age, gender and race features for more precise age estimation. They argue that the aging process is influenced by gender and race, and more precise age estimation can be obtained when gender and race information is present. They use a hybrid regression-ranking estimator to obtain the exact age values. The proposed model is relatively compact and suitable for use on mobile and embedded devices. Greco et al. [35] propose an effective training method for age estimation CNNs based on knowledge distillation. The goal of the method is to first learn age estimation using a more complex neural network and then distill the knowledge to a smaller, more compact model.

In our work, we are looking for a reasonably small model suitable for restricted hardware resources and embedded implementation, while providing state-of-the-art age estimation accuracy. Moreover, we try to use standard off-the-shelf models as building blocks in order to benefit from the availability of pre-trained weights obtained on large image classification datasets. We combine several of the ideas described above, such as representing classes with soft labels, expressing regression by computing expected value over softmax outputs, and penalizing the deviation of the mean value of the model output from the ground truth age [30]. Our best-performing model combines soft-label classification with regression through a new custom loss function.

III. METHODOLOGY

In this work we divide the age estimation pipeline into three phases as illustrated in Fig. 2. As part of the preprocessing, faces in the input image are detected and aligned. Then, each of the cropped face images is passed through a convolutional neural network responsible for facial feature extraction. Finally, one of the considered algorithms is applied to estimate either the exact age or the corresponding age group.

A. PREPROCESSING

The preprocessing stage starts with detecting faces in the input image using a model based on a convolutional neural network from *dlib* library [36]. Then, the image is aligned based on the located facial keypoints (also from *dlib*) with respect to the eyes' positions. The image is rotated around the center between the eyes so that the eyes become horizontal, and the face is cropped with the size of detection rectangle. Finally, the face image is standardized to have zero mean and variance one and scaled to 256×256 pixels (Fig. 3).

FIGURE 3. Input image preprocessing.

B. FEATURE EXTRACTION WITH CNN

We consider several standard CNN architectures and take their convolutional part for feature extraction. Depending on the variant of the age estimation task (e.g. classification or regression), we add the corresponding head consisting of two fully connected layers. The first dense layer with 256 neurons is followed by batch normalization and dropout (with probability 0.25). The last dense layer has the corresponding number of outputs.

The considered architectures are the following:

- VGG architecture [18]. It has several configurations with the same architectural pattern that differ only in network depth, varying from 11 to 19 layers. All convolutional layers have kernel size 3×3 with stride of 1. In this work, we use the convolutional part of the VGG-16 architecture (13 convolutional layers) as feature extractor.
- GoogLeNet architecture [37]. The network consists of 22 layers and it is based on Inception modules. Features in these modules are obtained by combining convolutional layers with different kernel sizes. This is based on the idea that convolution filters of different sizes would handle objects at multiple scales better. In addition, GoogLeNet has two auxiliary classifiers used exclusively during training. The purpose of these additional classifiers is to prevent the vanishing gradient problem and to provide regularization.
- ResNet architecture [38]. In that model a technique called skip connections is used to solve the vanishing gradient problem that appears when training very deep architectures. The model is composed of blocks of convolutional layers with the addition of residual shortcut connections. In this work, the ResNet50V2 architecture is used as a feature extraction backbone.

- MobileNet [39] is a family of architectures proposed specifically with the efficiency of execution on mobile and embedded hardware in mind. The basic idea, introduced with the first generation of the model is the use of depthwise separable convolutions – the regular convolutions are replaced by a series of two computationally lighter convolutions: the spatial 3×3 convolution applied to each input channel separately, followed by a 1×1 cross-channel convolution. The MobileNetV2 [40], used in our experiments, improves the basic building block module by using linear bottlenecks and inverted residuals.

C. AGE ESTIMATION ALGORITHM

We experiment with both age group and exact age estimation. For exact age estimation, we consider three models: (i) regression, (ii) fine-grained classification with soft labels, and (iii) our novel model with custom loss.

1) AGE GROUP CLASSIFICATION

For age group estimation, images are labelled with one of the age groups. In our experiments the Adience dataset is used, that defines eight age groups (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+), so that we use a classification model with eight output neurons, and softmax activation function. Weights of the classification model are learned by optimizing the categorical cross-entropy loss. Categorical cross-entropy measures the difference between two probability distributions and can be defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_j \log p_j \tag{1}$$

where N is the total number of samples, C is the number of classes, y is ground truth distribution vector (one-hot) and p is probability vector obtained by the output of the model.

2) REGRESSION MODEL

For the regression model, images are labeled with the exact age of a person. The model backbone is the same as in the classification model, but now the model has only one output neuron, with ReLU activation function. The model parameters are learned by optimizing the mean squared error (MSE) loss function defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2 \tag{2}$$

where N is total number of samples, y_i the estimated age and t_i the ground-truth age for the example i.

3) FINE-GRAINED CLASSIFICATION WITH SOFT LABELS

Apart from estimating the exact age from images using a regression model with one output, it can also be done using a discrete classification model. By increasing the number of output classes (i.e., shrinking the age ranges), the discretization error gets smaller. In [17], output layer of the model has 101 output neurons corresponding to the ages from 0 to 100. Final prediction can be made as an *argmax* of the output layer, or it can be calculated as the expected value of the output layer. We implemented a similar approach, with the difference that instead of one-hot encoded labels used in [17], we use soft labels [28]. The motivation for introducing soft labels is the fact that classes in this problem are not independent, but they are ordered – a misclassification into a neighboring age class should be penalized less than a larger error. One way of achieving this is by forming ground truth vectors by assigning non-zero values to neighboring classes. We adopt the ground truth label vectors computation from [28]:

$$y_i = \frac{e^{-\phi(r_t, r_i)}}{\sum_{k=1}^{K} e^{-\phi(r_t, r_k)}}, \forall r_i \in Y \tag{3}$$

where $\phi(r_t, r_i)$ is a metric loss function that penalizes the distance between the true value r_t and the class value r_i and K is the number of classes. We use the squared difference of class values as function ϕ. Final prediction is calculated as the expected value of the *softmax* layer outputs:

$$y_p = \sum_{k=1}^{K} r_k p_k \tag{4}$$

where p_k are *softmax* output probabilities, and r_k are discrete years corresponding to each class k. For training, the standard cross-entropy loss is used (Eq. 1).

4) HYBRID APPROACH WITH CUSTOM LOSS

The soft-label classifier described above addresses the ordinality of the classes in the age estimation problem. However, the shape of the ground truth "distribution" is somewhat arbitrary and not necessarily optimal for learning the model parameters. Moreover, the signal of the ground truth exact age is weakened compared to one-hot classification, as well as to regression. Therefore, we propose a new hybrid approach, which combines (i) the cross-entropy loss of predicted and ground truth soft labels (Eq. 1) with (ii) mean absolute error (Eq. 5) between the real age and the age computed from output probabilities.

$$L = \frac{1}{N}\sum_{i=1}^{N} |y_i - t_i| \tag{5}$$

A similar approach is presented in [30], where a classification model with one-hot encoded labels is trained. There, besides maximizing the probability of the ground truth class, the loss function penalizes the difference between the mean value of the model output and the ground truth age, as well as the variance of the estimated distribution (for the purpose of obtaining a narrow output distribution).

We propose to combine the fine-grained classification with soft labels approach and the regression through a linear combination of the cross-entropy loss (L_1) and the mean absolute error loss (L_2). Instead of fixing the linear combination coefficients α and β, we propose to define them as additional

model parameters and learn them from data. The combined loss function can be defined as:

$$L = \alpha^2 L_1 + \beta^2 L_2 \tag{6}$$

with the constraint on α and β:

$$\alpha^2 + \beta^2 = 1 \tag{7}$$

The learnable parameters α and β are used squared in order to prevent negative loss contribution. One way to include the constraint (7) into the optimization procedure is to add it to the loss function as an additional regularization term using the Augmented Lagrangian [41]:

$$L = \alpha^2 L_1 + \beta^2 L_2 + \mu_1 \left(\alpha^2 + \beta^2 - 1\right) + \mu_2 \left(\alpha^2 + \beta^2 - 1\right)^2 \tag{8}$$

where μ_1 and μ_2 are hyperparameters of the model. The first three terms of the function cause α^2 and β^2 values to approach zero, while the last term penalizes the deviation from the constraint (7).

However, this formulation exhibits one obvious weakness: if the values of one loss component have significantly larger values than the other, the corresponding coefficient will tend to become zero in the optimization process. This situation would eliminate the possible benefits of combining the two different approaches. Therefore, we introduce additional two terms that penalize approaching α^2 or β^2 values to zero, and obtain the final loss function:

$$L = \alpha^2 L_1 + \beta^2 L_2 + \mu_1 \left(\alpha^2 + \beta^2 - 1\right) + \mu_2 \left(\alpha^2 + \beta^2 - 1\right)^2 + \mu_3 \left(1 - \alpha^2\right)^2 + \mu_4 \left(1 - \beta^2\right)^2 \tag{9}$$

where μ_1, μ_2, μ_3 and μ_4 are hyperparameters of the model.

We choose an appropriate set of hyperparameter values based on analysis of their relative relations and with respect to the magnitudes of the loss functions, as well as experimental validation. We follow [41] in their suggestion that μ_2 should be significantly greater than μ_1. Furthermore, by looking at the values of the loss functions during training, we concluded that the values of μ_3 and μ_4 should be somewhere in between the values of μ_1 and μ_2 in order to influence the loss function in a meaningful way, defining thereby the relative relationships between the four hyperparameters. Finally, we performed a series of experiments with different combinations of values of the four hyperparameters, that confirmed our qualitative analysis. As an illustration we show training progress for parameters α^2 and β^2 for two characteristic sets of hyperparameters in Fig. 4. The experiments also show that the final age estimation performance is quite similar for different values of the hyperparameters, as long as they are kept in meaningful ranges. This indicates that the loss function is relatively robust with respect to the hyperparameter valus. However, in the experiments, the best results were obtained for the combination $\mu_1 = 0.1$, $\mu_2 = 200$, $\mu_3 = 10$, $\mu_4 = 10$.

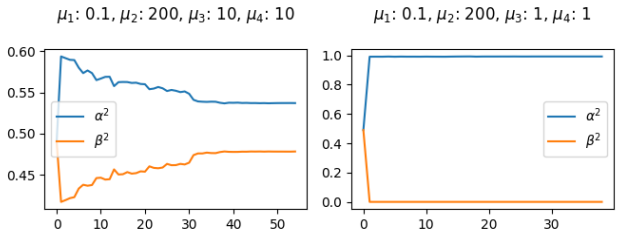

FIGURE 4. Training progress for parameters α^2 and β^2 for the chosen hyperparameters (left) and the insufficiently high values of the hyperparameters μ_3 and μ_4 (right).

IV. EXPERIMENTS AND RESULTS

Our goal is to investigate which approach and model are best appropriate for age estimation. We prefer lighter models suitable for embedded implementation. Our intended application is adapting multimedia content to the age of a viewer, so that even a rough estimation is satisfactory. To that end, we explore two main problem formulations: age group estimation vs. exact age estimation, employing several typical CNN architectures of different complexity, and different approaches to the age estimation problem.

We evaluate the considered methods on two typical datasets: Adience for age group estimation and FG-NET for exact age.

To reduce the risk of overfitting, data augmentation techniques are applied during the training of the models. Input images (obtained from the preprocessing step) are first scaled to 256 × 256 and then randomly cropped to 227 × 227 and randomly flipped horizontally, similarly to [12] and [28]. For testing, central crop of size 227 × 227 is taken. Adam optimizer is used for training of the models. Initial learning rate is set experimentally to 0.001 for all models, and reduced by factor of 0.1 if the loss on the validation set does not reduce for consecutive five epochs. Also, early stopping is applied if the result on the validation set does not improve for consecutive ten epochs. The models are trained for maximally 100 epochs with batch size 64.

All experiments have been performed on a standard PC with Intel i7 processor and NVIDIA RTX 2080Ti GPU. However, the final model is intended to be used on a variety of embedded platforms, ranging from embedded PCs without a GPU to specialised embedded computers with a dedicated GPU such as NVIDIA Jetson family.

A. DATASETS

There are many age-related datasets tailored to various age estimation tasks (real and apparent age estimation, exact age or age group estimation). A comprehensive overview can be found in e.g. [4]. To train and evaluate our models, we used three datasets: IMDB-WIKI for model pretraining, and two typical benchmark datasets: FG-NET and Adience for final training and evaluation.

- IMDB-WIKI [17] is the largest publicly available dataset with age annotations, containing images taken in unconstrained conditions. It contains 523,051 images of 20,284 subjects with ages ranging from 0 to 100 years. Images were automatically crawled and labelled so that there are many incorrect or missing labels. As we use this dataset for pretraining purposes only, we decided to improve the quality of the dataset by using a simple filtering technique: images with age annotations obviously erroneous (age under 0 or above 100 years) are removed, as well as the images with the provided face detector scores indicating that face detection was not unique (either none or multiple faces were detected). The cleaned version of the dataset contains 221,641 images.
- FG-NET [5], [15] is a publicly available dataset containing 1,002 images of 82 different subjects with ages ranging from 0 to 69. Each image is labeled with the exact age of the person in years. Images were collected by scanning photographs of subjects found in personal collections, so it contains variations in head poses, facial expressions, and illumination. For evaluation on the FG-NET dataset, we follow the widely used leave-one-person-out (LOPO) protocol [15].
- Adience [42] is a collection of images captured in real, unconstrained conditions. As the source of images, Flickr albums were used, so that noise is present in the images and there are variations in appearance, pose and lighting. The dataset contains 26,580 images of 2284 individuals. Images are labeled with one of eight age groups: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+. For evaluation, the 5-fold cross validation is applied, using the folds prepared by the authors of the dataset.

During age group estimation experiments, we observed that in the pretraining dataset (IMDB-WIKI) some age groups are significantly underepresented (age groups 0-3, 4-7, 8-13 represent less than 2% of the dataset). We tried to alleviate that problem using several standard procedures: oversampling underrepresented classes, undersampling overrepresented classes, weight balancing, and focal loss [43]. None of those approaches to pretraining improved accuracy on Adience dataset so we decided to extend the pretraining dataset with additional images for underrepresented classes collected from additional face datasets (UTKFace, AgeDB and APPA-REAL). This extension improved the accuracy on Adience, so for the rest of the experiments the extended dataset was used for pretraining.

B. EVALUATION MEASURES AND PROTOCOLS

We evaluate two age estimation settings: the exact age and the age group estimation, for which we use two different measures.

For exact age estimation the mean absolute error (MAE) is used:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_t - y_p\right| \tag{10}$$

where n is the total number of samples, y_t is the ground truth age and y_p is the estimated age.

Another often used performance measure for exact age estimation is the so-called cumulative score, which represents the classification accuracy for different allowed deviations of the estimate from the ground truth age. That allows to display the performance as a graph, giving a better insight into the estimator behavior (cf. Fig. 7). Sometimes only a single value from the graph is given, usually the cumulative score for error within 5 years – $CS(5)$.

Age group estimation is evaluated by accuracy:

$$\mathrm{Acc} = \frac{n_T}{n} \tag{11}$$

where n is the total number of samples and n_T the number of correctly classified samples.

Considering the task at hand, the classification accuracy is not very informative since it does not take into account the amount of estimation error in the case of misclassification. For some applications, a misclassification into a neighboring class could be acceptable. Therefore, the so-called *1-off* measure is used, representing the accuracy when the classification into a neighboring class is considered correct.

The age estimation algorithms are usually evaluated using some variant of cross-validation. The Adience dataset has five prepared subject-exclusive folds that are used for standard 5-fold cross validation [42]. Each fold is used for validation, while the remaining four are used for training, and then the average score is reported, together with the standard deviation. The FG-NET dataset is rather small (1002 images, 82 subjects), so that we follow the leave-one-person-out (LOPO) cross validation protocol [14], [15].

C. AGE GROUP ESTIMATION EXPERIMENTS

In the first set of experiments we applied several common backbones (GoogLeNet, VGG-16, ResNet50V2, MobileNetV2). For all experiments we add the same classification head having 8 output neurons as dictated by Adience age groups. Following good practices described in [17], we pretrain our models on IMDB-WIKI. For pretraining, we split the dataset randomly in the ratio of 80:20 for training and validation. Here we also try to explore the effects of initializing the pretraining backbone with ImageNet weights, compared to random initialization.

The results (Table 1) show that all considered architectures (with suitably tuned hyperparameters) achieve similar results, regardless of the model complexity and even of the initialization. The results are similar to other state-of-the-art approaches, but somewhat worse (Deep Attention [21] achieves 61.8%, DEX [17] 64%, and Deep RoR [19] 67.34% accuracy).

TABLE 1. Results on adience dataset.

Model (initialization)	Adience accuracy (%)	1-off	Parameters
ResNet50v2 (random)	58.6 (±4.7)	93.8 (±1.4)	24.26M
ResNet50v2 (Imagenet)	58.9 (±4.3)	93.3 (±1.2)	24.26M
GoogLeNet (random)	57.3 (±4.0)	94.0 (±4.4)	8.5M
GoogLeNet (Imagenet)	59.7 (±4.1)	94.3 (±5.0)	8.5M
VGG-16 (random)	58.0 (±5.3)	93.1 (±1.7)	27.6M
VGG-16 (Imagenet)	59.4 (±4.1)	92.9 (±1.6)	27.6M
MobileNetV2 (random)	60.0 (±4.8)	94.1 (±1.2)	2.6M
MobileNetV2 (Imagenet)	**61.1 (±3.9)**	**94.9 (±1.3)**	**2.6M**
Deep Attention [21]	61.8 (±2.1)	95.1 (±0.03)	>138M
DEX [17]	64.0 (±4.2)	96.6 (±0.9)	138M
Deep RoR [19]	**67.34 (±3.56)**	**97.51 (±0.67)**	–

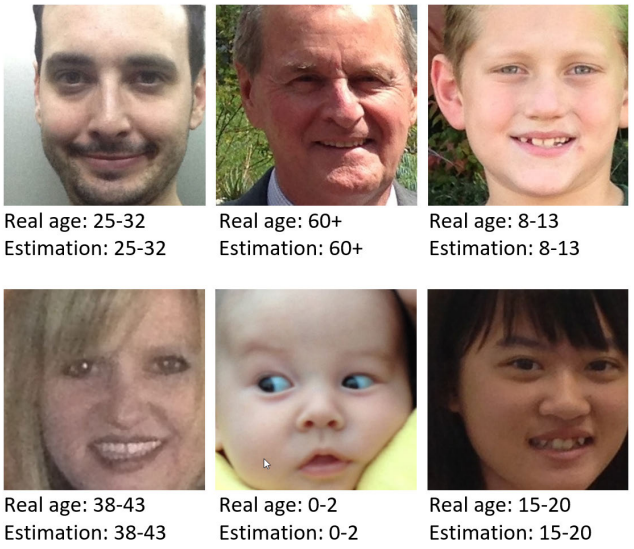

FIGURE 5. Examples of images with correct age group predictions on Adience dataset obtained by the MobileNet-based model.

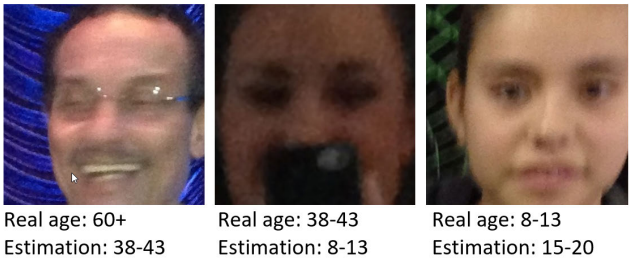

FIGURE 6. Examples of images with wrong age group predictions on Adience dataset obtained by the MobileNet-based model.

TABLE 2. Results on FG-NET dataset.

Model	FG-NET MAE (LOPO)
Regression – MobileNet	2.92 (±1.36)
Regression – ResNet	2.77 (±1.16)
Soft labels – MobileNet	2.67 (±1.39)
Soft labels – ResNet	2.55 (±1.37)
Custom loss – MobileNet	**2.40 (±1.34)**
Custom loss – ResNet	2.46 (±1.44)
DEX [17]	3.09
Age difference [44]	2.80
Mean-Variance Loss [30]	2.68
C3AE [33]	2.95 (±0.17)
AL-RoR-34 [20]	2.39
Deng et al. [34]	2.59
Guehairia et al. [23]	3.05
Adaptive Mean-Residue Loss [31]	3.61
Moving Window Regression [26]	**2.23**

Still, the *1-off* measure (almost 95%) shows that such an approach can be used for rough age estimation in our target application. Since we aim the execution on embedded hardware, we prefer a lighter model such as MobileNet, which moreover obtains the best accuracy in our experiments. Some examples of age predictions obtained by the MobileNet-based model are shown in Fig. 5 and Fig. 6.

D. EXACT AGE ESTIMATION EXPERIMENTS

For exact age estimation the three models described in section III-C are compared on FG-NET dataset, using LOPO protocol. For all exact age estimation experiments we chose (i) MobileNetV2 architecture because it achieved the best results on the age group estimation task, while being the lightest model, and (ii) ResNet-50V2 as a representative of large capacity models. The obtained results are given in Table 2, together with a few state-of-the-art results.

The table shows that performance of our hybrid model with custom loss is better than both regression and classification with soft labels, regardless of the applied feature extraction backbone. Moreover, the lighter model performs better than the more complex one, further encouraging our intention to use it for embedded application. In comparison with other state-of-the-art models, our model performs very competitively. The obtained MAE is close to that of the best performing Moving Window Regression approach [26], while our model is significantly simpler. Their model is based on VGG-16 backbone, followed by fully connected layers and iterative MWR process for ordinal regression, while our model uses MobileNetV2, which is much more appropriate for embedded platforms. It is worth noting that our loss function modification does not influence the inference complexity. Some examples of age predictions obtained by our hybrid approach are shown in Fig. 8 and Fig. 9.

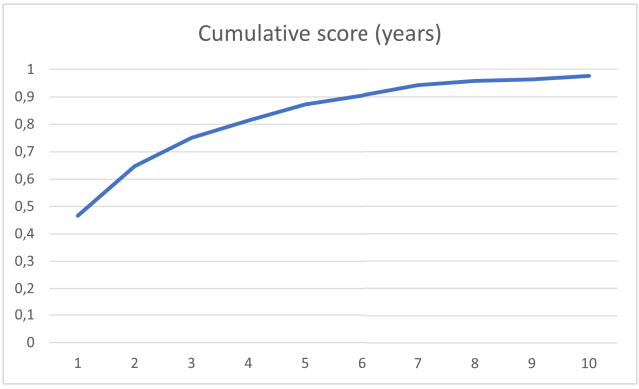

FIGURE 7. Cumulative score graph for MobileNet-based model with custom loss on FG-NET dataset.

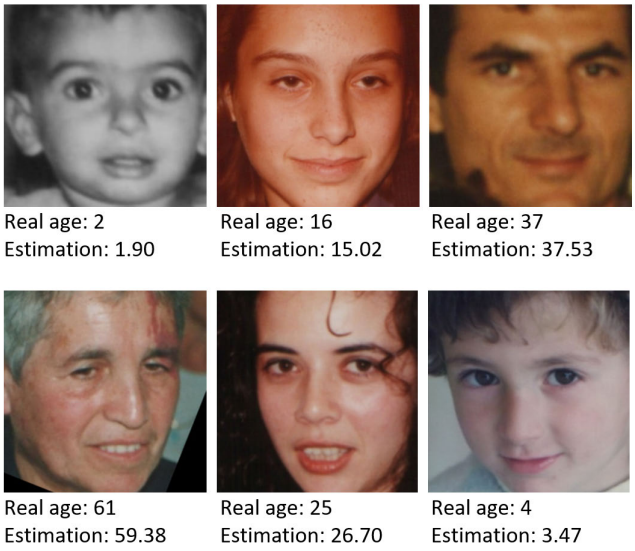

FIGURE 8. Examples of images with small age estimation errors on FG-NET dataset obtained by the MobileNet-based model.

FIGURE 9. Examples of images with large age estimation errors on FG-NET dataset obtained by the MobileNet-based model.

To get some more insight into the performance of our hybrid model with MobileNet backbone, we show the cumulative score graph (Fig. 7). We can see from the graph that $CS(5) = 87.3\%$. For comparison, the best performing Moving Window Regression approach obtains $CS(5) = 91.1\%$.

TABLE 3. Classification results on FG-NET dataset.

Model	Accuracy (%)	1-off
MobileNetV2 classification	70.3 (±15.6)	96.0 (±7.4)
MobileNetV2 regression	62.5 (±14.1)	**97.8 (±4.8)**
MobileNetV2 custom loss	**73.7 (±15.1)**	97.6 (±6.9)

E. COMPARING EXACT AGE WITH AGE GROUP ESTIMATION

In previous two sections we tried to evaluate the capability of a light convolutional architecture for age estimation task in order to select the most appropriate approach for our application. However, the two problem formulations (age group vs. exact age) are tested on different datasets and with different quality measures, so that it is still difficult to compare the results. While 61% accuracy obtained on age group estimation task seems modest, the *1-off* measure of 95% gives a much more optimistic view having in mind that we need only a rough age estimation. On the other hand, the results with exact age estimation approach (MAE of 2.5 years, which is current state-of-the-art on FG-NET dataset) seems very promising. However, in order to get a better insight, it would be interesting to apply both approaches to the same dataset. One way to do that would be to apply an exact age estimator to the age group estimation task: the exact age estimate could be simply put into the corresponding age group, thus obtaining classification.

Since Adience dataset does not have exact age labels, we cannot use it for training the exact age estimator. Therefore, we adjusted the FG-NET dataset by adding the age group annotations using the same groups as in Adience, and used it to train our age group classifier. Then we compared the classification performance of the age group classifier with classification based on exact age estimation. We present only the results for MobileNet based models, which are best suited for our application.

The results show that classification model performs better on FG-NET dataset than on Adience, i.e. we can conclude that FG-NET can be considered an easier dataset. It is not surprising since Adience images are taken in less restricted conditions (''in-the-wild''), while FG-NET contained scanned document photographs. Regarding the considered methods, simple regression model performed worse on the classification task than the model trained specifically for classification. However, the exact age estimation model with custom loss (that combines soft label fine-grained classification with regression) obtained the best results.

V. CONCLUSION

We considered several approaches to age estimation problem. All evaluated architectures used standard convolutional backbones for feature extraction, while the output head was configured according to the defined task. In all experiments we pretrained the backbone on a large face image dataset (IMDB-WIKI). The first approach was based on

classification into predefined age groups and it was evaluated on Adience dataset. The experiment showed that models using backbones of very different capacity obtained similar results, thereby supporting the application of a lighter model appropriate for embedded implementation. The next series of experiments explored several model configurations for exact age estimation: simple regression, fine-grained classification with soft labels, and our novel hybrid approach combining regression with soft-label fine-grained classification through (our original) custom loss. The hybrid approach performed the best, obtaining the state-of-the-art result on FG-NET dataset. A final experiment was designed to compare different approaches on a common task and dataset. To that end we adapted the exact age dataset FG-NET for age group estimation task. Our hybrid approach outperformed both classification and regression models.

REFERENCES

[1] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1955–1976, Sep. 2010.

[2] H. Han, C. Otto, and A. K. Jain, "Age estimation from face images: Human vs. machine performance," in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2013, pp. 1–8.

[3] V. Carletti, A. Greco, G. Percannella, and M. Vento, "Age from faces in the deep learning revolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2113–2132, Sep. 2020.

[4] O. Agbo-Ajala and S. Viriri, "Deep learning approach for facial age classification: A survey of the state-of-the-art," *Artif. Intell. Rev.*, vol. 54, no. 1, pp. 179–213, Jan. 2021.

[5] A. Lanitis, "Comparative evaluation of automatic age progression methodologies," *EURASIP J. Adv. Signal Process.*, vol. 2008, no. 1, 2008, Art. no. 239480.

[6] R. Angulu, J. R. Tapamo, and A. O. Adewumi, "Age estimation via face images: A survey," *EURASIP J. Image Video Process.*, vol. 2018, no. 1, pp. 1–35, Dec. 2018.

[7] A. S. Al-Shannaq and L. A. Elrefaei, "Comprehensive analysis of the literature for age estimation from facial images," *IEEE Access*, vol. 7, pp. 93229–93249, 2019.

[8] A. Othmani, A. R. Taleb, H. Abdelkawy, and A. Hadid, "Age estimation from faces using deep learning: A comparative analysis," *Comput. Vis. Image Understand.*, vol. 196, Jul. 2020, Art. no. 102961.

[9] P. Punyani, R. Gupta, and A. Kumar, "Neural networks for facial age estimation: A survey on recent advances," *Artif. Intell. Rev.*, vol. 53, no. 5, pp. 3299–3347, Jun. 2020.

[10] A. A. Shejul, K. S. Kinage, and B. E. Reddy, "Comprehensive review on facial based human age estimation," in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput. (ICECDS)*, Aug. 2017, pp. 3211–3216.

[11] K. ELKarazle, V. Raman, and P. Then, "Facial age estimation using machine learning techniques: An overview," *Big Data Cognit. Comput.*, vol. 6, no. 4, p. 128, Oct. 2022.

[12] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 34–42.

[13] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 534–541.

[14] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, Apr. 2002.

[15] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "Overview of research on facial ageing using the FG-NET ageing database," *IET Biometrics*, vol. 5, no. 2, pp. 37–46, May 2016.

[16] D. Deb, L. Best-Rowden, and A. K. Jain, "Face recognition performance under aging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 548–556.

[17] R. Rothe, R. Timofte, and L. Van Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, pp. 144–157, Apr. 2016.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14. [Online]. Available: https://arxiv.org/abs/1409.1556

[19] K. Zhang, "Age group and gender estimation in the wild with deep RoR architecture," *IEEE Access*, vol. 5, pp. 22492–22503, 2017.

[20] K. Zhang, "Fine-grained age estimation in the wild with attention LSTM networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 3140–3152, Sep. 2020.

[21] P. Rodríguez, G. Cucurull, J. M. Gonfaus, F. X. Roca, and J. González, "Age and gender recognition in the wild with deep attention," *Pattern Recognit.*, vol. 72, pp. 563–571, Dec. 2017.

[22] A. Garain, B. Ray, P. K. Singh, A. Ahmadian, N. Senu, and R. Sarkar, "GRA_Net: A deep learning model for classification of age and gender from facial images," *IEEE Access*, vol. 9, pp. 85672–85689, 2021.

[23] O. Guehairia, F. Dornaika, A. Ouamane, and A. Taleb-Ahmed, "Facial age estimation using tensor based subspace learning and deep random forests," *Inf. Sci.*, vol. 609, pp. 1309–1317, Sep. 2022.

[24] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 742–751.

[25] W. Cao, V. Mirjalili, and S. Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *Pattern Recognit. Lett.*, vol. 140, pp. 325–331, Dec. 2020.

[26] N.-H. Shin, S.-H. Lee, and C.-S. Kim, "Moving window regression: A novel approach to ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18739–18748.

[27] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2401–2412, Oct. 2013.

[28] R. Díaz and A. Marathe, "Soft labels for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4733–4742.

[29] G. Antipov, M. Baccouche, S. Berrani, and J. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognit.*, vol. 72, pp. 15–26, Dec. 2017.

[30] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5285–5294.

[31] Z. Zhao, P. Qian, Y. Hou, and Z. Zeng, "Adaptive mean-residue loss for robust facial age estimation," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[32] Q. Li, J. Wang, Z. Yao, Y. Li, P. Yang, J. Yan, C. Wang, and S. Pu, "Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20481–20490.

[33] C. Zhang, S. Liu, X. Xu, and C. Zhu, "C3AE: Exploring the limits of compact model for age estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12579–12588.

[34] Y. Deng, S. Teng, L. Fei, W. Zhang, and I. Rida, "A multifeature learning and fusion network for facial age estimation," *Sensors*, vol. 21, no. 13, p. 4597, Jul. 2021.

[35] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "Effective training of convolutional neural networks for age estimation based on knowledge distillation," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21449–21464, Dec. 2022.

[36] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[39] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[41] H. Hajiabadi, D. Molla-Aliod, R. Monsefi, and H. S. Yazdi, "Combination of loss functions for deep text classification," *Int. J. Mach. Learn. Cybern.*, vol. 11, no. 4, pp. 751–761, Apr. 2020.
[42] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2170–2179, Sep. 2014.
[43] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1115–2196, Dec. 2019.
[44] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan, "Facial age estimation with age difference," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3087–3097, Jul. 2017.

ZORANA DOŽĐOR received the M.Sc. degree from the University of Zagreb, Croatia, in 2021, where she is currently pursuing the doctoral degree.

She is a Junior Researcher with the Faculty of Electrical Engineering and Computing, University of Zagreb. Her research interests include computer vision and deep learning.

TOMISLAV HRKAĆ (Member, IEEE) received the Ph.D. degree in computer science from the University of Zagreb, Croatia, in 2009.

He is an Associate Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb, where he is a member of the Laboratory for Pattern Recognition and Biometric Security Systems and the Center of Excellence for Computer Vision. His research interests include computer vision and deep learning.

KARLA BRKIĆ received the Ph.D. degree in computer science from the University of Zagreb, Croatia, in 2013.

She is an External Associate with the Faculty of Electrical Engineering and Computing, University of Zagreb. Her research interests include deep learning and computer vision and their applications in security and privacy.

ZORAN KALAFATIĆ (Member, IEEE) received the Ph.D. degree in computer science from the University of Zagreb, Croatia, in 1999.

He is an Associate Professor with the Faculty of Electrical Engineering and Computing, University of Zagreb, where he is a member of the Center of Excellence for Computer Vision and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems. His research interests include computer vision and deep learning.