Pose and Expression Robust Age Estimation via 3D Face Reconstruction from a Single Image

Nedko Savov[1,†], Minh L. Ngo[1,2,†], Sezer Karaoğlu[1,2], Hamdi Dibeklioğlu[3], and Theo Gevers[1,2]
[1] 3DUniversum, Amsterdam, The Netherlands
[2] Computer Vision Lab, University of Amsterdam, Amsterdam, The Netherlands
[3] Department of Computer Engineering, Bilkent University, Ankara, Turkey
n.savov@3duniversum.com, l.m.ngo@uva.nl, s.karaoglu@3duniversum.com,
dibeklioglu@cs.bilkent.edu.tr, th.gevers@uva.nl

Abstract

In this paper, we present a deep learning architecture that exploits 3D face reconstruction to obtain a robust age estimation. To this end, effective representation is learned through an expression-, pose-, illumination-, reflectance-, and geometry-aware deep model reconstructing a 3D face from a single 2D image. The 3D face reconstruction network is combined with an appearance-based age estimation network, where the face reconstruction features are jointly learned with the visual ones. Experiments on large-scale datasets show that our method obtains promising results and outperforms state-of-the-art methods, especially in the presence of strong expressions and large pose variations. Furthermore, cross-dataset experiments show that the proposed method is able to generalize more effectively as opposed to state-of-the-art methods.

1. Introduction

The human face is an important source of information. Face properties may reveal different important cues such as emotion, intent, ethnicity, identity, gender, and age. The focus of this paper is age estimation. Age estimation has many potential applications in daily life. For instance, in marketing, it can be employed for analyzing which age groups are interested in what kind of products, services, or entertainment. Vending machines of tobacco and alcohol can use age estimation to determine if the user is of legal age.

However, due to the large variation of aging patterns, addressed by [2, 10], age estimation is a challenging task. Existing methods mostly rely on 2D information by exploiting appearance-related features. These features are either handcrafted [9, 13, 30, 39]) or obtained in a learning manner (*e.g.* through Convolutional Neural Networks (CNNs) [15, 20, 31, 34, 38, 40]). Other methods use pose dependent distances between 2D facial landmarks [7, 19] or learn manifolds to directly map 2D images to age.

Methods relying on 2D features have difficulties when the face appearance changes. For instance, a change in expression may introduce disturbing age-related patterns, like wrinkles, and may negatively influence the accuracy of age estimation methods [14]. Head pose variations that drastically change the facial appearance may also degrade the accuracy of age estimation algorithms [8, 22]. These variations cause issues for other visual facial analysis tasks as well, like expression recognition [32] and landmark detection [8]. Robustifying methods for dealing with these variations are extensively explored for face identification [5]. One subset of solutions attempt to remove the variations from the input image by face frontalization or expression normalization, as a pre-processing step for face identification [1, 42]. In that case, any failure from the normalization, being the inability to normalize or the presence of artifacts on the generated image, negatively affects the performance. Such approaches may help to preserve the identity related dominant face features which makes them suitable for identification. However, the reconstructed images lose important high-frequency information such as skin texture detail (i.e. wrinkles) which would reduce age estimation accuracy.

In this paper, effective representation is learned through an expression-, pose-, illumination-, reflectance-, and geometry-aware deep model, reconstructing a 3D face from a single 2D image. The goal is to minimize the negative influence of pose and expression variations and to obtain a face representation which is suited for robust age estimation. The proposed model also learns the changes in fa-

[†] The first two authors contributed equally.

cial appearance (2D image) through an appearance subnet. These subnets (2D and 3D) are trained to jointly optimize the 3D reconstruction and age estimation.

The main contributions of this paper are as follows: (1) To the best of our knowledge, we are the first to exploit 3D face reconstruction and 2D appearance features to jointly model pose and expression robust age estimation through multi-task learning. (2) The proposed multi-task learning model for age estimation achieves state-of-the-art accuracy on the Wiki database, as well as on cross-dataset experiments using UTK and AgeDB.

2. Related Work

Age estimation. Until recently, the predominant methods for performing age estimation are based on handcrafted features, focusing on wrinkles, skin texture and 2D shapes such as Local Binary Patterns [30, 39], Bio-Inspired Features [13], and Gabor features [9]. However, while different hand-crafted features handle some adversarial conditions, none of them are fully robust against expressions, head pose, and illumination variations. More specifically, such approaches are quite sensitive to facial pose since it causes drastic changes in facial appearance.

Convolutional Neural Networks (CNNs) performs better than previous methods for age estimation [15, 20, 31, 34, 38, 40]. Instead of mapping a full image to a certain age, as in manifold learning, CNNs aim to automatically learn efficient age-related features. Exploiting the benefits of CNNs, [31] proposes the Deep Expectation (DEX) algorithm, an age estimator that classifies age and, for more robust predictions, refines the inference prediction with a softmax expectation.

Pose and Expression Robustness. The effect of pose and expression on face analysis tasks is well studied. For instance, to provide pose robustness in face identification, [23, 27, 28, 29] augment their data by synthesizing face images for varying head poses using statistical 3D face models. In a similar way, [1] applies expression neutralization, and [42] employs pose normalization before face identification. These approaches may help to preserve the identity related dominant face features which makes them suitable for face identification. On the other hand, reconstructed/synthesized facial images lose important high-frequency details of skin appearance such as wrinkles, which would negatively influence age estimation. Nevertheless, our model is able to simultaneously learn multiple robust features, does not require labels other than age, and it is not influenced by face smoothing on neutralized images. In [21], age and facial expressions are modeled jointly to achieve expression robustness in age estimation.

Monocular 3D Face Reconstruction. Monocular face reconstruction is the task of decomposing a face into its components (i.e. 3D facial geometry, expression, head pose, skin reflectance, and scene illumination). Computing these components for a single RGB image is an ill-posed problem. To this end, methods use statistical 3D models that represent 3D faces with a low dimensional parameter code vector [3, 4, 12, 28]. This code vector contains the encoded face components such as geometry, expression, skin reflectance, and additional parameters depending on the statistical 3D model.

Conventional 3D face reconstruction methods employ iterative optimization of an energy function. For instance, [3] optimizes the parameters by minimizing the error between the reconstructed and original face. [36] also uses an iterative approach, yet, it is designed to transfer facial expressions –in videos– between faces. In addition to being computationally expensive, energy minimization approaches have the problem of being reliant on favorable initialization because of typically non-convex functions to optimize. Deep learning methods exist using data augmentation techniques to produce results closer to ground truth fitting [11, 16]. Some other studies apply an analysis-by-synthesis approach to train the neural network using a physically plausible image formation model [35]. We base our model for extracting pose and expression features on [35] with a number of modifications further discussed in this paper. To the best of our knowledge, we are the first to use 3D face reconstruction for the age estimation problem.

3. Methodology

An overview of our method is shown in Fig. 1. Given a cropped face image $\mathbf{I}$, our AlexNet-based CNN model learns to jointly produce the age prediction $\hat{y}$ (Section 3.2) and the 2D-to-3D reconstruction parameterized in a low dimensional latent space $\mathbf{z}$ (Section 3.1).

The appearance and 3D reconstruction features are combined using multi-task learning. Two methods are explored (Section 3.3). In the hard parameter sharing approach, a single shared CNN is adopted. The optimized loss is the weighted sum of a 3D fitting loss $\mathcal{L}_{fit}$ and an age estimation class distance loss $\mathcal{L}_{dist}$. In the soft parameter sharing approach, we use two separate CNN backbones and mutually connect multiple of their layers to allow sharing. The loss is the sum of $\mathcal{L}_{fit}$ and $\mathcal{L}_{dist}$. As a backbone, we use AlexNet [18] with a removed last fully connected layer.

3.1. Monocular 2D-to-3D Face Reconstruction Subnet

The employed monocular 2D-to-3D face reconstruction (fitting) model jointly decomposes a given 2D face image into its underlying components represented in a low dimensional code vector $\mathbf{z}$: face rotation $\boldsymbol{\omega} \in \mathbb{SO}3$ and translation $\boldsymbol{\tau} \in \mathbb{R}^3$, face identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, face expression $\boldsymbol{\delta} \in \mathbb{R}^{64}$, skin reflecntance $\boldsymbol{\beta} \in \mathbb{R}^{80}$ and illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$. A

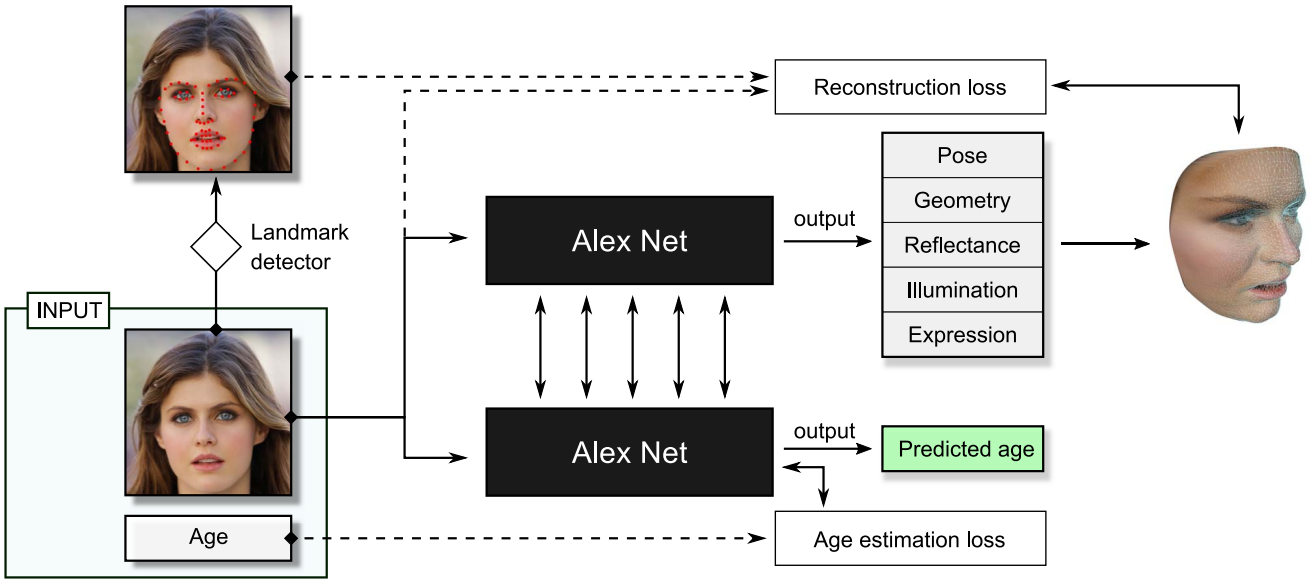


Figure 1. Our multi-task learning architecture for combining visual and 3D face reconstruction to perform robust age estimation. Two approaches are explored: (1) with hard parameter sharing - sharing the weights of a single AlexNet CNN, and (2) soft parameter sharing - two AlexNet CNNs have mutually connected layers.

fully connected layer with linear activation is added on top of the AlexNet backbone to infer $\mathbf{z}$.

$$\mathbf{z} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\tau}\} \tag{1}$$

Reflectance and geometry. The facial geometry $\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) \in \mathbb{R}^{N\times 3}$ and reflectance $\mathbf{L}(\boldsymbol{\beta}) \in \mathbb{R}^{N\times 3}$ are represented as a multilinear PCA model using the Basel Face Model 2017 [12].

$$\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) = \boldsymbol{\mu}_{geom} + \mathbf{E}_{id}[\boldsymbol{\alpha} \cdot \boldsymbol{\sigma}_{id}] + \mathbf{E}_{exp}[\boldsymbol{\delta} \cdot \boldsymbol{\sigma}_{exp}] \tag{2}$$

$$\mathbf{L}(\boldsymbol{\beta}) = \boldsymbol{\mu}_{ref} + \mathbf{E}_{ref}[\boldsymbol{\beta} \cdot \boldsymbol{\sigma}_{ref}], \tag{3}$$

where $\boldsymbol{\mu}_{geom}$, $\boldsymbol{\mu}_{ref} \in \mathbb{R}^{N\times 3}$ represent the mean neutral geometry and skin reflectance; $\mathbf{E}_{id}$, $\mathbf{E}_{ref} \in \mathbb{R}^{N\times 3\times 80}$, $\mathbf{E}_{exp} \in \mathbb{R}^{N\times 3\times 64}$ correspond to the linear bases of the PCA model together with their standard deviations $\boldsymbol{\sigma}_{id}$, $\boldsymbol{\sigma}_{ref} \in \mathbb{R}^{80}$, $\boldsymbol{\sigma}_{exp} \in \mathbb{R}^{64}$.

Camera model. We model the face transformation to the camera space by a rigid transformation consisting of rotation $\mathbf{R}(\boldsymbol{\omega}) : \mathbb{R}^3 \rightarrow \mathbb{R}^{3\times 3}$ and translation $\boldsymbol{\tau}$ together with a full perspective transformation Π to obtain vertex coordinates $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ on the camera plane.

$$\begin{aligned} \mathbf{u}, \mathbf{v} &= \{u_i, v_i\}, i \in \{1..N\} \\ &= \Pi \circ (\mathbf{R}(\boldsymbol{\omega})\mathbf{G}(\boldsymbol{\alpha}, \boldsymbol{\delta}) + \boldsymbol{\tau}) \end{aligned} \tag{4}$$

Illumination model. We model illumination using the first B = 3 bands of Spherical Harmonics [26] bases $H_b(\mathbf{n}) : \mathbb{R}^{N\times 3} \rightarrow \mathbb{R}^N$ assuming the face surface to be Lambertian with a distant illumination ignoring self-occlusion and cast-shadows. Illumination coefficients are predicted separately for the RGB channels. Vertex normals $\mathbf{n}$ are estimated using 1-ring neighborhood. Shaded colour is computed as a Hadamard product between reflectance and shading:

$$\begin{aligned} \mathbf{C}(\boldsymbol{\beta}, \mathbf{n}, \boldsymbol{\gamma}) &= \{c_i\}, i \in \{1..N\} \\ &= \mathbf{L}(\boldsymbol{\beta}) \cdot \sum_{b=1}^{B^2} \gamma_b H_b(\mathbf{n}) \end{aligned} \tag{5}$$

Fitting. The energy formulation of [35] is used to train the proposed pipeline to predict the code vector $\mathbf{z}$. Our loss consists of a landmark loss $\mathcal{L}_{lan}$, a photometric loss $\mathcal{L}_{photo}$ and a regularization term $\mathcal{L}_{reg}$ balanced using weights λ_{lan} and λ_{photo}.

$$\mathcal{L}_{fit} = \lambda_{lan}\mathcal{L}_{lan} + \lambda_{photo}\mathcal{L}_{photo} + \mathcal{L}_{reg} \tag{6}$$

Photometric loss. We use the $L_{2,1}$ loss [6] to penalize the difference between the predicted per vertex shaded colour (Eq. 5) and the ground truth colour at positions $\{\mathbf{u}, \mathbf{v}\}$. The loss is defined for a subset of vertices $\mathcal{V}$ with normals directed toward the camera screen.

$$\mathcal{L}_{photo} = \frac{1}{|\mathcal{V}|} \sum_{i\in\mathcal{V}} \|I(u_i, v_i) - c_i\|_2 \tag{7}$$

Landmark Loss. We annotated 48 landmarks with vertex indexes $k_j, j \in \{1..48\}$ on the BFM model and penalize the L_2 difference between ground truth landmarks $\mathbf{l}_j$ and

their corresponding prediction $\mathbf{p}_{k_j} = \{u_{k_j}, v_{k_j}\}$ from the 3D model.

$$\mathcal{L}_{lan} = \sum_{j=1}^{48} \left\| \mathbf{p}_{k_j} - \mathbf{l}_j \right\|_2^2 \tag{8}$$

Regularization. We regularize the model using Tikhonov regularization to enforce the model to predict faces closer to the mean.

$$\mathcal{L}_{reg} = \lambda_{alpha} \sum_{i=1}^{80} \alpha_i^2 + \lambda_{beta} \sum_{i=1}^{80} \beta_i^2 + \lambda_{delta} \sum_{i=1}^{64} \delta_i^2 \tag{9}$$

3.2. Appearance Subnet

We refer to our age estimation method, that learns visual age features, as the appearance subnet. Our appearance model is derived from [31] where the cross-entropy loss is used for resistance to outliers. Following [31], for further outlier resistance, we calculate the expectation over the softmax distribution to obtain a prediction $\hat{y}$ during the testing time:

$$\hat{y} = \sum_{i=0}^{M-m} (i+m) \cdot a_i \tag{10}$$

A fully connected layer is added with an output activation vector $\boldsymbol{a}$ on the backbone. The ground truth age is denoted by y and its one-hot encoding by $\widetilde{y}$. The minimum and maximum age that a model can predict are $m = 0$ and $M = 80$. In contrast to [31], a distance term is added to the cross-entropy loss to penalize the probability mass which is different from the correct classes. The modified loss is defined as:

$$\begin{aligned} \mathcal{L}_{dist} = \lambda_{dist} \sum_{i=0}^{M-m} a_i \cdot d(i,y) - \sum_{i=0}^{M-m} \widetilde{y}_i \cdot \log(a_i) \\ \text{where } d(i,y) = |(i+m) - y| \end{aligned} \tag{11}$$

It is assumed that each class corresponds to one year of age and that the classes are indexed in order of monotonic increase. $d(.)$ is a distance function. Absolute distance is chosen to be used in this work, as it is a natural choice to represent distance and is outlier resistant. λ_{dist} is a constant used to tune the balance between the two terms.

3.3. Multi-Task Learning

Both subnets (3D and appearance) have AlexNet as a backbone. This establishes a correspondence between the layers of the two pipelines. The features of the tasks are different. However, as they are processed by the same filter size, the features are at the same scale of detail. The combination of these features is then suitable for processing by both pipelines following the shared layer. In this paper, we attempt both hard and soft parameters sharing for multi-task learning.

Hard Parameter Sharing. A single AlexNet is shared for both tasks. The idea is that by joint training, the features are enforced to be suitable to age, and to pose and expression information. The last layer contains the refined informative features for each task. The loss is a weighted sum of both tasks with weight w:

$$\mathcal{L}_{HPS} = (1-w) \cdot \mathcal{L}_{fit} + w \cdot \mathcal{L}_{dist} \tag{12}$$

Soft Parameter Sharing. The hard parameter model forces the tasks to share all of their CNN features. This may not be optimal. Therefore, we employ a soft parameter sharing technique that can learn which layers to share. In this way, the tasks can produce independent high-level layers. For this, we use Cross-stitch Networks [24]. The approach is to have two instances of a backbone, i.e. A and B, one for each task. We choose to mark the 3D reconstruction subnet by A and the appearance subnet by B. So-called *cross-stitch* layers are then inserted in key positions in the deep network. Stitch layers take activations x_i from two layers, one from A and one from B, and blends them together as follows:

$$\begin{bmatrix} \bar{x}_A^i \\ \bar{x}_B^i \end{bmatrix} = \begin{bmatrix} \kappa_{AA}, \kappa_{AB} \\ \kappa_{BA}, \kappa_{BB} \end{bmatrix} \begin{bmatrix} x_A^i \\ x_B^i \end{bmatrix} \tag{13}$$

The κ parameters are trained together with the architecture. They are common for all activations in a pair of layers. [24] provides information about the positions for the cross-stitch layers inside AlexNet which we use after all max-pooling layers and fully connected layers. The final architecture is shown in Fig. 1.

In our implementation, Adam is chosen as an optimizer, but the κ parameters are trained separately with the Adagrad optimizer, to enforce a higher learning rate. This choice is meant to address the κ parameters receiving very small updates because of the magnitude of AlexNet activations, as noted by [24]. To avoid overfitting, we apply L_2 regularization but only on the age estimation branch, since the 3D reconstruction subnet already has its own regularization term.

4. Datasets

The training data is based on the large scale IMDB-Wiki [31] dataset. Different from other datasets, it contains in-the-wild faces with a variety of poses and expressions. Only the Wiki-Cropped subset is used, as it holds more accurate age annotations.

Figure 2. 3D reconstructions from our 3D face reconstruction model on samples from the Wiki test set. Shown are the original and the projected on them predicted 3D models.

Wiki-Cropped is cleaned by filtering out data crawled from unregulated Wikipedia sandbox and user pages, black-and-white images and photos with undetected face by the dlib face detector [17]. We keep images labeled below 80 years of age and a maximum of 600 images per age label, to balance the data distribution. Our test set (referred to as Wiki test set) consists of 10% of the cleaned data. To ensure sufficient training data, the test set is distributed as closely as possible to the training set. We alternate between 5 age groups when building a training batch to enforce label diversity. The boundaries of the groups were chosen to be the 20th, 40th, 60th and 80th percentiles of the dataset distribution.

The landmarks are extracted by the dlib face detector [17] and used for landmark loss $\mathcal{L}_{lan}$ in training.

For cross dataset evaluation, we choose the manually annotated in-the-wild AgeDB dataset [25] and the UTKFace dataset [41].

5. Experiments

The success of our approach heavily relies on the success of each subnet, therefore we first demonstrate the qualitative results of our monocular 3D face reconstruction subnet. In Fig. 2, original images and their reconstructions can be seen. The reconstructions are visually accurate even under high pose and expression variations.

5.1. Evaluating the appearance subnet for age estimation

In this experiment, we compare the performance of our appearance subnet to two other recent age estimation approaches: Deep Regression Forests [33] and SSR-Net [37]. Like the proposed appearance baseline, both models are trained on the cleaned Wiki dataset. Training of the appearance subnet is performed with a learning rate of 10^{-5}, Adam optimizer, batch size 5, step learning rate decay and L_2 regularization with weight 0.01. The λ_{dist} parameter of the loss is set to 0.2 which results in close values of the distance loss component and the Cross-Entropy component.

We report on mean absolute error (MAE) between estimated and ground-truth age in Table 1. Our appearance subnet outperforms the other methods. We use it as a baseline for further experiments.

Method	MAE
SSR-Net [37]	7.33
Deep Regression Forests [33]	13.21
Appearance Subnet (Standalone)	**5.86**

Table 1: Best MAE test score of different age estimation methods trained on the Wiki dataset. The appearance subnet used as the visual baseline in this paper outperforms the other two methods.

5.2. Joint learning of Age Estimation and 3D face reconstruction

In this section, we study the performance of the appearance subnet with a joint classification of age estimation and 3D face reconstruction. We show that the performance of age estimation increases by exploiting features learned from the monocular face reconstruction.

In this and subsequent experiments, for soft sharing, we load pre-trained Alexnet weights for the 3D face reconstruction subnet in the joint model. For other cases, we load ImageNet classification pre-trained AlexNet weights. We apply L_2 regularization with weight 10^{-5}, dropout with rate 0.7 on the final layer of age estimation. For comparison, we

used the same regularization scheme for the standalone appearance subnet and hard parameter sharing.

After tuning, the MAE score from the hard parameter sharing model (5.74 MAE) marks an age prediction improvement over the independent appearance subnet, as evident in Table 4, and shows the benefit from sharing the 3D reconstruction features.

Age estimation weight	MAE
Appearance Subnet	5.86
$w = 0.1$	5.95
$w = 0.3$	**5.74**
$w = 0.5$	5.78
$w = 0.7$	5.83
$w = 0.9$	5.89

Table 2: Best MAE test scores of the hard-parameter sharing model after training on Wiki dataset with different weights w for the age estimation loss. The weight of the 3D face reconstruction is $1 - w$. Hard parameter sharing outperformed the appearance subnet.

Soft sharing parameters	MAE
$\kappa_{AA} = 0.9, \kappa_{AB} = 0.1$	5.58
$\kappa_{AA} = 0.8, \kappa_{AB} = 0.2$	5.68
$\kappa_{AA} = 0.7, \kappa_{AB} = 0.3$	5.52
$\kappa_{AA} = 0.5, \kappa_{AB} = 0.5$	**5.47**

Table 3: Best MAE test scores from tuning soft parameter sharing model's κ parameters on the Wiki dataset.

We obtain better MAE scores with κ parameters that encourage large sharing in the soft parameter sharing model. Table 2 gives an overview of the test performance of hard parameter sharing for different choices of the loss weight w. The MAE scores are decreasing with decreasing of the weight for age estimation, which means higher sharing with 3D face reconstruction.

For soft parameter sharing, we assess different choices for the amount of sharing by κ. Our initialization follows the rules $\kappa_{BB} = \kappa_{AA}$ and $\kappa_{AB} = \kappa_{BA} = 1 - \kappa_{AA}$. We chose non-sharing ($\kappa_{AB}$ and κ_{BA}) values from the range $[0.5, 1]$ in order to follow the predetermined rules. If smaller values are chosen, the branches would just switch the CNNs they rely mostly on. The test MAE scores after training on Wiki are shown in Table 3. The results show that age estimation benefits from large sharing. Significance of the best results ($\kappa = 0.5$) is confirmed p-value $2.70 \cdot 10^{-5}$ from t-test after confirming normality with a normality test.

Having outperformed the hard parameter sharing, as shown in Table 4, the soft sharing age estimation seems to benefit from the independence of higher layers offered by the soft sharing architecture. As shown in Table 4, after 5 repeated training sessions per model, MAE score distributions are narrow and not overlapping. We can conclude that our age prediction is stable. For further experiments, we consider only the much better performing soft parameter sharing model.

5.3. Analyzing the age estimation improvements by pose and expression

In this experiment, we evaluate the performance of the proposed soft parameter sharing model on varying pose and expression and compare it to the standalone appearance subnet. Each image in the test set is associated with expression (i.e. using predicted expression parameters) and head pose (i.e. using predicted head pose angle). We obtain an expression extremeness metric from the Euclidean norm of the expression vector $\boldsymbol{\delta}$. Our pose extremeness metric is based on the maximum of the exponential coordinates that parameterize a rotation $\boldsymbol{\omega} \in \mathbb{SO}3$. Separately for each of these metrics, we cluster the images into equally balanced groups. For each of the groups, the mean of the MAE differences over all the images falling in the group is computed and plotted to analyze the impact of our model on each challenge.

Fig. 3 (a) visualizes the expression strength of each group by showing a number of samples. Fig. 3 (b) shows how the MAE changes throughout the groups. The appearance subnet's MAE increases with increasing expressiveness whereas the soft sharing method always scores better and performs similarly for the different ranges of expressiveness. Therefore, our proposed algorithm is more robust to expression variations. It improves over the appearance subnet the most on the most extreme expressions group (improvement is up to 1.8 MAE).

Fig. 4 (a) visualizes the head poses contained in each group. Looking at fig. 4 (b), the appearance subnet is much more likely to fail on more extreme poses than the soft sharing model. Moreover, the trend is that increasing the head pose extremeness leads to higher improvement over the appearance subnet. Therefore, the proposed algorithm is more robust to head pose variations. Notably, the improvement is highest for the most extreme head pose variations (1.4 MAE).

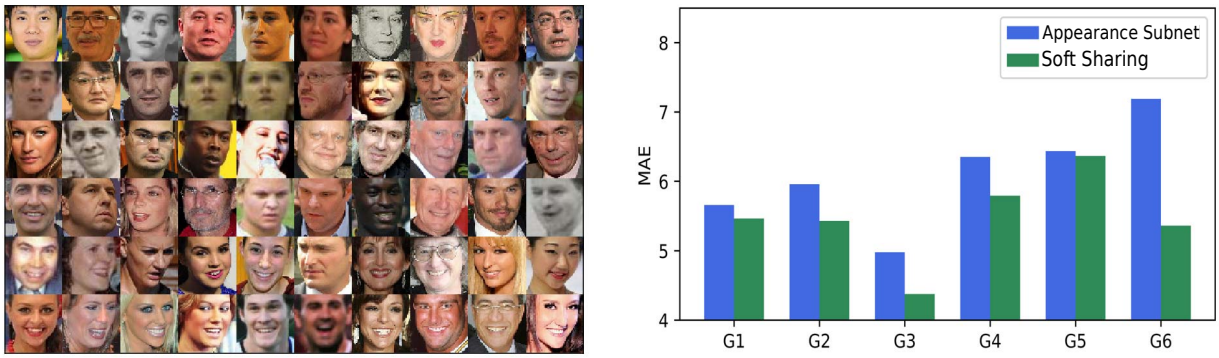
Fig. 5 further demonstrates the pose and expression robustness of the soft parameter sharing model by visually showing its superior predictions to the appearance subnet on the extreme pose and expression examples.

5.4. Cross-dataset evaluation

To show if the results extend beyond the dataset used for training, evaluation is done on UTKFace and AgeDB. The expectation is to obtain MAE scores with soft param-

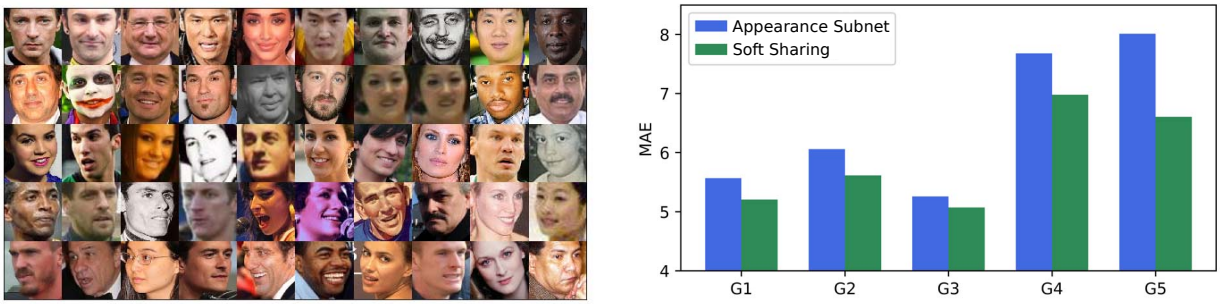
Method	Mean ± Std
Appearance Subnet (Standalone)	5.86 ± 0.04
Proposed: HPS (Appearance + 3D Reconstruction Subnets)	5.74 ± 0.04
Proposed: SPS (Appearance + 3D Reconstruction Subnets)	**5.47** ± 0.03

Table 4: Mean best MAE test scores and deviations calculated from 5 training sessions on Wiki dataset of the appearance subnet, the proposed soft parameter sharing (SPS) and hard parameter sharing (HPS), combining the Appearance subnet with the 3D Face reconstruction subnet.



(a) (b)

Figure 3. (a) Samples from the expression intensity groups. Each row contains samples from one group. Groups are sorted by increasing metric from top to bottom; (b) The MAE for the soft sharing model and the standalone appearance subnet over the expression extremeness groups. The expression extremeness metric is increasing in the groups from left to right. The results show that the proposed model shows robustness to expression in contrast with the appearance subnet.



(a) (b)

Figure 4. (a) Samples from the rotation groups. Each row contains samples from one group. Groups are sorted by increasing metric from top to bottom. (b) The MAE measures for soft parameter sharing model and appearance subnet over the rotation extremeness groups. The rotation extremeness metric is increasing in the groups from left to right. The results show the robustness of the proposed algorithm to head pose in contrast to the appearance subnet.

Method	UTKFace	AgeDB
Appearance Subnet (Standalone)	9.73	10.27
Proposed: SPS (Appearance + 3D Reconstruction Subnets)	9.54	10.01
t-test p-value	$3.22 \cdot 10^{-9}$	$1.78 \cdot 10^{-8}$

Table 5: MAE scores from cross dataset evaluation of the appearance subnet and the soft parameter sharing model (SPS).

Figure 5. Age predictions of the Appearance subnet (denoted as Baseline) and the soft parameter sharing model (denoted as SS) on non-frontal and non-neutral faces from the Wiki test set. The improvement of age prediction under extreme pose and expression conditions is visible.

eter sharing, which are significantly lower than the MAE scores of the standalone appearance subnet. It is not common practice to provide results on cross-dataset evaluation for age estimation since the performance may largely deteriorate. Results are shown in Table 5. It can be derived the improvements are significant. It shows the generalizable power of the soft sharing multi-task learning model.

6. Conclusion

In this paper, we have shown that 3D reconstruction features can significantly improve the age estimation performance when jointly learned with appearance features. Our method takes a single 2D image and derives 3D reconstruction features as a new source of pose and facial expression robustness by employing a monocular 3D face reconstruction model. After evaluation, our method has shown to be consistently more robust across variation and improved over the baseline the most with extreme head poses (1.4 MAE) and intensive expressions (1.82 MAE).

References

[1] B. Amberg, R. Knothe, and T. Vetter. Expression invariant 3d face recognition with a morphable model. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–6, 2008.

[2] R. Angulu, J. R. Tapamo, and A. O. Adewumi. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, 2018.

[3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999.

[4] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2-4):233–254, 2018.

[5] C. Ding and D. Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on Intelligent Systems and Technology*, 7(3):37, 2016.

[6] C. Ding, D. Zhou, X. He, and H. Zha. R1-pca: Rotational invariant l1-norm principal component analysis for robust subspace factorization. In *International Conference on Machine learning*, pages 281–288, 2006.

[7] L. G. Farkas. *Anthropometry of the Head and Face*. Raven Pr, 1994.

[8] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2018.

[9] F. Gao and H. Ai. Face age classification on consumer images with gabor feature and fuzzy lda method. In *International Conference on Biometrics*, pages 132–141, 2009.

[10] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007.

[11] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[12] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open

framework. In *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 75–82, 2018.

[13] G. Guo, G. Mu, Y. Fu, and T. S. Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009.

[14] G. Guo and X. Wang. A study on human age estimation under facial expression changes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2553, 2012.

[15] Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong, and S. Yan. Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097, 2017.

[16] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. InverseFaceNet: Deep monocular inverse face rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[19] Y. H. Kwon et al. Age classification from facial images. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–767, 1994.

[20] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42, 2015.

[21] Z. Lou, F. Alnajar, J. M. Alvarez, N. Hu, and T. Gevers. Expression-invariant age estimation using structured learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 40(2):365–375, 2018.

[22] J. Lu and Y.-P. Tan. Ordinary preserving manifold analysis for human age and head pose estimation. *IEEE Trans. on Human-Machine Systems*, 43(2):249–258, 2013.

[23] I. Masi, F.-J. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, et al. Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41(2):379–393, 2019.

[24] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3994–4003, 2016.

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

[26] C. Müller. Spherical harmonics, volume 17 of lecture notes in mathematics, 1966.

[27] T. Napoléon and A. Alfalou. Pose invariant face recognition: 3d model from single photo. *Optics and Lasers in Engineering*, 89:150–161, 2017.

[28] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 296–301, 2009.

[29] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *IEEE International Conference on Computer Vision*, pages 1623–1632, 2017.

[30] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[31] R. Rothe, R. Timofte, and L. 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.

[32] O. Rudovic, M. Pantic, and I. Patras. Coupled gaussian processes for pose-invariant facial expression recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(6):1357–1369, 2013.

[33] W. Shen, Y. Guo, Y. Wang, K. Zhao, B. Wang, and A. L. Yuille. Deep regression forests for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2304–2313, 2018.

[34] L. Sun, S. Qiu, Q. Li, H. Liu, and M. Zhou. Age estimation via pose-invariant 3d face alignment feature in 3 streams of cnn. In *Pacific Rim Conference on Multimedia*, pages 172–183, 2017.

[35] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision*, volume 2, page 5, 2017.

[36] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[37] T.-Y. Yang, Y.-H. Huang, Y.-Y. Lin, P.-C. Hsiu, and Y.-Y. Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *International Joint Conference on Artificial Intelligence*, pages 1078–1084, 2018.

[38] X. Yang, B.-B. Gao, C. Xing, Z.-W. Huo, X.-S. Wei, Y. Zhou, J. Wu, and X. Geng. Deep label distribution learning for apparent age estimation. In *IEEE International Conference on Computer Vision Workshops*, pages 102–108, 2015.

[39] Z. Yang and H. Ai. Demographic classification with local binary patterns. In *International Conference on Biometrics*, pages 464–473, 2007.

[40] K. Zhang, C. Gao, L. Guo, M. Sun, X. Yuan, T. X. Han, Z. Zhao, and B. Li. Age group and gender estimation in the wild with deep ror architecture. *IEEE Access*, 5:22492–22503, 2017.

[41] S. Y. Zhang, Zhifei and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[42] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.