University of Amsterdam

MSc Artificial Intelligence
Master Thesis

Aiding Age Estimation with 3D Facial Geometry

by
Nedko Dimitrov Savov
11404345

September 28, 2018

36 EC
March 26 - September 7, 2018

Supervisor:
Dr Sezer Karaoglu

Assessor:
Prof. Dr Theo Gevers

Informatics Institute
University of Amsterdam

Abstract

In this work, we present deep network architectures that aim to improve age prediction by making use of 3D facial geometry. The exploited 3D geometry features were learned by a model which predicts a 3D face model from a 2D image. In addition, we propose a novel class distance loss for age estimation that penalizes high probability of distanced classes from the ground truth one. This allows age estimation to learn age-related features for a small interval around the ground truth age. We show that monocular 3D face reconstruction implicitly learns 3D facial geometry features that are age representative. Jointly learning age and 3D facial geometry lead to a significant improvement in the accuracy of age estimation. The sources of improvement are identified to be the features from monocular face reconstruction and the capability of multi-task learning. The proposed model helps to give better predictions the most for extreme poses and intensive expression.

Acknowledgments

Foremost, I would like to express my deep gratitude to my supervisor Sezer Karaoglu for always being there to patiently answer my questions and to help guide me through the thorny path of this research. I want to thank Hamdi Dibeklioglu for the regular meetings, where he provided me with invaluable advice and insights about my work. I would like to thank Theo Gevers for taking part in my defence committee. Special thanks to Minh Ngo for his work, which was used as an initial codebase for this thesis.

In the past two years, I have greatly benefited from the collaborations with my fellow students in the course of this Master's program. From them I have gained a lot of domain knowledge, team work and research skills which I came to use in this research. This thesis would not be possible without the strong support from my mother, my father and my sister, who never hesitated to provide me with resources and advice.

Contents

Chapter 1

Introduction

Human faces are a source of important information about a person's background and mind state. They encode emotion, intent, ethnicity, identity, gender, and age. Two sets of problems have been recognized regarding the information that a face contains. One deals with extracting the encoded information and the other - with generating faces with certain characteristics. Among other uses, automated systems that are aware of a person's identity can track people in a large audience [Turk and Pentland, 1999] or make a decision if a person can have access to a certain location [Coffin and Ingram, 1999]; robots that are able to track the user's emotional response can behave accordingly [Liu et al., 2017].

One of the tasks for extracting information from a face is age estimation. Its goal is to predict the age of an individual by a picture of the person's face. Age estimation has many potential uses. In management, it can be employed for getting to know what age groups are interested in what kind of products, services, and entertainment by monitoring the clients. Currently, when refugees without sufficient documentation arrive, advanced dental age estimation is performed [Sykes et al., 2017]. It would be helpful to have a fast automatic estimation from an image instead. Vending machines for tobacco and alcohol must also be able to tell if a client is old enough. Age can serve as a soft biometric to improve upon the primary biometrics. Other tasks, like face recognition, are affected by the progressing of age and can adopt age estimation features or networks in their architectures [Zheng et al., 2017].

Age estimation is a challenging task. The reasons for that are the properties of aging, addressed by[Angulu et al., 2018, Geng et al., 2007]. It is irreversible and uncontrollable, which makes collecting well-distributed data problematic. Also, due to genetics, environment, and lifestyle, aging can go very differently for different people. Being a temporal process, the current state of appearance affects future appearances but not the past ones. To capture the aging pattern of one individual many images have to be available over small intervals of time. As aging is a slow process, collecting complete datasets of many people from childhood to old age is a complex task.

Until not long ago, a large part of predominant methods for performing age estimation was based on handcrafted features, sensitive to wrinkles, skin texture and 2D shapes - Local Binary Patterns [Phillips et al., 2000, Yang and Ai, 2007],

Bio-Inspired Features [Guo et al., 2009], Gabor features [Gao and Ai, 2009]. In these solutions, the accuracy of age prediction is capped to the capabilities of the designed features. It would be more helpful to learn and develop features automatically to do better at prediction. Another set of solutions were based on biological theories. The distances between predefined face landmarks on a frontal face were used [Kwon et al., 1994, Farkas, 1994] and the idea that aging is a process that is different for different individuals [Geng et al., 2007]. The disadvantages are respectively sensitivity to face pose and requiring more complete data. Manifold learning was also employed to find a low dimensional representation of age based on images [Fu et al., 2007], but the manifold mapping functions were too simple to represent well the face-age relationship.

In the recent years, an unprecedented increase of development and improvement of deep learning models for a wide variety of tasks has been observed. With large enough datasets of labeled faces becoming available, deep learning also proved very effective with processing and generation of digital face imagery. Deep learning algorithms were found to be a good fit for age estimation as it can generalize over incomplete data [Zhang et al., 2016]. Convolutional neural networks (CNNs) gave higher accuracy than before for age prediction [Yang et al., 2015, Hu et al., 2017, Levi and Hassner, 2015, Zhang et al., 2017, Rothe et al., 2018, Sun et al., 2017]. Instead of mapping a full image to a certain age, as in manifold learning, CNNs reason by automatically learning efficient age-related features inside their structure, derived from the input 2D image. The automatic learning removes the need to manually represent human domain knowledge in the solution. Because of the convolutions, the features are local to parts of the image.

The development of a face can be divided into two stages: from birth to adulthood and from adulthood to old age [Angulu et al., 2018]. Fig. 1.1 shows some examples of aging faces. In the first stage, changes are very dynamic and to the largest extent related to geometry. A baby's face covers a small, roundish area, the eyes are relatively large, the profile is more concave than an adult face, with a protuberant forehead. The nasal bridge is lower and the skin is soft and smooth [Alley, 1988]. With age progressing, the size of the head increases and the bone structure continues developing. The face is extending upwards, making the forehead slope back while the chin becomes more protruding. With the expansion of the face cheeks, eyes, nose, and mouth also expand to cover the new area.

In craniofacial development theory, the face extension is modeled by the cardioid strain transformation [Todd et al., 1980], shown in Fig. 1.2. It describes facial growth by transforming a circle. This transformation was claimed to be self-sufficient for describing age in young faces as perceived by people. However, it is certainly not the only characteristic for aging.

The second stage is more related to skin appearance changes - wrinkles, blemishes, sagging [Angulu et al., 2018]. The skin properties can be well captured from age estimation making use of only a 2D image. Many research works are focused on making better use of wrinkle and skin texture features specifically [Choi et al., 2011, Hayashi et al., 2002, Ng, 2015]. In this stage of development, fat is depositing in different areas of the face while disappearing from others, which also changes the geometry in a subtle way. The corners of the mouth move to form a slight frown.

Figure 1.1: Examples of age progression patterns from the FGNET dataset. Source: [Lanitis, 2008]

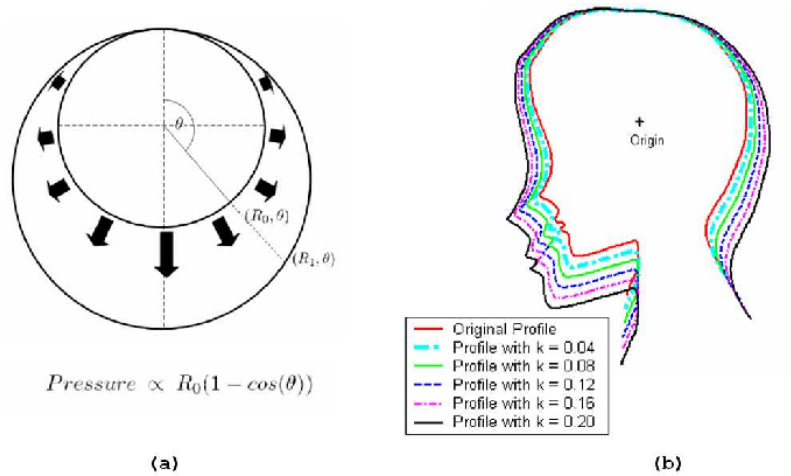


Figure 1.2: (a) visualization of the cardioid strain transformation (b) the craniofacial development modeled by the cardioid strain - k is a growth related parameter. Source: [Ramanathan and Chellappa, 2006]

The overall shape of the face also changes, especially in women. [Zhuang et al., 2010] compare a group of people older than 45 years old with an age group between 18 and 29 years old. The comparison is made with the principal components calculated with PCA on the two age groups. Principal components correspond to facial shape properties, derived from distances between facial landmarks. They report that older people have significantly larger faces, with longer narrower features.

It can be noticed from the age characteristics that were discussed above that 3D facial shape changes a lot with aging. Additionally, by having 3D face information, that is independent of extremeness of expression and pose, higher age accuracy can be expected with non-frontal poses and strong expressions. Classic age estimation from a 2D image has to learn how a face pose and expression can change in images without any given cues and by having a large and diverse dataset. However, age estimation can benefit from the expression and pose invariance of the 3D facial geometry by handling the concepts of moving in 3D space and changing the facial expression from the method used to obtain the 3D face information to handle. Then age estimation can focus on age informative features. Despite the advantages, in current age estimation solutions, 3D information is not included. [Sun et al., 2017]

uses 3D geometry only to align properly facial landmarks, which were a source of information for their age estimation. [Booth et al., 2018] prove that their framework for representing a 3D face is able to capture age. However, they use a very coarse age classification into 4 age groups and don't explore if the 3D geometry is able to help standard 2D age estimation. Our approach is to provide a representation of the 3D facial geometry to age estimation and explore to what extent it can benefit from the 3D information. It is also important to make this 3D facial geometry information supplemental to the features learned given the 2D image. In this way, the model can learn to consider wrinkles and skin texture for the age prediction - features especially important for the more mature age groups, as discussed previously.

One way to obtain 3D face information is by employing a 3D scanner - a device used to capture depth information. However, as nowadays standard cameras are by default incorporated in mobile devices, they are much more widely available than 3D scanning equipment. As a result, there is limited 3D scan data available for training and the application of a model working with 3D scans is also limited to 3D scanners being available. Therefore, instead of working with 3D scans, we choose to use a model that obtains 3D facial geometry from a 2D image. In this way, we can still use a single image as an input. Monocular 3D face reconstruction is a task that aims at generating a 3D model of a face given a single image. It has been developing in an effort to create virtual environment avatars [Bouaziz et al., 2013] and to advance video editing - e.g to be used for face replacement and facial reenactment [Thies et al., 2016, Tewari et al., 2017], without the costly effort of a 3D artist.

Approaches, tackling this problem, use statistical 3D models that represent 3D faces with a low dimensional parameter code vector [Blanz and Vetter, 1999, Paysan et al., 2009, Booth et al., 2018, Gerig et al., 2018]. This code vector contains the encoded facial shape, skin texture and additional parameters depending on the statistical 3D model - e.g. illumination and identity. Statistical 3D models achieve biological correctness by modeling variation of the positions of a large set of points on a face. Thus, they can also capture biological attributes, including age, weight, gender, nose shape, distinctiveness, etc [Paysan et al., 2009, Booth et al., 2018]. Monocular 3D face reconstruction learns the code vectors that lead to the most likely-looking 3D reconstruction of an input image. In this thesis, we show that monocular 3D face reconstruction is able to capture age information in the parameter code vector. It would be helpful to employ the implicitly learned age information, contained in the face shape features, for age estimation.

The optimization can be done iteratively [Blanz and Vetter, 1999, Vlasic et al., 2005, Thies et al., 2016] or with a convolutional neural network [Sela et al., 2017, Wang et al., 2017, Kemelmacher-Shlizerman and Basri, 2011, Tewari et al., 2017]. Having solutions with convolutional neural networks for both age estimation and monocular face reconstruction, we can combine their features with multi-task learning. In this way 2D age features, encoding skin texture, and wrinkles can be combined with 3D facial geometry features, containing information about the already discussed 3D facial shape characteristics of age progression. To the best of our knowledge, this has not been attempted before.

Regarding the multi-task learning, we use a soft parameter sharing method [Misra et al., 2016]. This means that the model is able to learn which part of

the architecture to share and to what extent. A significant improvement of the age predictions compared to the baseline was observed. Cross-dataset evaluation is performed to confirm these better scores in multiple datasets and further experiments are performed to find the source and conditions for improvement. The facial geometry features are shown to be a very good choice of initialization for regular age estimation.

1.1 Research Questions

The subject of this research is to find out if 3D facial shape features, generated from monocular face reconstruction algorithm, can be helpful to age estimation by combining them with the features from a traditional age estimation approach. Thus, the research questions to be answered are:

1. **Can monocular 3D face reconstruction produce 3D facial geometry that is age discriminative?**
2. **Can combined traditional age estimation with 3D facial geometry lead to improved age predictions?**

1.2 Contributions

The main contributions of this thesis are:

- We are the first to show that 3D face geometry learned from a single image is age informative.
- We are the first proposing to combine traditional age estimation and 3D facial geometry through multi-task learning.
- We propose a new loss for introducing distance in classification tasks.
- We perform extensive analysis of where the improvement of age prediction by 3D facial geometry is focused.
- We show the benefit of using 3D facial geometry features as a pretraining for age estimation.

1.3 Organization

This thesis starts by giving a theoretical background in Chapter 2. The described works are related to monocular face reconstruction, statistical 3D models representing a face, age estimation, and multi-task learning. Having all the necessary knowledge, in Chapter 3 the choices for the monocular face reconstruction pipeline and the age estimation models are discussed, together with the age prediction losses that are considered and the evaluation metric. Also, the multi-task learning in the context of this work is discussed in detail. In Chapter 4 the datasets

that were used are described, with the details about their cleaning and building balanced batches for training. Then, the performed experiments' setup and results are discussed. Chapter 5 contains the conclusions reached, the identified problems and challenges, and gives ideas for future work on the topic.

Chapter 2

Background

2.1 Age Estimation

Age estimation can be trained to predict real or apparent age. In this work, we focus on real age estimation, as we do claim that the used 3D facial geometry features are age-related and not that they are directly related to human perception. Also, to make the problem more bearable, it is not uncommon to simplify it to a classification into age groups. Current solutions attempt to deal with up to the year predictions and to be relevant, the age estimation in this thesis has the same granularity.

Anthropometric models consider the distances between facial points as a source of information for age [Kwon et al., 1994]. [Farkas, 1994] defines those distances based on 57 landmarks on the face. Anthropometric models are discriminative only in young age ranges, as the distances change mostly in this time period. They consider the distances between the landmarks on a 2D image. Taking only this information, however, leads to the problem of pose sensitive results, as the metrics are available for a frontal face.

Active Appearance Models [Cootes et al., 2001] were then developed to consider the appearance of the face (texture) in addition to its geometry. Both were modeled statistically with PCA, as will be explained in Section 2.2.2. The result is two low dimensional vectors, representing face geometry and appearance. However, shape here is based purely on information from landmarks. Age can be then estimated on the shape and texture vectors representing an input image [Lanitis et al., 2002]. Linear embeddings of interpretable parameters containing age information is also a centerpiece of this thesis. However, our approach makes use of lower level features used to estimate parameters for a much more advanced statistical model and it is the additional source of information to a standard deep learning age estimation. [Eidinger et al., 2014] deals with age estimation from in the wild images by adopting a dropout support vector machine.

[Fu et al., 2007] introduced manifold learning for age estimation. They work with an image space I, represented by a set of aligned face images, and a ground truth labels set A. It is aimed to find an approximation Y of a low-dimensional manifold M in the image space, informative of age-face relationship. For the manifold estimation,

they test four different linear methods separately. Having the manifold, the goal is to find a function f, such that $f(M) = age$. Since an approximation Y of M is available from the limited amount of examples, an approximation $\hat{f}$ is estimated by regression, such that $\hat{f}(Y) = A$. The regression estimates a quadratic function with the least squares error. To infer the age of a new face image, it is first mapped to the estimated manifold and then ran through the learned function to obtain the prediction. However, a quadratic function is too simple to represent well the age-face relationship.

[Geng et al., 2007] focus on aging as a personalized process in their AGES algorithm. They worked with aging patterns - features extracted from an image sequence of an aging individual. They estimated a low dimensional space with aging patterns. The prediction was made by projecting the image in this space and selecting the age and aging pattern that give minimal reconstruction error. However, AGES models full face aging patterns instead of separate cues for aging. This is why it requires a dataset of individuals, each being photographed at different ages. Current deep learning algorithms can find common age cues between different faces with similar age and do not have this requirement.

Another approach is to employ handcrafted features which are discriminative for texture, shape, and wrinkles. For every image in the dataset, these features are computed and then age estimation is trained on them. Existing methods utilize LBP (Local Binary Patterns) [Phillips et al., 2000, Yang and Ai, 2007], BIF (Bio-Inspired Features) [Guo et al., 2009] and Gabor features [Gao and Ai, 2009]. However, deep learning has the power to automatically learn more efficient task-specific features and outperforms these methods.

[Levi and Hassner, 2015] and [Wang et al., 2015] develop the first CNN architectures for estimating age and gender and show that they significantly outperform previous state-of-the-art methods. Their work clearly shows the advantages of employing deep learning for age estimation. [Zhang et al., 2017] designed a deep architecture with residual blocks to deal with the overfitting problem and report improved age predictions over other CNN architectures. This thesis is focused on up to the year predictions, which is not the case with [Levi and Hassner, 2015] and [Zhang et al., 2017], where evaluation is performed on the Adience benchmark for age group classification.

[Yang et al., 2015] takes up the task of apparent age estimation by modeling the uncertainty of the labels in the final loss, preceded by a CNN. A lot of image data is available, labeled only with a date of creation but no subject age information. In this case, the amount of time between taking two photos can be informative. [Hu et al., 2017] trained an age difference estimator between two images, given this kind of data and combined it with a ground truth age estimator.

[Rothe et al., 2018] proposed the Deep Expectation (DEX) algorithm - a state-of-the-art age estimator that classifies age and refines the inference prediction with a softmax expectation. It is the winner of the LAP 2015 challenge. In DEX each class is a range of consecutive ages. There are a lot of classes with short ranges - usually, each class represents one integer age. The use of expectation is to consider all the probabilities in the softmax - the smoothing can regularize training on noisy data.

Age can be seen as a regression or classification problem. Regression methods are sensitive to outliers and classifiers have the problems that by default they do not consider distance and easily overfit. [Angulu et al., 2018] discusses that which one is better depends on the quality of the images, data distribution, and feature extraction. However, [Rothe et al., 2018] clearly shows that an advantage of classification, together with expectation over the softmax, is dealing better with noise in the more widely available data with noisy age labels.

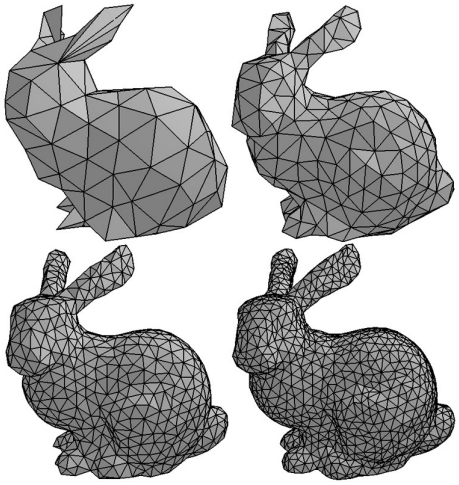
All the research addressed so far is based on working with 2D images only. There is a limited amount of work on age estimation that considers 3D facial information in some way. [Booth et al., 2018] presents age estimation ran on the vector, representing the shape of different Statistical Shape Models (SSM). SSMs are frameworks to generate 3D face models given a few low dimensional vectors, one of which represents the 3D geometry. The vectors were learned from a dataset of 3D face scans. More details on SSMs are available in Section 2.2.2. An SVM was trained to distinguish between four age groups, covering their large dataset. The achieved test accuracy on Basel Face Model 2009 was 71% and on their LSFM model - 74%. The high accuracy shows that the SSM's representation of 3D geometry carries age information. In this thesis, we perform an up-to-the-year age estimation, in contrast with the just 4 age groups in [Booth et al., 2018]. While they estimate age directly from given SSM code vectors, we take as input a face image, the 3D geometry features of which we obtain with a monocular face reconstruction solution. [Sun et al., 2017] created architectures, which directs the attention to local informative face patches, surrounding specific landmarks. To ensure pose invariance, the landmarks are projected from an estimated 3DMM. In one channel, a CNN processes the stacked patches, in another - they are processed from separate CNNs without weight sharing and in a third - the whole face is processed by another CNN. The three channels are combined with late fusion. Despite the impressive results, the architecture is quite elaborate. Also, the 3D model is only used to map landmarks more precisely and the focus of this thesis is to introduce the 3D facial geometry as a source of information for age estimation.

2.2 3D Face Models

2.2.1 3D Models

A 3D model represents an object by a set of points in a 3D space and connections between those points, which guide the forming of the surfaces of the object. Although there are different ways to define the surfaces, the used one in this work is the polygonal modeling. The appropriate neighboring points, also called *vertices*, are connected with lines, thus forming polygons. The configuration of these polygons is called a *polygon mesh*. A typically chosen polygon for the mesh is the triangle, mainly because the generation performance for triangle meshes is very high. An example of a triangle mesh is shown in Fig. 2.1.

To add a color texture to the model, each of the vertices is assigned a color - 3 values for the red, green and blue channels of the color. An arbitrary point inside a triangle in a triangle mesh can be colored by interpolation of the color of the three



(a) Triangle mesh with different levels of coarseness. Source: http://studentsrepo.um.edu.my/6408/1/NgoCheeGuan_SGJ130008.pdf

(b) A completed model with applied texture and estimated normals. Source: https://www.cc.gatech.edu/~turk/bunny/bunny.html

Figure 2.1

vertices. Sometimes it is more useful to have a separate mesh for texturing than the one for geometry - in that case, a different texture resolution is possible compared to geometry.

To model lighting, a type of function, called Bidirectional Reflectance Distribution Function (BRDF), is needed. This function describes how light is reflected at an opaque surface. BRDFs, for example, can regulate if the surface has specularities or is diffuse. BRDFs require an information about the surface at the point, at which the reflectance distribution is estimated - the perpendicular unit vector to the surface, called the *normal vector*. To include this information, the 3 coordinates of the normal vector are stored for each geometry vertex. For the surface to appear smooth and not divided into polygons, the normals are interpolated in the triangle mesh in the same way as the colors in texturing.

2.2.2 Statistical Shape Models

3D models with fixed positions of vertices, colors, and normals can represent a face. However, faces are not rigid bodies, meaning that their surface can dynamically change - for example with different facial expressions and aging. Changing identity characteristics can also be thought of as a dynamic transformation of a face. Identity and expression transformations can be modeled by statistical methods on a dataset of many different scanned faces. *Statistical Shape Models* (SSM) are approaches that all use the following stages of generation: first, all the 3D models in the dataset must be aligned by having the same vertices at the same exact anatomical position on the face; second, a multivariate distribution is used to capture variability of the aligned vertices' positions; third, a linear dimensionality reduction is performed; finally, new faces can be created by adding variance, according to the result from the previous step, to a mean face. The expressiveness of such models is demonstrated in Fig 2.2.

It has to be noted that these approaches can represent any kinds of shapes.

Early SSMs are the Active Shape Models [Cootes et al., 1995]. They were designed to represent a range of possible 2D contours of an object for the purpose of object detection at the time.

An especially popular family of SSMs are the 3D Morphable Models (3DMM), originally proposed by [Blanz and Vetter, 1999]. They can represent the shape and texture of a face. The building process of 3DMMs is next described, as it is the basis for how the selected morphable model for this thesis works.

First, it has to be ensured that the vertices in all the scans from the dataset exist and correspond to the same spot in all of the 3D models. This is important for anatomical correctness, which later gives PCA the power to recognize how those anatomical characteristics jointly change between faces. The original approach is to parameterize the surface by projecting the 3D face (3D Cartesian coordinates) onto a cylinder that surrounds it (2d cylindrical coordinates). This transforms the problem of 3D models mapping into 2D image mapping (registration), which is better studied. Then the correspondences are found with optical flow - an algorithm which computes a flow field that minimizes the differences between the two images. Originally, only 200 faces were used, making the model fragile. [Patel and Smith, 2009] and [Cosker et al., 2011] attempted to reduce the effects by the use of thin plate splines (TPS).

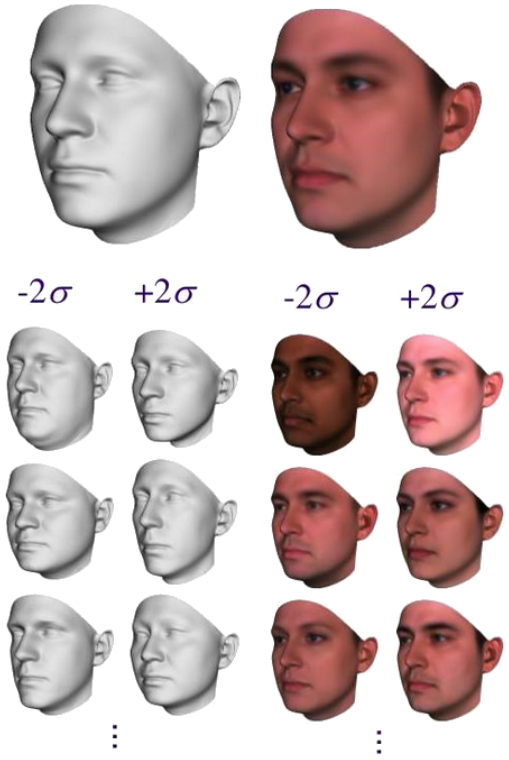


Figure 2.2: Visualization of the expressiveness of a statistical shape model. Deviating in specific directions from the mean shape and texture (on top) lead to gradually changing semantic face attributes. Source: https://www.slideserve.com/apria/3d-face-modeling

An alternative approach is to directly find correspondences in 3D space by employing versions of Iterative Closest Point (ICP) algorithm [Besl and McKay, 1992]. The Basel Face Model 09 [Paysan et al., 2009] is a 3DMM that adopts Optimal Step Nonrigid Iterative Closest Point (NICP) [Amberg et al., 2008] to align the dataset with a template.

The shape is established to be the vector of Cartesian coordinates of the aligned vertices S and the texture - the vector of RGB values T. New faces can be obtained by linear combinations of the shapes $S_{i i=1}^{m}$ and textures $T_{i i=1}^{m}$ of the aligned faces:

$$S_{model} = \sum_{i=1}^{m} a_i S_i \qquad T_{model} = \sum_{i=1}^{m} b_i T_i \qquad \sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i = 1 \tag{2.1}$$

The appearance of the new face is controlled by the parameter vectors a and b. Now a distribution over those parameters is computed, which is the likelihood of the resulting appearance and it will ensure that the generated model looks like a face. It is a multivariate normal distribution with mean - the average shape $\bar{S}$, and texture $\bar{T}$ and covariances containing the shape and texture differences of each of the models from the mean.

So far the size of the parameter vectors is the size of the dataset. A more compact representation is required. Principal Component Analysis (PCA) is a technique for dimensionality reduction that preserves most of the information (variance) of the original dataset. [Wold et al., 1987] discusses in detail how PCA works. After applying PCA with the covariances of the estimated distribution, a and b are reduced to α and β, which are now controlling the eigenvectors s_i and t_i of the covariance matrices. New face models are computed with:

$$S_{model} = \bar{S} + \sum_{i=1}^{m} \alpha_i s_i \qquad T_{model} = \bar{T} + \sum_{i=1}^{m} \beta_i t_i \qquad \sum_{i=1}^{m} a_i = \sum_{i=1}^{m} b_i = 1 \tag{2.2}$$

In the paper, they also identify linear combinations of parameters that change specific attributes of the face - weight, mouth width, femininity, more or less bony facial structure, some facial expressions. As these attributes were automatically learned without special guidance, this shows promise that age can also be potentially represented from this kind of statistically built models.

The underlying data, used for building the model, has a great impact on how expressive it is. In [Booth et al., 2018] they use a large proprietary dataset (10000 3D scans) to create LSFM - an accurate 3DMM. Their approach was to find correspondences with NICP and outlier pruning. They show that LSFM allows for two times more identity variation than the then available state-of-the-art models, with improved representation of ethnicity and age.

Additionally, expressions can be modeled with blendshapes. They are a collection of 3D face models with different expressions. They are blended together to constitute new expressions. Just as texture and shape, they are represented as a variation on top of a mean - in this case, the neutral face. This technique is done by [Bouaziz et al., 2013] to transfer expressions of real faces to a virtual avatar.

2.2.3 Gaussian Process Morphable Model

Gaussian Process Morphable Models (GPMM) [Lüthi et al., 2017] are a generalization of SSMs. In contrast with SSMs, GPMMs are not limited to model shape variation linearly - this is done by a Gaussian process.

Basel Face Model 2017 (BFM 2017) [Gerig et al., 2018] is a publicly available morphable model, the successor of the Basel Face Model 09. It employs Gaussian

processes because they allow much easier control of the face shape variation. They find a linear approximation of this model to make it tractable. Because of the superior quality and expressiveness of this morphable model, it was chosen as a major component of the monocular face reconstruction pipeline. For this reason, it is next described in detail.

In GPMMs, by applying a deformation field $u : \Gamma_R \to \mathcal{R}^d$ on a reference shape Γ_R, a new shape $\Gamma_T = \{x + u(x) | x \in \Gamma_R\}$ is derived. The assumption is that any shape can be represented in that way. The deformation field is a Gaussian process $GP(\mu, k)$, where μ is a mean function and $k : \Gamma_R \times \Gamma_R \to \mathcal{R}^{3\times3}$ - a covariance function, also known as kernel. Γ_R is normally chosen to be the mean face computed from the dataset and μ is zero in that case, as no mean deformation is needed for this choice of reference. Then, to find the best transformation, it comes down to solving the MAP problem:

$$argmax_u p(u|\Gamma_T, \Gamma_R) = argmax_u p(u)p(\Gamma_T|u, \Gamma_R) \tag{2.3}$$

The likelihood $p(\Gamma_T|u, \Gamma_R)$ is chosen to be based on the vertex-wise distance between the original shape Γ_T and the result of the deformation field u. Then, an approximation is made of u with the truncated Karhunen-Loève expansion [Huang et al., 2001]. This results in a linear combination of known basis functions ϕ_i and weights λ_i, parameterized by a vector α:

$$\tilde{u}(\alpha, x) = \mu(x) + \sum_{i=1}^{r} \alpha_i \sqrt{\lambda_i} \phi_i(x), \alpha_i \ N(0, 1) \tag{2.4}$$

The probability of the deformation $p(u)$ is now fully determined by the alpha parameters, which are normally distributed: $p(\tilde{u}) = p(\alpha) = N(0, I_{r\times r})$. Now the MAP problem can be solved.

A big advantage of Gaussian processes is that it allows controlling the modeling of the surface through the choice of the kernel k. In BFM 2017, they have a kernel that defines the deformation and makes it stronger at certain non-smooth parts of the face, a kernel for preserving face symmetry and another one for modeling expressions. A combination of those forms the kernel of the Gaussian process. BFM 2017 models separately not only shape and texture but also facial expression by employing the methods in [Amberg et al., 2008].

2.3 Monocular 3D Face Reconstruction

3D Face reconstruction is the task of creating a 3D representation of a face given sensor data, capturing the face. Monocular 3D face reconstruction deals with creating the model, given only a single RGB image. Such reconstructions are visualized in Fig. 2.3, where the resulting 3D model is aligned to the face in the original image.

Obtaining a 3D model given a single RGB image is an ill-posed problem. Solutions attempt to divide the scene into multiple components - illumination, face reflectance, albedo, shape, expression, only based on the color information from the

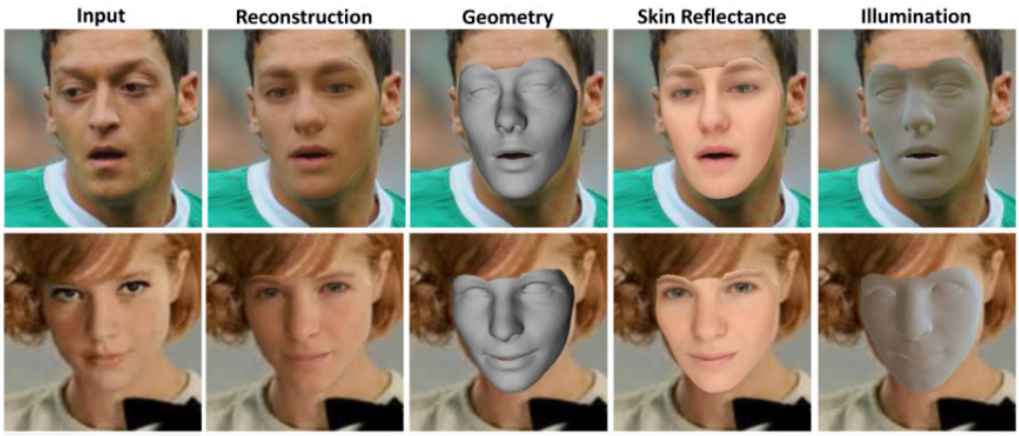


Figure 2.3: Monocular face reconstructions produced by MoFA and visualization of the separate parameters. Source: [Tewari et al., 2017]

image. Assumptions and simplifications are made when obtaining the components. For example, lighting is usually assumed to be distanced and low frequency and it is modeled by the Spherical Harmonics framework, which is limited concerning physical correctness and requires Lambertian assumption for reflectance.

When taking a picture, 3D geometry is being projected on a plane, according to the properties and position of the camera. More precisely, a point x is first mapped from a global coordinate system to the relative coordinate system of the camera by applying rotation R and translation t (camera extrinsics): $\hat{x} = Rx + t$. Then it is projected to a plane by the operator Π: $\Pi(\hat{x})$. Finally, the effect of the properties of the camera K is applied to get the point p in the image: $p = K\Pi(\hat{x})$. This whole processed is called *image formation*. When reconstructing the 3D geometry from the image, this process has to be reversed.

The general approach for solving the task at hand is to find the parameters of the separate scene components that constitute a 3D model. The loss or energy functions in these algorithms consist of multiple components. A typically used component is photometric consistency (photometric loss) - the RGB difference between the original image and the rendered aligned predicted face reconstruction. It can be a pixel-wise difference of all the pixels in both images (full photometric loss), or the difference only of the pixels, corresponding to a vertex from the predicted 3D model (vertex-wise photometric loss). Another loss component is landmark alignment - the distance between ground truth landmarks and their predicted positions.

Not long ago, the predominant methods for estimation were with iterative optimization of an energy function. Upon introducing the 3DMM, [Blanz and Vetter, 1999] optimizes a photometric consistency term by optical flow to create and align a reconstruction. [Vlasic et al., 2005] gives an early solution of the face transfer problem by parametrizing a 3D model space by identity, expression and viseme attributes with a multilinear model. [Thies et al., 2016] brings the Face2Face architecture, designed for expression transfer between faces. the source's and target's morphable models' shape, albedo and expression parameters are estimated iteratively, considering photometric consistency and distance between original and rendered landmarks. The expression parameters of the source are then assigned to the target to enable

the facial reenactment. Refer to [Zollhöfer et al., 2018] for an extensive review of iterative methods in the field.

The more recently explored approach is by adopting regression to estimate the parameters. [McDonagh et al., 2016] improve face tracking by training a regressor from synthetic data with the limitation of being actor-specific. [Sela et al., 2017] employs an image-to-image translation network to estimate a pixel-wise depth map and a mapping of pixels to 3D model vertices. Given this data, a non-rigid registration is performed iteratively to create the reconstruction. In [Wang et al., 2017] they take a different path - representations of expression, pose and 3D information are directly disentangled from the image by a deep latent variable model.

When making use of a statistical model, it is important to note that the predicted face surface is smooth - details like wrinkles are not present. [Tran et al., 2017] has a coarse shape estimation and an encoder-decoder for applying the missing details by learning bump maps. It can also deal with occlusion well. [Guo et al., 2017] employ CNN to compute the parameters for a coarse geometry model and another CNN to transfer fine geometry details onto it. Although the setup can work with a single image, they improve the reconstruction from an entire video by having a version of a coarse CNN that considers results from previous frames. [Cao et al., 2015] explores using a regressor trained on local image patches to add details on a low-resolution face surface.

[Kemelmacher-Shlizerman and Basri, 2011] uses the Shape from Shading (SfS) approach that requires the image normals, albedo and illumination to produce a reconstruction. It iteratively changes a single template to fit the image. [Sengupta et al., 2018] construct SfSNet - an end-to-end fully deep learning based model for obtaining facial shape, reflectance and illumination, again based on SfS. Features for normals, albedo, and Spherical Harmonics lighting are learned by training on labeled and unlabeled images with photometric loss.

[Tewari et al., 2017] propose MoFA - an end-to-end model, consisting of a CNN, learning the parameters of a statistical 3D model, and a fully differentiable decoder that constructs and renders the face, with the photometric loss on top. Its greatest advantage is the interpretability of the learned parameters from the CNN. This model was chosen to integrate into the architectures of this work. It is suitable because all of the features are contained in a single CNN, which can be tied to an age estimation architecture for it to make use of. Also, the interpretable compact parameters help with the analysis of the results. The reconstructions of this model depend a lot on the chosen morphable model. However, in any case, variation is modeled with PCA which leads to the limitation that the resulting shape and albedo are smoothed out, missing out on small details. However, age is greatly expressed in the overall face parts proportions and their shape changes, as discussed in Chapter 1.

The final goal of the field is to produce photorealistic complete heads with details based on an image. Therefore, research is done on realistically represent eyes, eyelids [Bérard et al., 2016] and gaze [Wang et al., 2016], mouth interior [Cao et al., 2016, Thies et al., 2016], lips [Garrido et al., 2016], hair [Chai et al., 2016] and facial hair [Beeler et al., 2012]. A stride towards this goal is developed in [Kim et al., 2018], where many face components are implemented for an expression transfer task.

Notably, head translation and rotation of the source actor are also transferred. In the current work, only the facial reconstruction is necessary for aiding age estimation.

2.4 Multi-Task Learning

A typical machine learning model solves one task by optimizing a metric through an error function. After tuning the parameters, a threshold is reached where performance does not increase anymore. However, by considering another task at the same time and optimizing its metric, a new source of information becomes available. It can help the model to capture more informative features for the first task and to make it generalize better. Optimizing multiple loss functions at the same time is called Multi-Task Learning (MTL) [Ruder, 2017]. An advantage of this approach is increased regularization capability. This is achieved by finding representation close to all the tasks, preventing the capturing of dataset-specific biases.

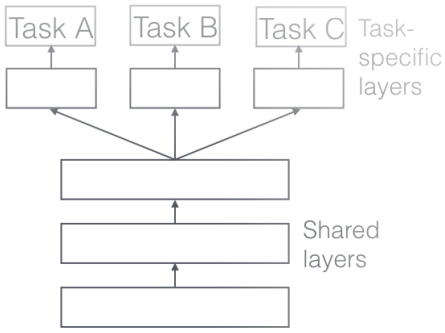
Early multi-task learning algorithms apply block-sparse regularization. The idea is that, if the parameters of all the tasks to be optimized are organized in a matrix, where each column contains all the features from one task and each row represents a feature, an assumption can be made that the important shared features between all the tasks are few and hence the matrix is sparse. To enforce the assumption [Argyriou et al., 2007, Zhang et al., 2008] use different combinations of l_1, l_2 and l_{inf} norms for most of the features - the remaining ones contain the shared information.[Negahban and Wainwright, 2008] indicates that if the features do not have much in common between tasks (e.g. because one task is not closely related to all the others), applying this method can actually harm the final performance.

The research effort was then focused for learning and representing the relatedness of the tasks and use this information as a prior. [Evgeniou et al., 2005, Jacob et al., 2009] used the variance of the features between tasks as a measure of clusteredness of the tasks. Bayesian methods were also very popular in the field. [Heskes, 2000] enforces feature similarity for different tasks by introducing a prior. In [Bakker and Heskes, 2003] they make use of a mixture of Gaussians for this similarity, based on separate Gaussian priors for each task.

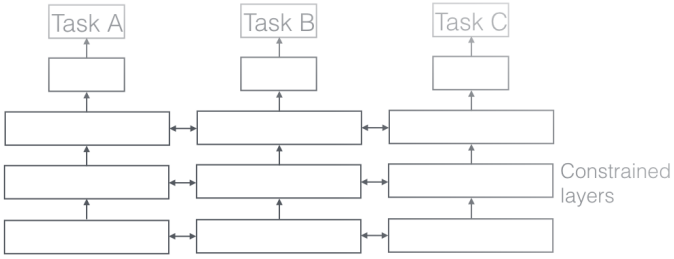
[Kumar and Daume III, 2012] base their solution on the assumption that each task is a linear combination of latent tasks, the scaling parameters of which is a sparse vector. The overlap between the sparse patterns is what is shared between two tasks. They form groups of shared tasks (but not disjoint, as in [Kang et al., 2011]). They use the sparseness and the assumption of low-dimensional representation of shared information in each group to form an optimization function for multi-task learning.

In deep learning, multi-task learning is divided into two categories. *Hard parameter sharing* sharing is the most basic kind and can be traced to [Caruna, 1993]. It is achieved by sharing a single deep network for all of the tasks, but each of the tasks has a separate subarchitecture, as shown in Fig. 2.4. As a single representation for all the tasks has to be found, overfitting is being reduced in a factor of the number of the tasks [Baxter, 1997]. In *soft parameter sharing* there are separate networks for all of the tasks and their parameters are forced to be closer together by minimizing a distance metric, as can be seen in Fig 2.4.

Deep Relationship Networks [Long and Wang, 2015] improve hard parameter



(a) Hard Parameter Sharing



(b) Soft Parameter Sharing

Figure 2.4: The two types of Multi-Task Learning. Source: [Ruder, 2017]

sharing architectures by weighting the task-specific layers with matrix priors to learn relatedness of the tasks. They have the disadvantage of being difficult to adapt for different tasks. [Lu et al., 2017] present an automatic way of building multi-task architectures given a set of tasks. The approach is to start from a fully shared until the output layers architecture of certain depth and gradually subdivide the shared layers from top to bottom. The division is determined by task similarity. The process is greedy and may find non-optimal solutions.

Multi-task learning requires multiple labels for the same input. However, especially with some less related tasks, suitable datasets are unavailable. Ubernet, developed by [Kokkinos, 2017], alleviates the problem, combining multiple datasets by carefully balancing a batch and adjusting the loss according to what labels are available. They proposed a version of SGD, which applies different learning rates to different datasets to address any label imbalance while staying memory efficient.

Hard parameter sharing requires the important but difficult choice of which layer to share. [Misra et al., 2016] introduced Cross-Stitch Networks, which automates this choice and makes it possible to share multiple layers while preserving a level of self-dependency for each of the tasks. The sharing of a pair of layers of two tasks is done by blending their features. The amount of blending is controlled by trainable parameters. Sharing layers at multiple levels allows a task to make use of features from another one and can control to what extent to rely on them and on its own learned features. This architecture was employed in this thesis, as learning insights can be gained from the blending parameters and it is shown to perform well. [Ruder12 et al., 2017] designed Sluice Networks, which harvests elements from block-sparse regularization, cross-stitch networks, and NLP hierarchical structures models (e.g. [Hashimoto et al., 2016]) to combine their benefits. Even though they

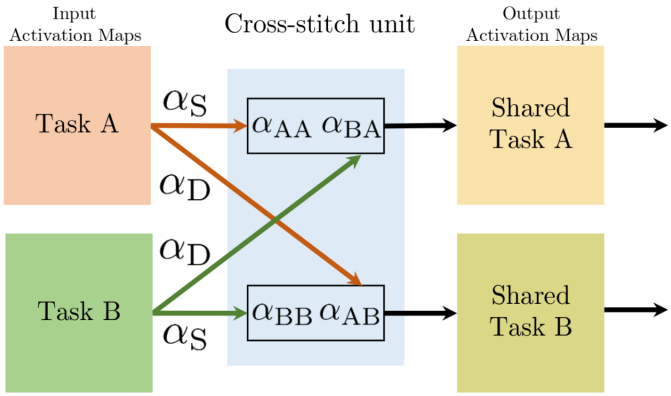


Figure 2.5: A stitch layer, as proposed by [Misra et al., 2016]. The α trainable parameters are controlling the blending of the features. Source: [Misra et al., 2016]

achieved better performance than cross-stitch networks, the improvements are small considering that the architecture is much more complex and difficult to analyze.

[Kendall et al., 2017] help to relieve the burden of determining the weights of each task in a hard parameter sharing setting by replacing the loss with a sum of task uncertainties.

As will be established, there is a relation between the tasks considered in this thesis. Multi-task learning has been long applied for benefiting from such relatedness, even if it is a weak one. [Zhang et al., 2014] do facial landmark estimation by jointly learning head pose and face attributes. [Girshick, 2015] jointly learns class and coordinates of an object. [Cheng et al., 2017] jointly predicts optical flow and object segmentation.

Chapter 3

Methodology

In this chapter, we present the detailed descriptions of the chosen age estimation, monocular reconstruction, and multi-task learning pipelines.

3.1 Monocular 3D Face Reconstruction

MoFA [Tewari et al., 2017] is a state-of-the-art model that combines the generative and regression-based approaches for 3D face reconstruction from an RGB image. The goal is to learn with a CNN a semantic code vector, containing the entire scene information. It is decomposed into facial expression $\delta \in \mathcal{R}^{64}$, shape $\alpha \in \mathcal{R}^{80}$, skin reflectance (albedo) $\beta \in \mathcal{R}^{80}$, camera rotation $T \in SO(3)$, camera translation $t \in \mathcal{R}^3$ and scene illumination $\gamma \in \mathcal{R}^9$.

$$x = (\alpha, \delta, \beta, T, t, \gamma) \tag{3.1}$$

The interpretability of these parameters is important to analyze to what extent the shape related features are age discriminative. While other deep learning methods also learn separate interpretable depth and albedo maps, this one stores the information in a low-dimensional vector, while performing end-to-end learning with the help of the differentiable parametric decoder. Another advantage is that all the learned information is stored in a single base CNN, unlike the configurations of CNNs in other approaches. This makes it possible to map the features to the ones in an age estimation CNN with multi-task learning.

The encoder CNN is AlexNet, as in the original implementation. For the shape, reflectance and expression parameters, the Basel Face Model 2017 is employed. The vertices V and reflectance at their positions R are then retrieved by:

$$V = A_s + E_s\alpha + E_e\delta \tag{3.2}$$

$$R = A_r + E_r\beta \tag{3.3}$$

, where E_s, E_e and E_r are PCA bases for shape, expression, and reflectance, and A_s and A_r are respectively the shape and reflectance means, all determined when building the BFM 2017 model.

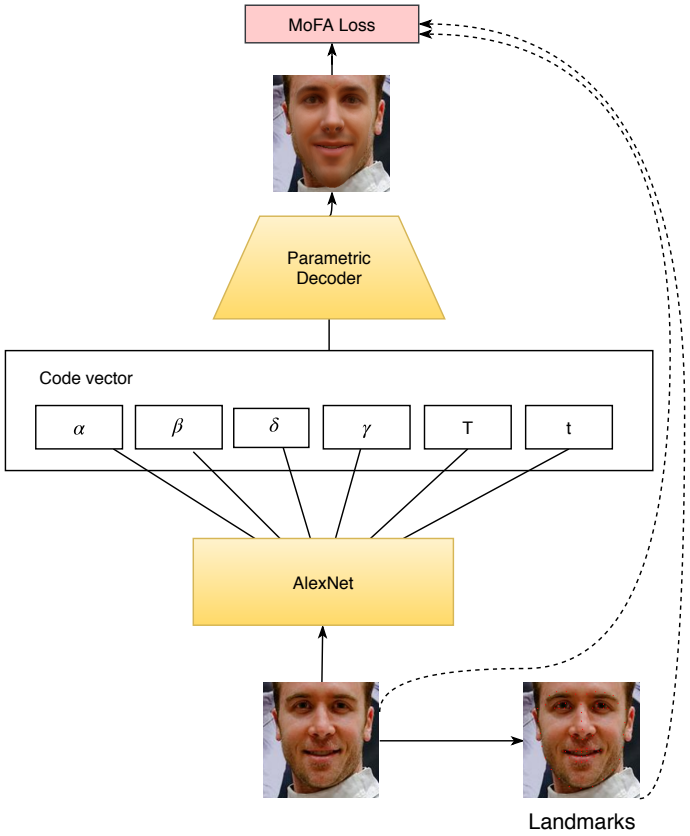


Figure 3.1: The MoFA architecture.

Given the vertices, the normal at each of them is estimated from a local 1-ring neighborhood. Spherical harmonics [Müller, 1966] is utilized for illumination modeling, where the γ parameters control colored illumination. The assumption is that illumination is always Lambertian, distant and low-frequency. Then, the radiosity at a vertice, given its surface normal n and reflectance r is computed with the $C(.)$ function:

$$C(r, n, \gamma) = r \cdot \sum_{b=1}^{B^2} \gamma_b H_b(n) \tag{3.4}$$

with $B = 3$ bands and basis functions $H_b(n)$.

For the rendering, the pinhole camera model is adopted with full perspective projection Π. The decoder performs standard image formation to create an output aligned image of the synthetic face. A vertex from V (Eq. 3.2) - V_i, is mapped to the camera plane by $u_i(x) = \Pi(T^{-1}(V_i - t))$ and its color is computed from spherical harmonics: $c_i(x) = C(R_i, Tn_i, \gamma)$.

The loss of MoFA is a linear combination of photometric loss E_{photo}, landmark distance loss E_{lan} weighted by the respective weights w_{photo}, w_{lan} and a regularization term E_{reg}:

$$E_R(x) = w_{lan}E_{lan}(x) + w_{photo}E_{photo}(x) + E_{reg}(x) \tag{3.5}$$

The regularization term penalizes faces with high deviations from the mean face. High deviations of the parameters mean that less likely faces are produced. Penalization is applied for the shape, albedo and blendshape parameters, the extent of which is controlled by the weights w_α, w_β and w_δ:

$$E_{reg}(x) = w_\alpha \sum_i \alpha_i + w_\beta \sum_i \beta_i + w_\delta \sum_i \delta_i \tag{3.6}$$

Although it is optional, for more accurate reconstructions, landmark loss is used in our implementation. The originally proposed vertex-wise photometric loss [Tewari et al., 2017], shortly described in Section 2.3, was chosen because it preserves better corner shape details than a full photometric loss.

The general architecture is shown in Fig. 3.1.

3.2 Age Estimation Visual Baseline

We refer to the age estimation pipeline, based completely on the image input without using geometry features, as the visual baseline.

The input of age estimation is processed as described in Section 4.1. This leaves us with well-distributed batches of cropped faces. They already have some light form of rotation augmentation, as no vertical alignment is performed by the adopted face detector. AlexNet architecture serves as the body of the network. Having the same base as MoFA makes it possible to utilize the optimized multi-task learning approach, described in Section 3.3.2. Considering the distribution of the available datasets, the number of output classes N was set to 81 to represent the ages in the interval $[0, 80]$.

As mentioned in Section 2.1, the state-of-the-art age estimation employs classification to outperform the existing best approaches in terms of prediction accuracy. They train their architecture on the large but noisy IMDB-Wiki dataset and show that they owe their success to the outlier resistance of classification. We choose classification in this work to benefit from this noise resistance, especially considering that our training set is a subset of IMDB-Wiki. However, with pure classification, no relationships between the classes are enforced. But in our case, there is a natural order and distance between the age classes - the further away the age prediction is from the ground truth, the worse it is. We explore losses that introduce distance between classes in classification. Being explicit about distance, the algorithm can focus on learning age descriptive features for an interval of years around the ground truth age.

In this section, we denote the final layer's activations vector with a, the ground truth age with y and its one hot encoding with $\bar{y}$. a is a single value for regression and a vector for classification - more specifically, the output of the softmax function. The minimum and maximum age that a model can predict are denoted accordingly with m and M. All the loss formulations are meant for a batch of examples. For a batch of size N, a^i and y^i are the i-th activation vector and label from the batch.

Loss functions

[Rothe et al., 2018] calculate the expectation over the softmax distribution to get a prediction $\hat{y}$, but only when testing:

$$\hat{y} = \sum_{i=0}^{M-m} (i+m) \cdot a_i \tag{3.7}$$

The smoothing helps for a better prediction. For that reason, we also employ this technique in our models.

The **cross entropy loss** is a standard classification loss, also applied by the DEX approach:

$$CE(a, y) = -\frac{1}{N} \sum_{k=0}^{N} \sum_{i=0}^{M-m} \bar{y}_i^k \cdot log(a_i^k) \tag{3.8}$$

It penalizes equally all the classes and so the distance between a predicted age and the actual age is not considered. The results in [Rothe et al., 2018] show that the DEX approach improves the MAE measure of the prediction. Even though a distance is not explicitly included, the mistakenly predicted classes are close to the real ones because the high-level features learned in the network correspond to age ranges.

However, introducing a distance measure in the loss can give the network the context to decrease this distance even further. A loss like this would have the benefits of its classification and regression components: resiliency to outliers and a properly represented relationship between the age classes.

Earth Mover's Distance (EMD). EMD is a classification loss with class distance. EMD is the minimum cost to transport the mass of one distribution to another. In our case, the source is the predicted softmax distribution and the target is the one hot encoding, which can be viewed as a distribution. The idea is to use the formulation of the problem for transporting mass from a set of supplier clusters to a set of consumer clusters. With the help of some basic properties of classification (the classes are ordered and sorted and the two distributions have equal mass), this formulation can be simplified. The full loss is shown below:

$$EMD(a, \bar{y}) = \frac{1}{N} \sum_{k=0}^{N} (\sum_{i=0}^{M-m} |CDF_i(a^k) - CDF_i(\bar{y}^k)| - \sum_{i=0}^{M-m} \bar{y}_i^k \cdot log(a_i^k)) \tag{3.9}$$

Here $CDF(.)$ is the cumulative distribution of the given vector. A squared version of EMD was shown to have a large impact with distance based classes [Hou et al., 2016]. We choose, however, the non-squared version because of the larger resistance to outliers and better stability.

Proposed **Class Distance Loss** (CDL). CDL penalizes having probability mass away from correct classes. It was constructed as an effort to introduce a version

of the expectation over the softmax for training. The penalization depends on the position of the class in respect to the ground truth label class position - the further, the higher the penalty. It is shown below:

$$D(a,y) = \frac{1}{N}\sum_{k=0}^{N}(\lambda P(a^k, y^k) + CE(a^k, \bar{y}^k)) = \tag{3.10}$$

$$\frac{1}{N}\sum_{k=0}^{N}(\lambda \sum_{i=0}^{M-m} a_i^k \cdot d(i, y^k) - \sum_{i=0}^{M-m} \bar{y}_i^k \cdot log(a_i^k)) \tag{3.11}$$

$$d(i,y) = |(i+m) - y| \tag{3.12}$$

It is assumed that each class corresponds to one year of age and that the classes are indexed in the order of monotonic increase. $d(.)$ is a distance function.

There are two separate terms in the loss that have the same aims as the terms in EMD loss. The first term is the distance penalty and the second one is cross entropy meant to guide the network towards the right class. λ is a constant that is meant to tune the balance between the two terms. Absolute distance is chosen to be used in this thesis, as it is a natural choice to represent distance and, as already discussed, outlier resistant.

The gradient of the distance penalty in respect to a single softmax output a_i is shown below:

$$\frac{\partial}{\partial a_i} P(a,y) = \frac{\partial}{\partial a_i} \sum_{i=0}^{M-m} a_i \cdot d(i,y) \tag{3.13}$$

$$= d(i,y) \tag{3.14}$$

$$= |(i+m) - y| \tag{3.15}$$

$$\tag{3.16}$$

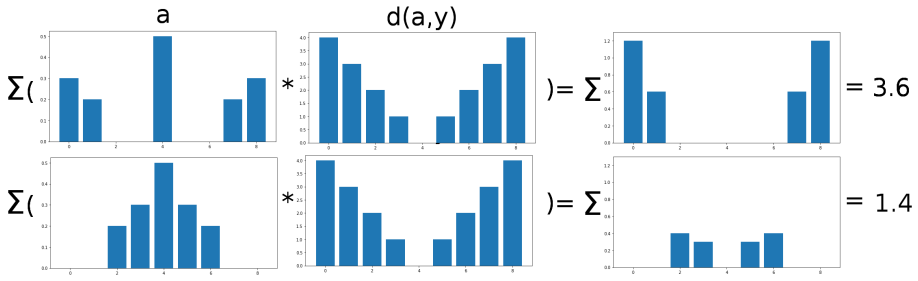


Figure 3.2: Two examples of how the distance term P of the class distance loss gives scores. In this setting, there are 9 classes and the ground truth class is the fifth. The plots in the first column are chosen softmax distributions, in the second - the penalization of each class for the current example (based on the linear distance) and in the third - the final penalty for each class.

It can be noticed that the penalty linearly increases with the distance. This is visualized on the examples provided in Fig. 3.2. The examples show that a more

spread or incorrectly skewed softmax distribution is penalized more. Meanwhile, cross entropy would give the same score to both softmax distributions.

3.3 Proposed Multi-Task Learning Models

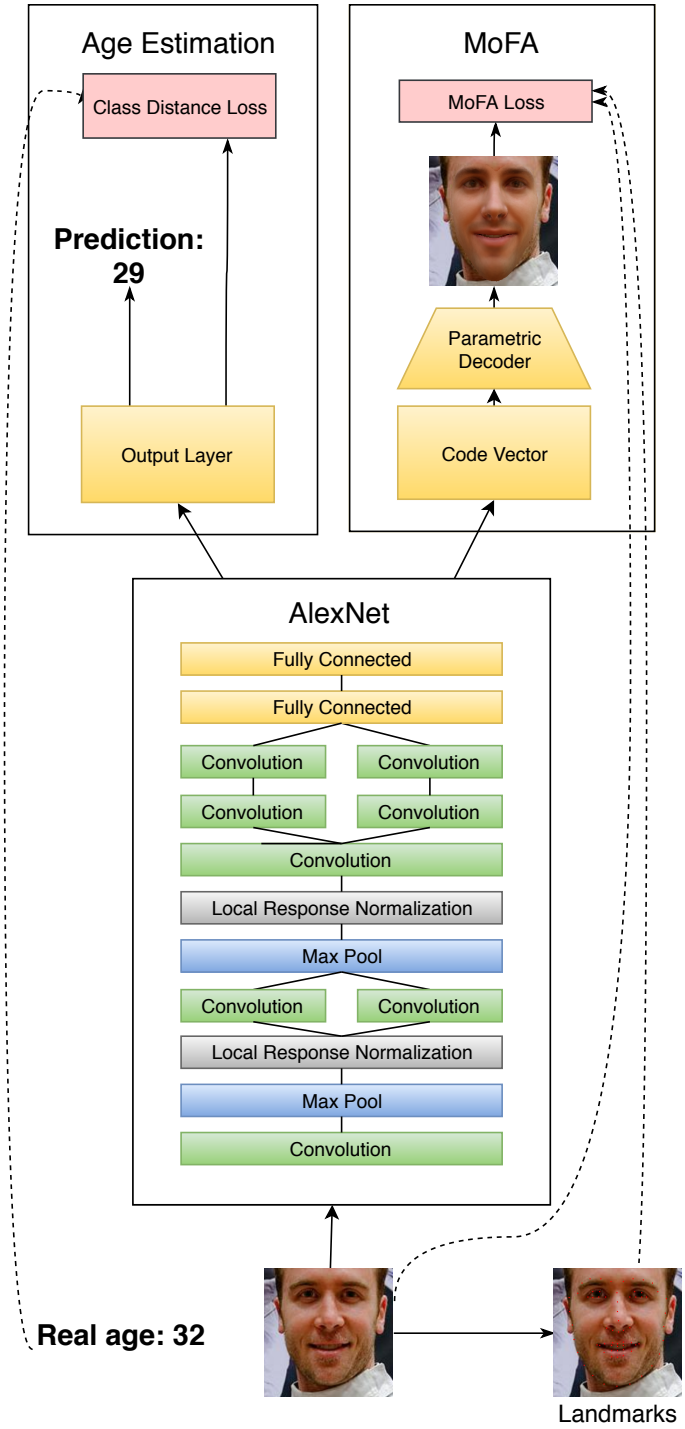


Figure 3.3: The hard parameter sharing architecture.

Both of the models have an AlexNet base CNN. This establishes a correspondence between the layers of the two pipelines. While the features between tasks will be different themselves, having been processed by the same filter sizes means they will be at the same scale of detail. The combination of these features is then suitable for processing by both pipelines following the shared layer. Also, it is especially important for many soft parameter sharing methods, because fundamentally different features may not benefit from the same prior or, in one of our cases, blending technique.

3.3.1 Hard Parameter Sharing

First, we build a hard parameter sharing model. A single AlexNet is shared for both tasks. The idea is that when jointly training, the features on each layer are enforced to be both rich to age data and to the 3D model related information, notably including shape. The last layer contains the refined informative features for each task. The architecture is shown in Fig. 3.3.

The loss is a weighted sum of both tasks with the weight w that should be tuned:

$$E_{HPS}(x, a, y) = (1 - w)E_R(x) + wCDL(a, y) \tag{3.17}$$

3.3.2 Soft Parameter Sharing

There is no guarantee that sharing all of the layers in the CNN base network is optimal. It is possible that enforcing learning relatedness on lower layers' more general features is important, but higher level layers are better left independently trained by each task to prevent competition between the tasks. One option is to try out training multiple architectures, in each the last shared layer is chosen to be different. However, in a deep network like AlexNet, this kind of experiment becomes infeasible, as it requires not only training an excessive amount of models but also tuning the weight parameters of the two tasks. This problem is solved by automatically determining which layers to share and to what extent with the use of a soft parameter sharing solution - Cross-stitch Networks [Misra et al., 2016].

The approach is to have two instances of a base neural network - A and B, one for each task. We choose to mark MoFA's CNN with A and age estimation's with 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:

$$\begin{bmatrix} \bar{x}_A^i \\ \bar{x}_B^i \end{bmatrix} = \begin{bmatrix} \alpha_{AA}, \alpha_{AB} \\ \alpha_{BA}, \alpha_{BB} \end{bmatrix} \begin{bmatrix} x_A^i \\ x_B^i \end{bmatrix} \tag{3.18}$$

The α parameters are trained together with the architecture. They are common for all activations in a pair of layers. In our implementation a cross-stitch layer calculates a single $\bar{x}$. This means for mutual sharing, one cross-stitch layer is inserted in architecture A to calculate $\bar{x}_A$ and another one in B to calculate $\bar{x}_B$.

A choice that has to be made is where to place cross-stitch layers in the architecture. [Misra et al., 2016] provides information about good positions for the new layers inside AlexNet. They place them after all max-pooling layers and fully connected layers. We make use of this choice and the final architecture is shown in Fig. 3.4.

If training is directly performed it would be noticed that the α parameters receive very small updates, as also noted by [Misra et al., 2016]. This is the case because the neurons from AlexNet give activations a few orders of magnitude lower than what the α parameters are needed to be. Therefore, a different learning rate is applied to them. In our implementation, as Adam is chosen as an optimizer, the α parameters are trained separately from the other parameters with the Adagrad optimizer, to enforce the higher learning rate.

To battle against overfitting, we apply L2 regularization but only on the age estimation branch. In this way, there is no extra unnecessary restriction for MoFA, which already has its own regularization term.

3.4 Evaluation Measure

The chosen age error evaluation metric for all models is Mean Absolute Error (MAE):

$$MAE(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N} |\hat{y}_i - y_i| \tag{3.19}$$

,where y is a prediction and $\hat{y}$ is a ground truth label. Here N is the number of samples that we average.

It is widely employed as an evaluation metric in age estimation research, as it is a natural way to represent distance. In classification, the measure typically shown is accuracy percentage of predicting the right class. However, choosing the exact class is not as important as how close the choice is to the right answer. Hence, the more informative MAE measure is shown also for classification.

When needed, comparison on MoFA performance is done with the photometric component of the loss.

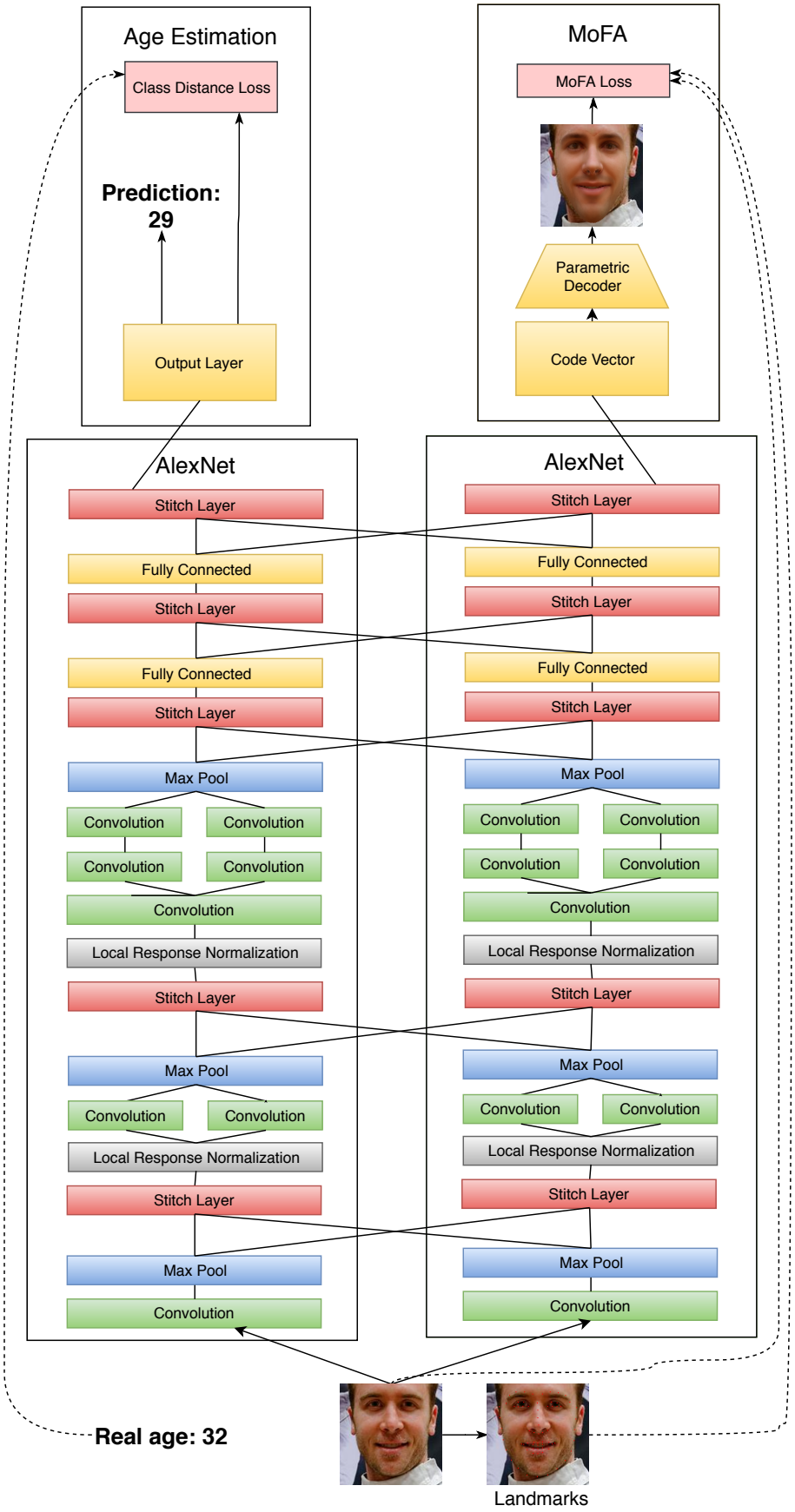


Figure 3.4: The CSN sharing architecture.

Chapter 4

Experiments

In this chapter, the contributions and the experiments needed to answer the research question are described. First, the datasets further used, their cleaning and balancing are described. Then the choices for age estimation are established and MoFA is trained and tested for implicit age information. The following experiments study the effects of multi-task learning with MoFA on age estimation. The experiments give insights about the source and the details of the results. For the implementation, the Theano and Lasagne frameworks were used. The shown MAE scores are computed per batch - the average score over all batch entries. For validation, it is the average over all batches of the validation. In all experiments, the same validation set of 960 Wiki images was used.

4.1 Data

In this section, all the datasets that are further used and their cleaning methods are presented. Then, the techniques that are applied for training and validation set balancing and training batch balancing are explained.

A most suitable dataset would be fairly large to accommodate for the large set of parameters to learn and the complexity of the monocular face reconstruction task. The labels must be as uniformly distributed as possible to prevent the model from being biased towards a specific range of label values. The number of identities should be as large as possible to prevent overfitting. However, finding a dataset that meets these requirements is a challenge. A number of fairly small datasets are available: FACES [Ebner et al., 2010], RaFD [Langner et al., 2010], 3DFE [Yin et al., 2006], Bosphorus [Savran et al., 2008]; for apparent age estimation: 2016 Looking at People CVPR Challenge - Track 3. However, their size makes them unsuitable for training in the case of this research. The AdienceFaces dataset [Eidinger et al., 2014] is larger but the ground truth age is discretized into bins, while we are interested in up to the year accuracy of the prediction. The datasets that were used are next described. Their sizes after the cleaning process are arranged in 4.1.

	Original size	Cleaned size
Wiki	62,000	15,384
UTKFace	23,095	23,095
AgeDB	16,488	11,683
FGNET	1002	990

Table 4.1: Sizes of the used datasets before and after cleaning.

4.1.1 Datasets

IMDB-WIKI

IMDB-Wiki [Rothe et al., 2018] is a large scale dataset, annotated with age and gender, consisting of two parts - IMDB Faces and Wiki Faces. IMDB Faces has around 460,000 face images collected and annotated from the actor profile pages at IMDB. Wiki Faces has 62,000 images collected and annotated from Wikipedia articles describing people.

The dataset is the largest dataset available with age annotations.

In this thesis, only the Wiki Faces subset was used, as it has much more accurate annotations than the IMDB Faces subset. Specifically, the provided set of cropped faces dataset is employed (Wiki Faces Cropped).

However, there were still a lot of incorrect labels and additional processing was required. All the people with annotated age greater than 100 were removed.

Some of the photos were crawled from sandbox pages and user pages, which contain dummy data and unregulated by the staff content - these were removed. As in black-and-white images albedo and illumination are difficult to be differentiated, all of them are removed. Detection of a black-and-white image is done by checking if it has more than 100 pixels which have a significant difference (above a specific threshold) of their channels. This check still doesn't detect some small amount of the images. Unsuccessful face detections (used for cropping the faces) and instances of a placeholder image were also removed

Even after cleaning there are still a lot of non-face images. Further cleaning was done by running the dlib face detector. In the process, however, some true faces are also deleted. After all the procedures, the size of the dataset is 15,384 images.

UTKFace

The UTKFace dataset [Zhang and Qi, 2017] consists of 23,095 images of faces, annotated with age, gender and ethnicity. It contains a wide range of ages with a fair amount of data points for each age group, which is an advantage compared to IMDB-WIKI. It is visually clean in terms of correctness of image content. The provided set of cropped faces was used and no processing was necessary. We use this dataset for cross-dataset MAE evaluation.

A disadvantage is that the human annotation is based on correcting the state-of-the-art age estimation of the image. Since the baseline and the proposed age estimation model are both evaluated on this same dataset for comparison, the pres-

ence of noise in the labels is not critical for this work.

AgeDB

AgeDB [Moschoglou et al., 2017] is an in-the-wild face dataset of 16,488 images of 568 subjects. It is manually annotated with labels accurate to the year and so it is noise free. The annotation was performed by manual searches in Google Images and saving only the ones with an explicitly mentioned year of taking in the accompanying caption. The age distribution is highest at the age interval $[30, 40]$. Data is very scarce for ages outside the interval $[15, 85]$.

The dataset was processed by removing black-and-white images and running the dlib face detector [King, 2009] to filter out images without facial landmarks detection. This reduced the dataset to 11683 images. It was used for cross-dataset evaluation for comparing the MAE measures between the age estimation baseline and a proposed model.

FG-NET

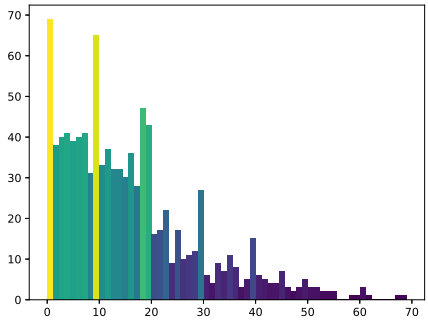


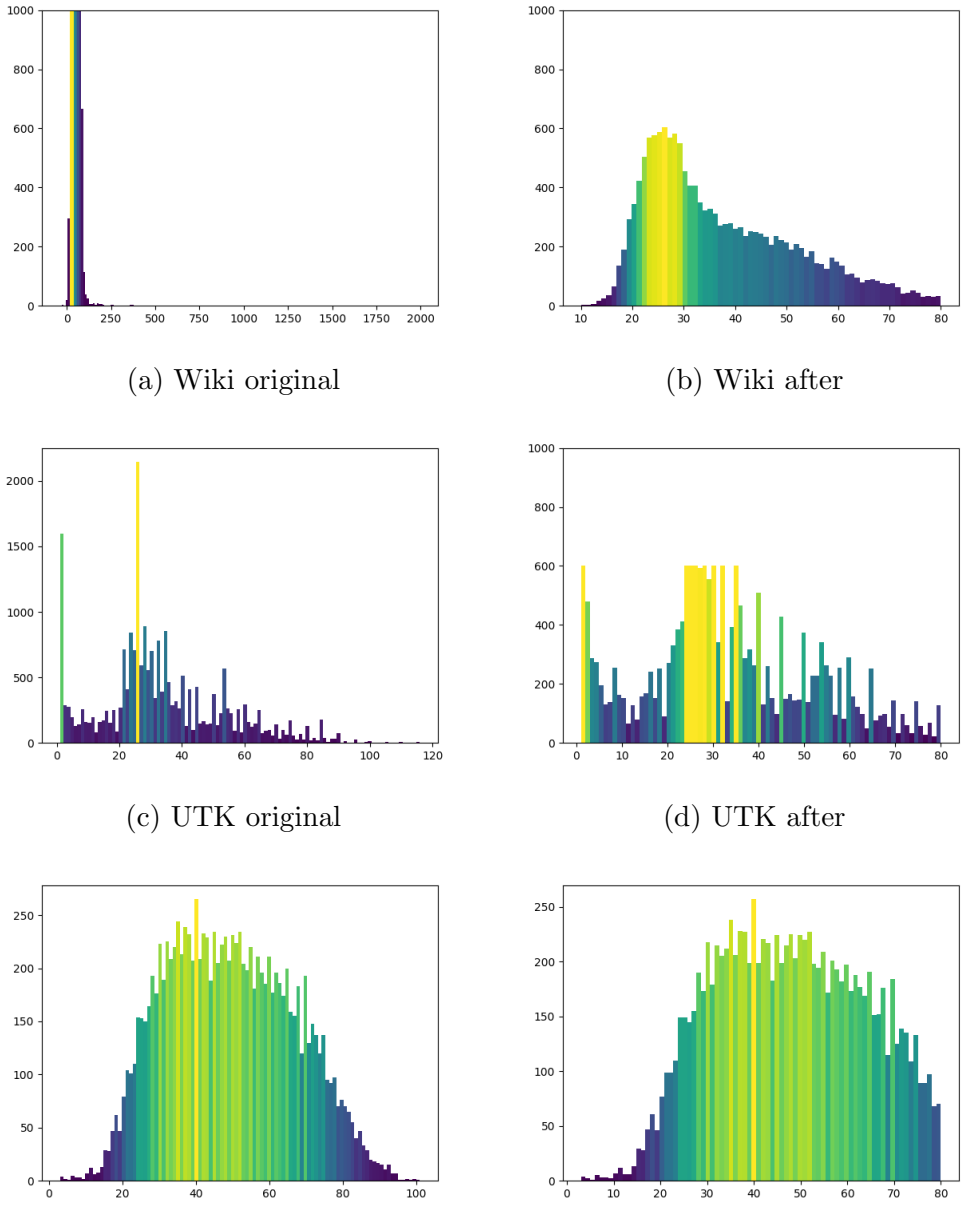
Figure 4.1: FGNET distribution.

FG-NET is a database of 1002 images of faces annotated with age. It is widely used for the evaluation and comparison of age estimation models. Its purpose in this work, as with UTKFace and AgeDB is also cross-dataset evaluation.

After extracting faces using dlib face detector [King, 2009], the dataset was reduced to 990 images. The distribution is depicted in Fig. 4.1. This dataset contains a disproportionate amount of children. This has to be kept in mind when evaluating a model trained on scarce children data.

4.1.2 Face extraction

For all the datasets, the dlib face detector [King, 2009] was adopted to detect facial landmarks. They were stored to be used in training later and their bounding boxes were stored for cropping the face and pass only this image for training. To be sure the full face is captured in the crop and that it has the right scaling for the face reconstruction pipeline, a 20% margin is added to all sides of a face box.



(a) Wiki original

(b) Wiki after

(c) UTK original

(d) UTK after

(e) AgeDB original

(f) AgeDB after

Figure 4.2: Datasets age distributions before and after cleaning and balancing.

4.1.3 Balancing

The distribution of a dataset is important to consider for the quality of model training and evaluation. First, having a very uneven distribution means that the model would be biased to predict in the range where the most mass is. It is best if the training dataset distribution is as uniform as possible.

On 4.2 the original (after processing) and the balanced distributions are shown. For all datasets, their distributions' maximum values were clipped to a threshold. This threshold was chosen, considering the tradeoff between uniformity and dataset size.

For all datasets, the maximum threshold was chosen to be 600 and the maximum age - 80.

A validation set for a model is built by taking a certain percentage of the entire data. To make sure that it is representative, it is always made sure that it is distributed as similarly as possible to the training set. Ideally, the validation set would be uniformly distributed, but if such a requirement is enforced, then there will not be enough data to train for some less represented labels in the full dataset. Instead, the same percentage of each bin in the distribution is taken for validation, which always ensures as sufficient as possible amount of training data.

Secondly, it is also important the order in which the data is taken for training. Training with too similar examples again causes a bias for the model prediction. This calls for balancing the training batches to have as diverse data in a batch as possible. [Kokkinos, 2017] perform training by placing an equal amount of examples for each label in a training batch. As age has many possible labels, five age groups were selected. It is necessary that each of the groups have about the same amount of examples. To ensure that, the boundaries of the groups were chosen to be the 20th, 40th, 60th and 80th percentiles of the dataset distribution. As the division is not exact, some of the groups can still have more data. When building a batch, an equal amount of examples is taken from each set. Drawing is done independently for each group by iterating over a sequence of data points without repetition. If there are no more elements in a group, iteration jumps to the start of the sequence. In this way, the notion of epochs is preserved in the context of a group.

4.2 Establishing baselines

Experiment 1: Age Estimation Visual Baseline

In this experiment, we test how the cross entropy and two distance-based classification losses - EMD and the proposed CDL, perform in comparison with each other. This will allow us to make reasonable choices for a baseline and for use in the models combining the 2D image with 3D face geometry.

Experimental Setup:

Wiki was chosen as the training dataset. It was noticed that validation score was oscillating too much at a later stage in training in general. This was the result of the learning rate η being too high at this later stage. For this reason, learning rate decay was implemented. The specific approach is step decay - linearly reducing the learning rate with the increase of iterations passed:

$$\eta = \eta_{init} \cdot \lambda^{\frac{t}{\delta}} \tag{4.1}$$

,where ν_{init} is the learning rate at the start of training, t is the number of iterations passed, $\lambda \in (0, 1)$ and δ are predetermined parameters. Their meaning is that the learning rate is gradually decreasing to reach one more full decay rate λ at every δ iterations. After experimenting, λ was set to 0.8 and δ - to 20000.

Another encountered problem was overfitting. To reduce this effect, L2 regularization penalty was added to the loss with weight 0.01.

The λ parameter of Class Distance Loss was set to 0.2 after carefully checking that in our case the distance loss component stays close to the value of Cross Entropy for the most part in training and in this way does not overtake it.

A model with each of the losses was trained on the cleaned Wiki dataset. The validation MAE through time is visualized in Fig. 4.3a.

Results:

Looking at the results in Fig. 4.3, the models are seen to be overfitting before reaching 10,000 iterations even with harsh regularization. This is especially visible from the rising validation loss in Fig. 4.3b. The point with the minimum score is chosen to avoid the overfitted results.

The models' MAE measures are listed in Table 4.3. The results from both EMD and the class distance loss show the benefit from explicitly implementing distance in the loss. Class distance loss outperformed EMD, although not with much. For further experiments, the chosen loss is the class distance loss for a combined approach.

learning rate (lr)	1e-5
lr decay rate (λ)	0.8
lr decay interval (δ)	20000
batch size	5
optimizer	Adam
L2 regularization weight	0.01

Table 4.2: Age estimation training parameters

Parameter tuning showed that the same learning rate and decay parameters as in the MoFA baseline work well. This is important because when the two baseline models are combined with multi-task learning techniques later, unified parameters for the entire network can be used. In this way, additional parameter tuning would not be needed to ensure better stability of the network. The chosen parameter values are listed on Table 4.2.

Loss	MAE
Cross Entropy	6.12
EMD	5.98
Class Distance	**5.86**

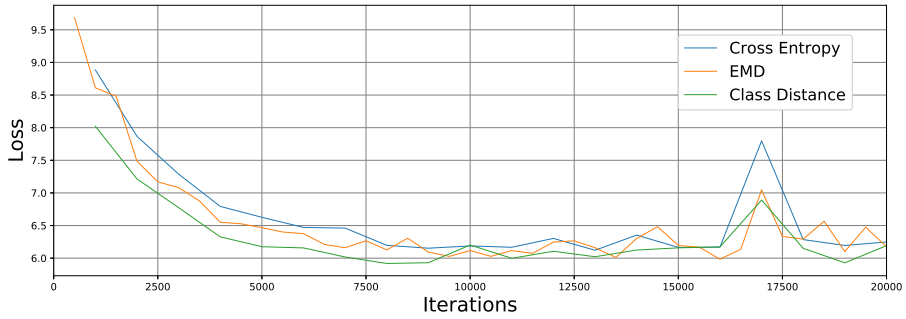
Table 4.3: Final MAE score of loss functions

Experiment 2: Monocular 3D Face Reconstruction Baseline

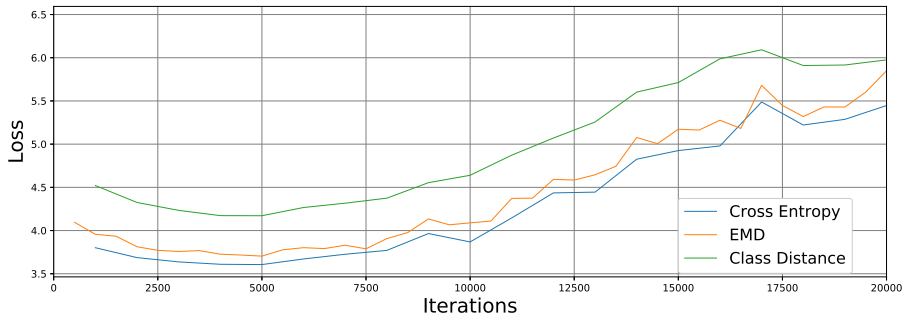
Experimental Setup:

In this experiment, the implementation of the monocular face reconstruction is trained to test how well it performs after training on the Wiki dataset and to keep the results for comparison with later models. The Wiki dataset is fairly small and this experiment will show if this impacts 3D facial geometry negatively.

The implementation sticks closely to the original MoFA model that was described in Section 3.1. The encoder has the original architecture of AlexNet. Weights from training for image classification on ImageNet were obtained and loaded before every

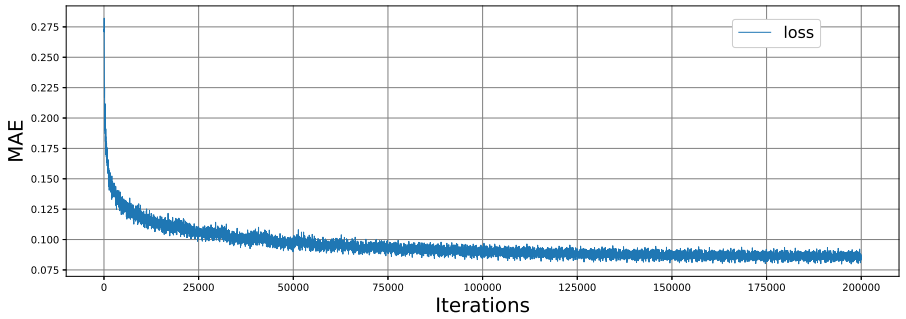


(a) Validation MAE

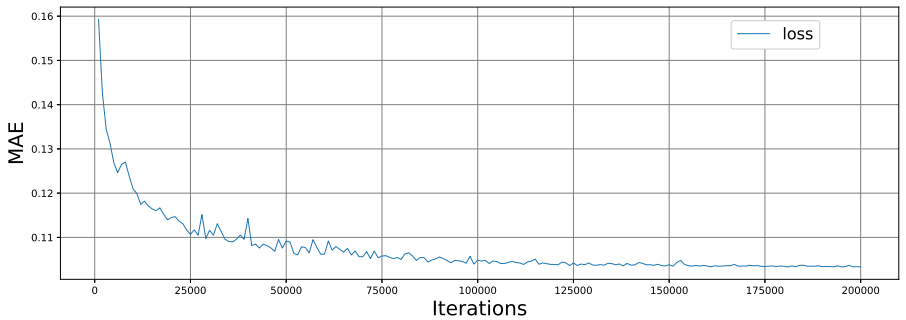


(b) Validation Loss

Figure 4.3: Classification losses experiment.



(a) Training photometric loss



(b) Validation photometric loss

Figure 4.4: MoFA baseline photometric loss.

experiment. The training parameters, established as default are as shown in Table 4.4.

MoFA was trained on Wiki for 200,000 iterations by loading standard pre-trained AlexNet weights on ImageNet classification. It is important to note that, while age estimation needs balancing of the batches, this task has no discrete labels, and so batch balancing was not performed. This lead to a higher amount of data available for training.

Results:

A measure of performance is the photometric loss, showing how close a predicted face pixel is to the corresponding one in the image. Fig 4.4 shows the photometric component of the loss from training and validation. In the end, the validation photometric loss is 0.103. For comparison, MoFA trained on the much larger CelebA dataset [Liu et al., 2015] gives a photometric loss of 0.104 in the end.

Fig. 4.5 visualizes some reconstructions from the validation set after training. First of all, the reconstructions are reasonable - the identity, illumination, expression, and pose are accurately reconstructed. The model is also doing well with quite extreme poses and faces of different weight, ethnicity, gender, and age. Examples show also older and younger people. Looking at the reconstructions, older age can be visually noticed. One feature is that older people have more prominent cheekbones and protruding cheeks than younger ones, especially directly to the left and right of the mouth corners. This is the result of fat depositing through aging. Noticing age on the reconstructed geometry already shows that it already contains age information.

learning rate (lr)	1e-5
lr decay rate (λ)	0.8
lr decay interval (δ)	20000
batch size	5
optimizer	Adam
w_{lan}	0.00192
w_{col}	1.92
w_{α}	$2.9 \cdot 10^{-5}$
w_{β}	$4.93 \cdot 10^{-8}$
w_{δ}	$2.32 \cdot 10^{-4}$

Table 4.4: MoFA training parameters

4.3 MoFA and age relationship

Experiment 3

The encoding learned when training a morphable model is known to encode facial attributes. With BFM 09 a specific vector is provided that, when multiplied by the principal components, alters the age attribute of the result. As BFM 2017 is designed to be even more descriptive, age estimation is expected to gain from the features required for an accurate hidden age encoding.

The question remaining is if monocular 3D face reconstruction can correctly learn to represent this hidden encoding, corresponding to the age of the person. As already discussed in Section 1, age has a notable impact on the face shape, and for older ages - on albedo and fine geometry, because of skin sagging and texture

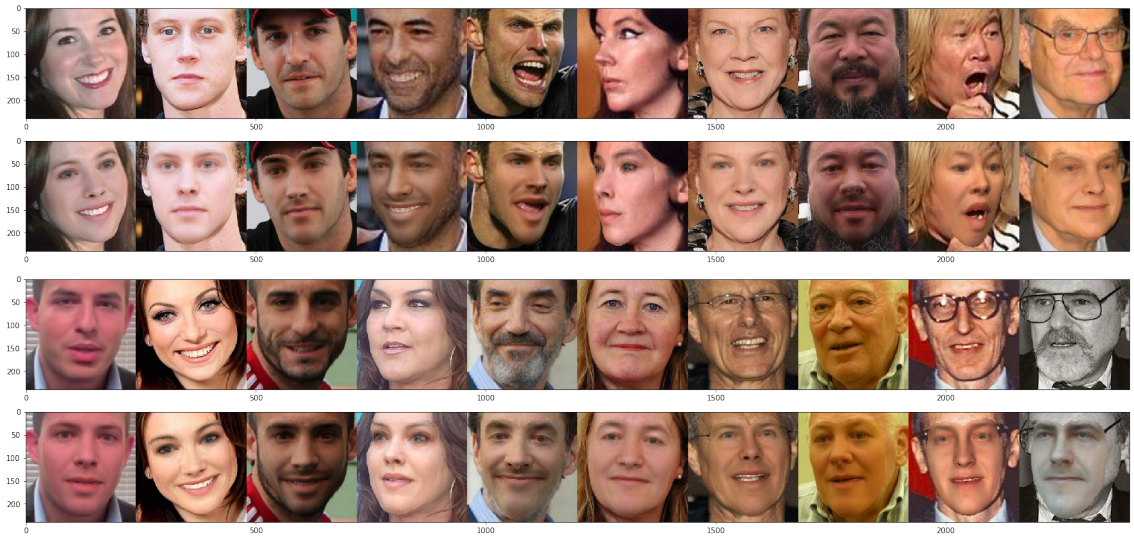


Figure 4.5: MoFA reconstruction examples from Wiki validation.

changes. MoFA does not model wrinkles and skin details well because of the linear model. This means we expect more general features to show age - like overall face shape, nose pointiness, prominence of nose flaps and the amount of protrusion of the cheeks. One claim to make is that the shape component can be expected to be a sufficient source of improvement of age estimation accuracy. Another one is that monocular 3D face reconstruction has to learn these age-related shape features, so it can achieve more likely-looking reconstructions with age characteristics matching the subject's age.

Experimental Setup:

Age estimation with MAE loss is trained on only the shape parameters of code vectors, generated by the pre-trained 3D face reconstruction from Experiment 2. It has to be noted that the layers of the 3D face reconstruction pipeline are frozen and only the regression layer is trained. This ensures that the work of estimating 3D facial geometry is unbiased by age estimation.

The training dataset is Wiki - the pre-trained MoFA was trained to reconstruct well on it. The learning rate had to be increased to 10^{-3} and no learning rate decay was used.

Results:

Fig. 4.6 shows the validation performance over time. After 40,000 iterations, the regressor scores 7.07 MAE undoubtedly showing that the shape parameters contain age information. Moreover, monocular 3D face reconstruction generated these code vectors, so it indeed successfully learned to express this age information.

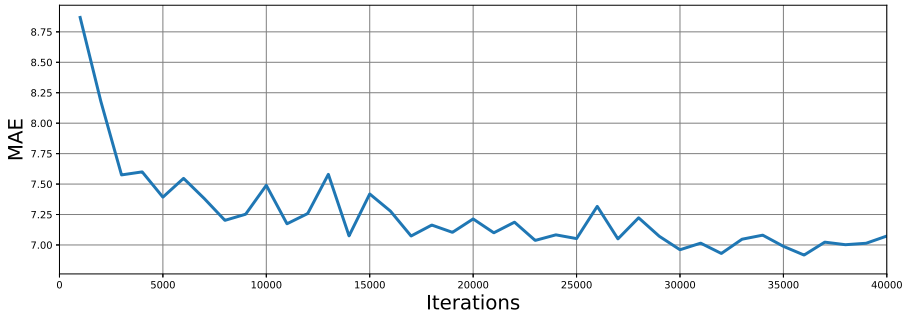


Figure 4.6: Age estimation validation loss on Wiki, of regression on shape parameters, generated by monocular 3D face reconstruction.

4.4 Proposed Multi-Task Learning Architectures

Experiment 4

Next, we attempt to train classification age estimation and 3D face reconstruction together. The expectation is that they will be able to benefit from each other by correcting and monitoring each other's features and from the regularization inherent to multi-task learning.

Hard parameter sharing introduces the weight parameter. A higher weight for age estimation will make its gradients dominate the network. This has two effects: 3D face reconstruction can be impaired by extreme unhelpful changes from age estimation and also would have a small influence on this second task. A too low weight means that the age estimation gradients are too small for the task to learn independently the important features that are unavailable in the estimated 3D facial geometry - e.g. wrinkles. We study the effect of changing the weight parameter.

Having a single CNN means that the features are constantly changed from both tasks. If the tasks are not close in terms of goal, it is beneficial to have separate features for each of the tasks at some layers. In this sense, this experiment also gives an understanding if there is a big overlap between the goals of the two tasks.

It is expected that if 3D facial geometry does not help in this architecture, the score will consistently increase by putting more emphasis (higher weight) on the age estimation because its gradients are consistently increasing. Otherwise, for some lower weight, the score will be better.

Experimental Setup:

The particular architecture that is trained, is described in Section 3.3.1.

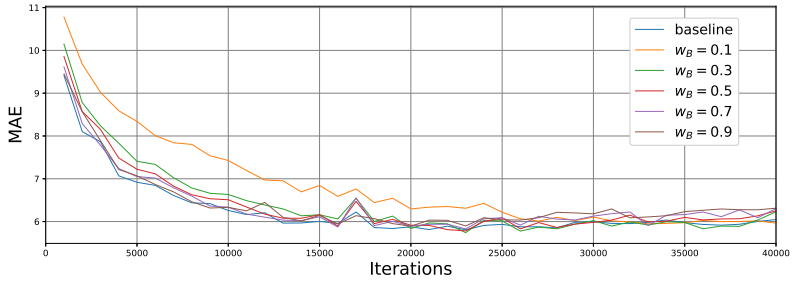
Training was performed with dropout with drop rate 0.7 and L2 regularization penalty with weight 0.00001 to be consistent with the following experiments with the soft parameter sharing. For a fair comparison, the age estimation baseline with Class Distance Loss was retrained with the same regularization choices. The single AlexNet was initialized with ImageNet weights.

Results:

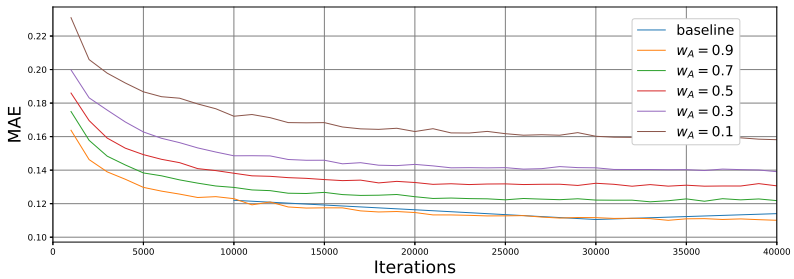
Fig. 4.7 and Table 4.5 gives an overview of the validation performance of models with different weights. The pattern of the 3D face reconstruction photometric loss

shows that age estimation has no sufficient influence on 3D face reconstruction in the current setup. However, it should be kept in mind that photometric loss is not explicitly and solely evaluating the 3D geometry of the reconstruction. This architecture showed small age prediction improvements over the baseline. The MAE scores were decreasing with putting lower weight but this is a sign that this task could use smaller gradient update steps. However, with a weight of 0.1, the gradients are too small and convergence is slower.

This experiment showed us that the simple sharing of the entire network is not a good enough choice of architecture and despite the presence of age cues for 3D facial geometry, some independence is required by the two tasks.



(a) Validation MAE



(b) Validation Photometric Loss

Figure 4.7: Hard parameter sharing with different task weights. Here we denote the 3D face reconstruction weight with w_A and the age estimation weight with w_B.

Age estimation weight	MAE
baseline	5.85
0.1	5.95
0.3	**5.74**
0.5	5.78
0.7	5.83
0.9	5.89

Table 4.5: Scores from hard parameter sharing.

Experiment 5

In the previous experiment, we attempted an architecture with fixed sharing of the entire CNN. However, a different part of the CNN may be more suitable to train, as discussed in Section 3.3.2. Instead of testing an extensive amount of architectures with hard parameter sharing, we train a soft parameter sharing architecture to learn what layers are helpful to share for a reduced loss on age estimation and 3D face reconstruction losses and to what extent. It is expected that the age estimation pipeline in this architecture will make use of the 3D facial geometry features. Also, classification has overfitting issues and can use additional forms of regularization. As discussed in Section 2.4, multi-task learning provides regularization. Therefore, we expect there to be a significant improvement over the baseline.

Experimental Setup:

We employ the Cross-Stitch Network architecture, as described in 3.3.2. First, we consider the regularization technique to use for classification with soft parameter joint learning. As there is a regularization term in the 3D face reconstruction, it is not helpful to do L2 regularization on its weights. This leaves us with the choice to do the L2 penalization only on the age estimation's branch. However, we chose the small weight of 0.00001 for the L2 penalization and also do dropout on the final layer of age estimation. In this way the effect of direct regularization cannot be traced from a single branch and age estimation would not try to overtake the unpenalized MoFA branch. The chosen drop rate was 0.7 after some experimentation. For comparison, we use the retrained Class Distance Loss baseline with the above-described regularization scheme.

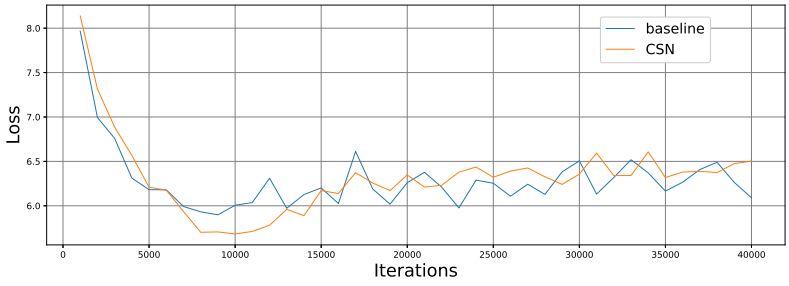


Figure 4.8: Soft parameter sharing - Validation MAE

Classification	MAE
baseline	5.85
CSN	**5.58**

Table 4.6: Scores from soft parameter sharing.

Results:

As seen from the results on Fig. 4.8 and Table 4.6, the architecture offers a sizeable reduction of the Mean Absolute Error on Wiki validation set.

Fig 4.9 visualizes how the α parameters change with the training of the classification model. To remind, although the parameters are initialized to do blending

with weights summing to 1, they are not constrained to do so while training, as part of the original architecture. Scaling up or down of the features of both layers can happen in any configuration. We judge the importance of a sharing layer by checking individually for each shared layer if the sharing parameter increases and how much. It can be noticed that the third shared layer gives the most benefit to age estimation. Sharing happens also on the first two layers. For the last two layers, it seems that the preference of age estimation is to stay separated. This separation allows the network to learn task-specific features with the fully connected layers but still make use of 3D face geometry features from the convolutional layers.

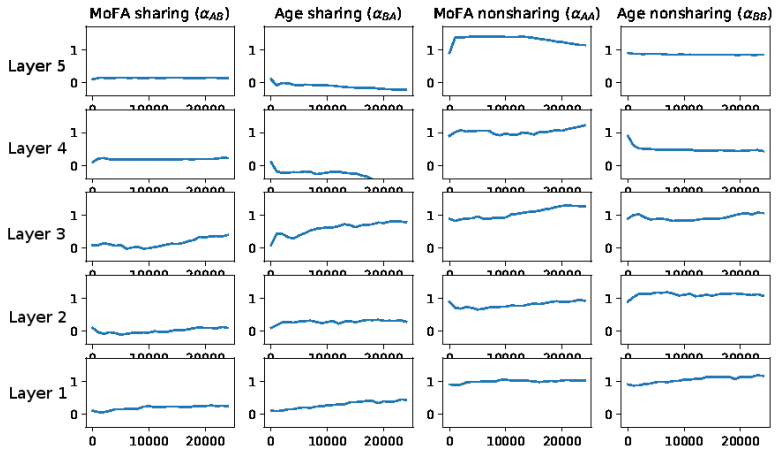


Figure 4.9: Visualization of the α parameters for 25,000 iterations of training.

Experiment 6

Experimental Setup:

For classification, 5 training sessions of the baseline and the soft parameter sharing with age classification each, were performed to test if the improvement is a random deviation.

	Mean ± Std
baseline	5.86 ± 0.04
CSN	5.53 ± 0.03

Table 4.7: Means and deviations calculated from 5 training sessions.

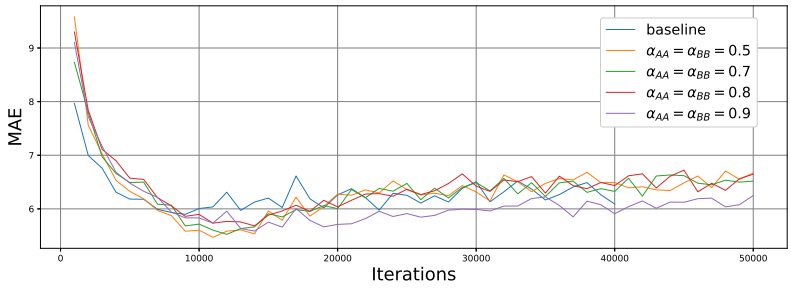
Results:

On Table 4.7 are displayed the means and the standard deviations. The scores of the soft parameter sharing model seem to be even lower than the one received from the last trained model (5.58 MAE). The standard deviations are small and there is no overlap between the baseline's and the proposed model's distributions of the MAE scores. We can conclude that the improvement of soft parameter sharing over the baseline is stable.

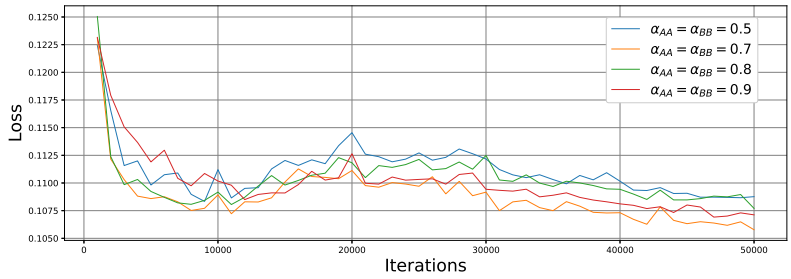
Experiment 7

Optimization can go in a different direction if we choose to share more between the branches at the beginning. The goal of this experiment is to check what is the optimal choice for initialization of the α parameters of the CSN.

Our initialization follows the rules $\alpha_{BB} = \alpha_{AA}$ and $\alpha_{AB} = \alpha_{BA} = 1 - \alpha_{AA}$. In that way, each branch is sharing the same amount from the other. Also, by summing to 1, it is ensured that the network is performing intuitively interpretable blending at the start and that the scale of the features is preserved. Non-sharing (α_{AB} and α_{BA}) values that are reasonable to test are in the range $[0.5, 1]$. If smaller values are chosen, the branches would just switch their base CNNs. If they are larger, the rule for the parameters summing up to one would be broken.



(a) Validation MAE



(b) Validation Photometric Loss

Figure 4.10: Tuning of the α parameters.

Training was done on Wiki with the same parameters and choices previously adopted for CSN. Two initializations close to the recommended one was tested (0.7 and 0.8 for the non-sharing alphas) and one enforcing much more sharing between the branches (0.5). The validation MAE for all the runs is plotted on Fig. 4.10 and the final scores are on Table 4.8.

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

Table 4.8: Final MAE scores from tuning α parameters.

The results show that age estimation immediately benefits from large sharing (0.5). However, MoFA suffers in this case as it has to compete with the other task - despite balancing MoFA has smaller gradients at the start of training. Moderate sharing proves most helpful for both tasks. We choose this best model for further experiments.

Experiment 8

Considering the reconstructed 3D facial geometry is age-related, it can be claimed that 3D face reconstruction alone learns internal features in its AlexNet that are age-related. This naturally would make these features suitable for direct use from age estimation. It can be argued that much of the success of soft parameter sharing for age classification comes from the pretraining. We study this claim by initializing the visual age estimation baseline with 3D face reconstruction pretraining and training. Then the obtained validation MAE is compared with the ones from the baseline and the soft parameter sharing model.

Experimental Setup:

The age estimation baseline with Class Distance Loss was trained with initially loading the same MoFA weights that were used as a pretraining in the soft parameter sharing architecture - the ones from pretraining on Wiki for 100,000 iterations. The rest of the parameters correspond to the ones in the proposed soft parameter sharing model and the baseline.

Results:

After training, the model achieved 5.62 validation MAE. This shows the major source of the success of the soft parameter sharing model - making use of the pre-trained 3D facial geometry features. Also, it explains why higher sharing α parameters work well. With soft sharing, however, a better MAE score was achieved (5.47 was the best). The multi-task learning properties to regularize the combined tasks account for this improvement. The comparison is shown in Fig. 4.11.

Note that the pretraining was done on the cleaned and clipped Wiki dataset, which is the same as the training data for the soft parameter sharing model, but without batch age balancing. Having the same dataset ensures that the 3D facial geometry does not have the advantage of assimilating a greater or different amount of data.

Figure 4.11: Validation MAE comparison of a pretrained on MoFA baseline.

Experiment 9

To show if the results extend beyond the dataset used for training, evaluation is done on FGNET, UTKFace, and AgeDB. The expectation is to obtain MAE scores with soft parameter sharing, which are significantly larger than the MAE scores of the baseline.

Experimental Setup:

	UTKFace	FGNET	AgeDB
baseline	9.73	16.89	10.27
CSN	9.54	16.54	10.01
t-test p-value	$3.22 \cdot 10^{-9}$	0.01	$1.78 \cdot 10^{-8}$

Table 4.9: MAE scores from cross dataset evaluation.

We obtained the weights from a soft parameter sharing model trained on Wiki and did not do any fine tuning on the target datasets - cross-dataset evaluation. This lack of fine-tuning means we do not expect low MAE scores but we expect to see CSN giving lower scores than the ones from the baseline.

Results:

Results are shown on Table 4.9. It can be seen that these expectations are met. The improvements are significant, their consistency shows the soft sharing multi-task learning model is generalizable.

Experiment 10

In this experiment, we attempt to confirm the statistical significance of the MAE improvement that the soft parameter sharing model brings.

Experimental Setup:

Statistical testing was performed on the model with the lowest score ($\alpha_{AA} = 0.5$). First, it has to be established that we want to compare 2 distributions - each being build from the age differences between validation labels and predictions. In the first distribution, the predictions are from the baseline and in the other - from the CSN model. We want to show that there are significant differences between the two distributions. The difference between both is plotted in Fig. 4.12.

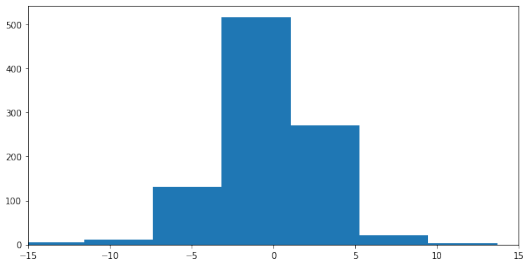


Figure 4.12: Distribution of differences between MAE measures of the baseline and the soft parameter sharing model.

Visually, this has a resemblance with the normal distribution. If normality is established, the most popular statistical testing method - t-test can be used without concerns. After running a normality test, normality was confirmed (the score is $1.79 \cdot 10^{-20}$ - far less than the threshold of 10^{-3}). The null hypothesis now is that the mean of the two distributions is not significantly different - in the specific case, the difference between the two distributions has mean 0. The difference is needed to enforce the pairing of the sample values and prevent variance confusion.

Results:

After running t-test, significance was confirmed. The p-value is $2.70 \cdot 10^{-5}$ - less than the generally used threshold of 0.05 for null hypothesis rejection, the statistic is 4.218 - larger than the threshold for 99% confidence.

Before normality was established here, testing with Wilcoxon non-parametric

test was also done. This test also checks if the distributions are the same but do not require normality. It also confirmed the significance of results with a p-value of $2.77 \cdot 10^{-4}$.

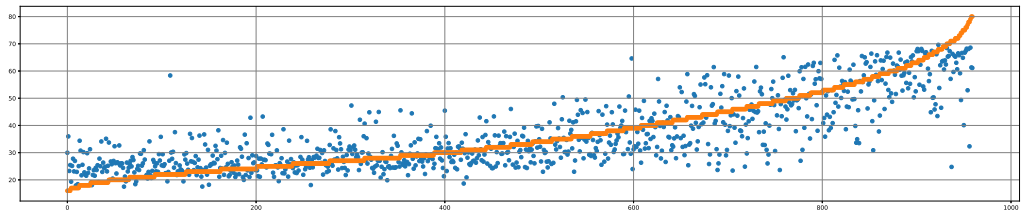
Experiment 11

In the current experiment, we explore if the MAE improvement observed from soft parameter sharing is expressed in certain age groups. We expect there to be an improvement in the groups with smaller but sufficient training data, where regularization can help.

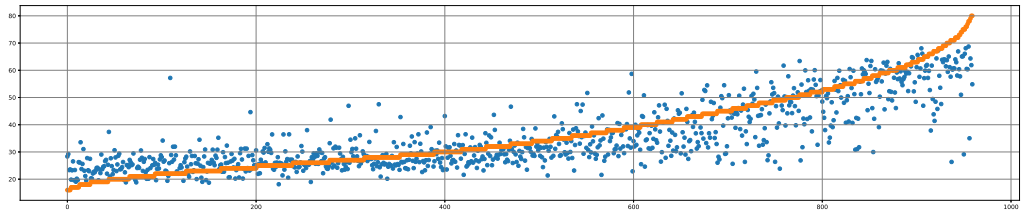
Experimental Setup:

All the predictions of the trained soft parameter model for Wiki Validation, AgeDB, UTKFace, and FGNET were extracted and stored. The same was done for the age estimation baseline. Using the collected predictions, a comparison between the two models is done, as described below.

Results:



(a) Baseline



(b) CSN

Figure 4.13: MAE Variance on Wiki validation. The x axis represents individuals, sorted by their ground truth age; the y axis represents the age. Orange points show ground truth age and blue points - predictions.

On Fig. 4.13 the variance is visualized for the baseline and the soft parameter sharing model. The differences are subtle, but it can be still noticed that the overall variance is reduced in the results from joint learning - the predictions are closer to the ground truth curve.

The boxplots in Fig. 4.14 and Fig. 4.15 show a comparison of the baseline and the proposed soft sharing model for age groups with a width of 10 years. For each group, the significance, the mean and the significance p-value are shown. The chosen statistical testing is Wilcoxon because then normality does not need to be established for each age group.

The FGNET results, although showing improvement, are not to be considered because of insignificance noticed from the high p-values on every age interval on Fig.

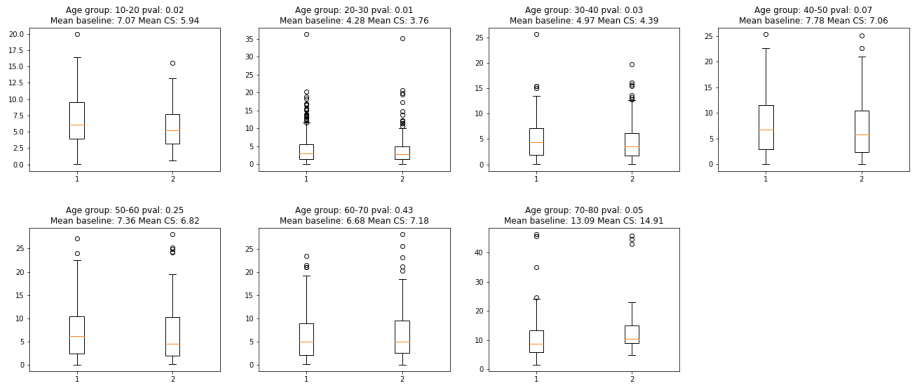


Figure 4.14: Age group boxplots of Wiki validation MAE results.

4.15b. One major problem with training our models is the lack of data for young people and FGNET's distribution is skewed towards children.

The too scarce data between 0 and 20 years and above 60 years gives also a small sample size for statistical testing which affects the MAE scores and the p-values. The improvements, noticed on the boxplots, are in the range between 30 and 50 years old. Most of the data is concentrated between 20 and 35 years old. Ranges that have less data but still enough for training seem to benefit more from the additional 3D face geometry and the regularization.

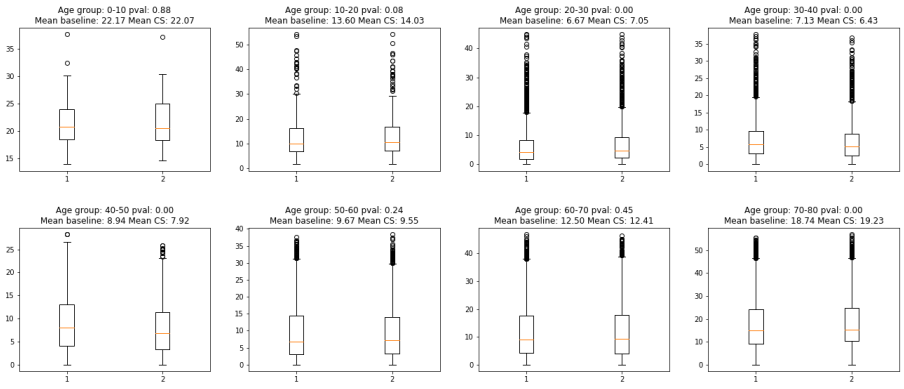
The conclusion is that the significant improvement comes from the age range 30 to 50 years old, which was shown with statistical significance on Wiki validation, AgeDB, and UTKFace.

Experiment 12

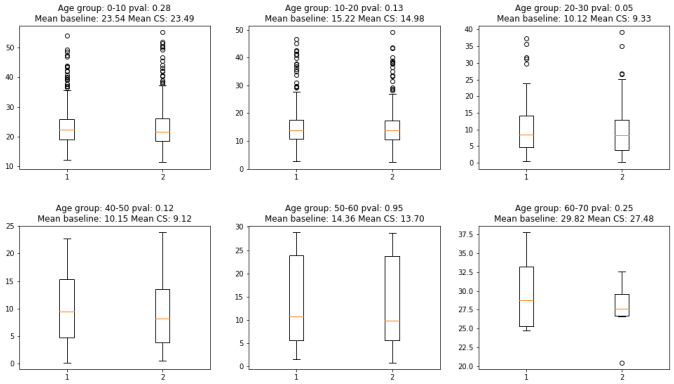
Here we show image examples and their age predictions from the baseline and the soft parameter sharing model in an attempt to find further insights when our proposed method succeeds and fails. In Experiment 2, it was already observed that 3D face reconstruction is doing well with non-frontal face poses. A reasonable reconstruction of non-frontal face image encodes 3D geometry in the same way as it would with a frontal face. Therefore, we expect improvement for non-frontal face poses when employing 3D face geometry features. Pure geometry also does not contain facial expression. Therefore we expect a noticeable improvement for extreme facial expression. To test these claims we design a procedure to divide the validation images into groups of expression and pose extremeness and then we evaluate the MAE improvement for each group.

Experimental Setup:

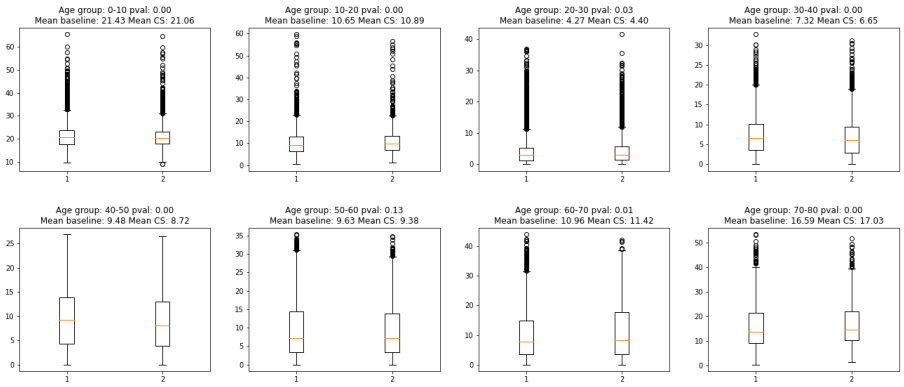
First, the MAE scores for each image in the Wiki validation is obtainedd for the baseline ($MAE_{baseline}$) and the soft parameter sharing model (MAE_{CSN}). The MAE difference between the two scores for each image is then computed with $MAE_{baseline} - MAE_{CSN}$. This gives us how much the proposed model is better than the baseline for this image. Negative value means that the soft sharing model



(a) AgeDB



(b) FGNET



(c) UTK

Figure 4.15: Age group boxplots of the cross-dataset evaluation MAE results.

is giving worse predictions than the baseline. Each image in the validation is associated with a blendshape vector and rotation. We obtain metrics for each of those items, which we will use to build groups of images.

For blendshapes we calculate the Eucledian norm of the vector: $\sqrt{\sum_{i=1}^{64} v_i^2}$, where v_i is one of the 64 values in the vector. Since the vector is the deviation from the mean face (and the mean expression is the neutral expression), this is a measure of how extreme the expressions are. For rotation, we take the maximum of the exponential coordinates that parameterize a rotation in SO3 space. It gives us a measure of how extreme the predicted pose is.

We sort all the images in the validation set by each of the metrics described above, in increasing order. The sorted validation set is next divided into intervals based on the measure. For blendshapes the metric is in the interval $[0.49, 3.31]$. We split this interval into 6 groups with approximate size 0.3, the values at which the interval is split are $[1, 1.3, 1.6, 1.85, 2.45]$. The small corrections in the intervals are made so there is enough data for each group. Rotation values go from 0.002 to 0.45. Each group has a width of 0.5, split values are $[0.1, 0.2, 0.3, 0.35]$. For each of the intervals, the mean of the MAE differences over all the images falling in the interval is computed and plotted.

Results:

Fig. 4.16 shows some samples from the validation set and their age predictions. Improvement is generally noticed even across expressions and ethnicity.

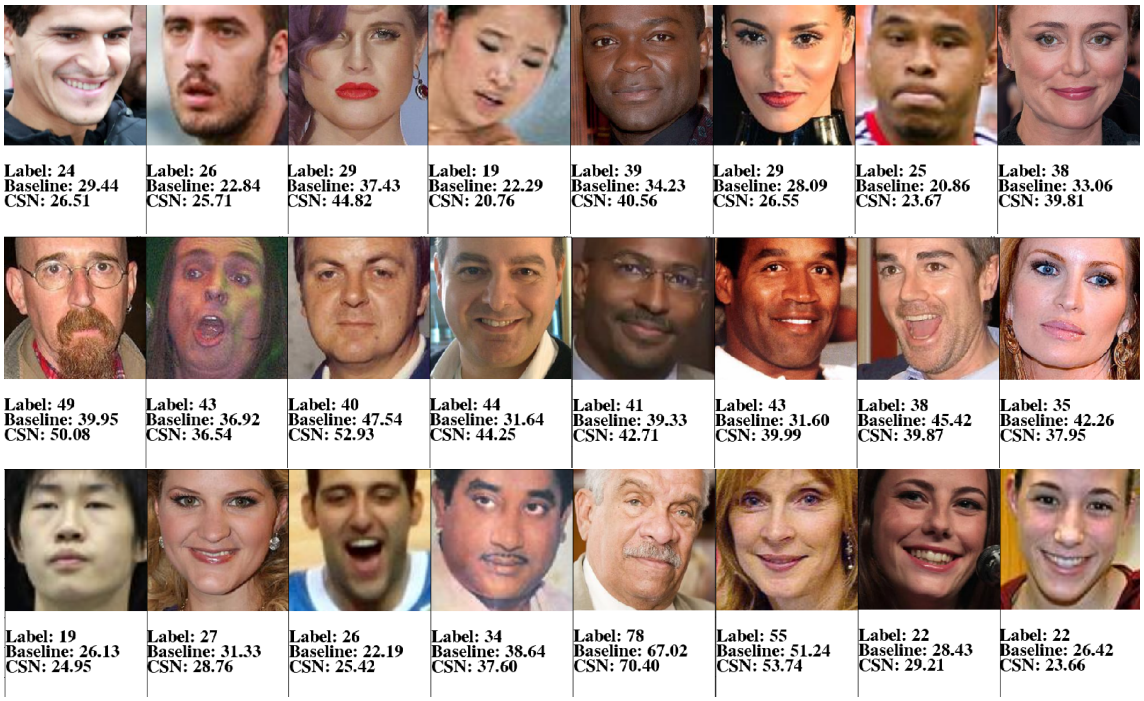


Figure 4.16: Samples from the Wiki validation set and their age predictions.

Fig. 4.17 visualizes the expression severity of each group by showing samples.It is seen that the expression starts from neutral and gradually introduces mouth opening and smiles until extreme expressions are reached - strong smiles or widely opened mouth. In Fig. 4.18 shows how the improvement over the baseline changes throughout the groups. It can be seen that most of the improvement is focused on the groups with non-neutral expression. It can be seen that given images with

very extreme expression, there is a large amount of improvement from the proposed model.

Fig. 4.20 visualizes what kind of poses each group contains. It is visible that the first groups contain frontal faces, while the last groups - faces with large angles to the camera. Fig. 4.21 visualizes the improvement offered by the soft parameter sharing model for each of the rotation groups. It is seen that the trend is that with increasing the amount of rotation, the improvement rises. The rotation invariance of the geometry features can come in handy with the lack of rotation invariance of the visual features. For more extreme poses, the visual baseline is likely to fail and the model that includes 3D geometry can offer better predictions in this case, as can be seen in Fig. 4.19.

Figure 4.17: Samples from the expression intensity groups. Each row contains samples from one group. Groups are sorted with increasing of the metric from top to bottom

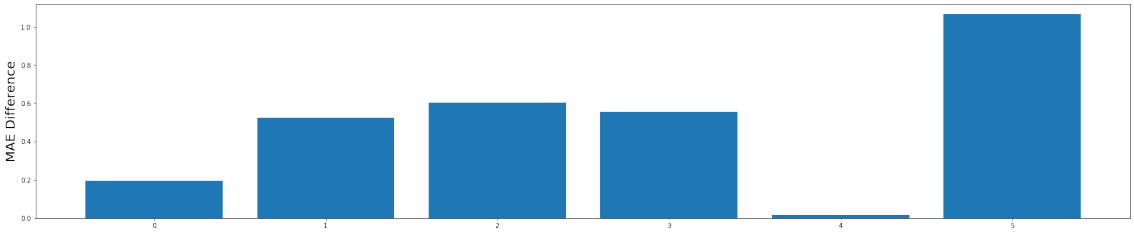


Figure 4.18: The MAE improvement of the proposed soft sharing model over the baseline over the expression intensity groups. The expression intensity metric is increasing in the groups from left to right.

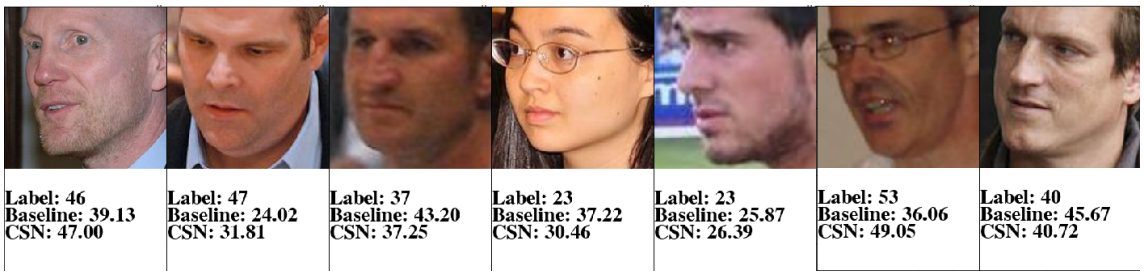


Figure 4.19: Samples from the Wiki validation with non-frontal faces.

Figure 4.20: Samples from the rotation groups. Each row contains samples from one group. Groups are sorted with increasing of the metric from top to bottom

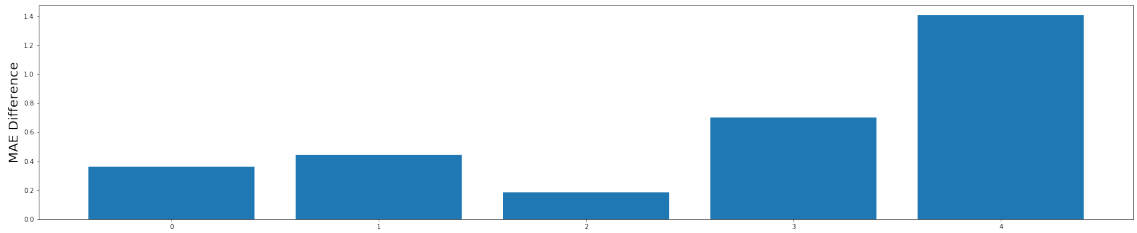


Figure 4.21: The MAE improvement of the proposed soft sharing model over the baseline over the rotation extremeness groups. The rotation extremeness metric is increasing in the groups from left to right.

Chapter 5

Conclusion

In this thesis, we set out to find the answer to the questions "Can monocular 3D face reconstruction produce 3D facial geometry that is age discriminative?" and "Can a combination of traditional age estimation with 3D facial geometry lead to improved age predictions?". Evidence shows that 3D facial geometry is age descriptive [Angulu et al., 2018, Alley, 1988, Todd et al., 1980, Zhuang et al., 2010]. Also, it is known that statistical 3D face models employed by monocular 3D face reconstruction solutions are biologically accurate. We argue that monocular 3D face reconstruction should then implicitly learn age-related features in order to produce quality geometry reconstruction. Standard age estimation relies only on 2D visual information, that models mostly wrinkle and skin texture information [Choi et al., 2011, Hayashi et al., 2002, Ng, 2015]. We are the first ones that seek to combine the traditional age estimation with 3D facial geometry in an attempt to utilize this new source of information and achieve improved age prediction as a result.

First, it has been shown that 3D a deep learning based 3D face reconstruction is able to learn age implicitly, by finding that there is a correspondence between age and the learned encoding for 3D facial geometry. This answers our first research question - monocular 3D face reconstruction can produce 3D facial geometry that is age descriptive.

Following the success of up-to-the-year age classification proposed by [Rothe et al., 2018], we also employ classification for the noisy training dataset. We proposed a new simple loss, which we refer to as Class Distance Loss, that is able to introduce the concept of distance between the discrete age classes. It makes it easier for the network to learn age-related features characterizing a small interval around the ground truth age. It outperformed cross entropy and EMD losses.

We have applied a multi-task soft parameter sharing architecture to combine age estimation and 3D facial geometry. The performance of this new model was studied for age estimation.

The soft parameter sharing model was able to give statistically significant improvement on the Mean Absolute Error (MAE) scores from validation and cross-dataset evaluation on the external datasets FGNET, UTKFace, and AgeDB. It was shown that the internal 3D geometry features are a great source of improvement, as they caused a decreased MAE measure when the 3D face reconstruction baseline weights were used as age estimation initialization.

Comparative analysis on the MAE results of the soft sharing model and the visual baseline showed that the proposed model excels in extreme expressions and poses, and to increasing expression intensity and rotation up to a certain level. The causes of the improvement is the invariance of the 3D facial features to pose and expression.

From the results of the proposed combined model and the fact that 3D facial geometry was identified as a main source of improvement, we have found the answer of the second research question: traditional age estimation can benefit from 3D facial geometry features as an additional source of information.

5.1 Future Work

Children were underrepresented in the available training set. Considering that age geometry changes occur at the early stages of life, 3D face reconstruction is expected to greatly influence age estimation in exactly these age groups. A large enough children's faces dataset is not publicly available and obtaining one is a difficult task. Even if collected data is not substantially larger than for other age groups, a specialized model for children age estimation can be trained and combined with the full model, as in [Antipov et al., 2016].

MoFA with BFM 2017 results in quite smoothed faces. Introducing additional fine details on top of the predicted geometry and including them in the age estimation procedure can help to capture important subtle features that define age.

This thesis focuses on improving the age estimation task with monocular 3D face reconstruction. However, the face reconstruction task can also benefit from age estimation. 3D face reconstruction can learn to create more likely facial geometries with the visible age of the person.

While we looked into age estimation, a range of other tasks can be improved by or used to improve estimation of monocular 3D face reconstruction. For example, the expression and emotion prediction tasks are directly related to the blendshape component of the code vector and gender classification - to geometry and albedo.

A promising direction is to further study how good are the weights of the 3D face reconstruction model as pre-training for age estimation. DEX owes its success to pre-training on IMDB-Wiki. In this thesis the benefit of pre-trained 3D face reconstruction weights for age estimation became clear. The two pre-training methods can be compared to see if 3D face reconstruction is actually better. Also, combining both methods by initializing with the 3D face reconstruction weights fine tuning on IMDB-WIKI could lead to more accurate age estimation.

Bibliography

Alley, T. R. (1988). The effects of growth and aging on facial aesthetics.

Amberg, B., Knothe, R., and Vetter, T. (2008). Expression invariant 3d face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE.

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

Antipov, G., Baccouche, M., Berrani, S.-A., and Dugelay, J.-L. (2016). Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 96–104.

Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning, in 'advances in neural information processing systems 19'.

Bakker, B. and Heskes, T. (2003). Task clustering and gating for bayesian multitask learning. *Journal of Machine Learning Research*, 4(May):83–99.

Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39.

Beeler, T., Bickel, B., Noris, G., Beardsley, P., Marschner, S., Sumner, R. W., and Gross, M. (2012). Coupled 3d reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics (ToG)*, 31(4):117.

Bérard, P., Bradley, D., Gross, M., and Beeler, T. (2016). Lightweight eye capture using a parametric model. *ACM Transactions on Graphics (TOG)*, 35(4):117.

Besl, P. J. and McKay, N. D. (1992). Method for registration of 3-d shapes. In *Sensor Fusion IV: Control Paradigms and Data Structures*, volume 1611, pages 586–607. International Society for Optics and Photonics.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co.

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

Bouaziz, S., Wang, Y., and Pauly, M. (2013). Online modeling for realtime facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):40.

Cao, C., Bradley, D., Zhou, K., and Beeler, T. (2015). Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):46.

Cao, C., Wu, H., Weng, Y., Shao, T., and Zhou, K. (2016). Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics (TOG)*, 35(4):126.

Caruna, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning: Proceedings of the Tenth International Conference*, pages 41–48.

Chai, M., Shao, T., Wu, H., Weng, Y., and Zhou, K. (2016). Autohair: fully automatic hair modeling from a single image. *ACM Transactions on Graphics (ToG)*, 35(4):116.

Cheng, J., Tsai, Y.-H., Wang, S., and Yang, M.-H. (2017). Segflow: Joint learning for video object segmentation and optical flow. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 686–695. IEEE.

Choi, S. E., Lee, Y. J., Lee, S. J., Park, K. R., and Kim, J. (2011). Age estimation using a hierarchical classifier based on global and local facial features. *Pattern Recognition*, 44(6):1262–1281.

Coffin, J. S. and Ingram, D. (1999). Facial recognition system for security access and identification. US Patent 5,991,429.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685.

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59.

Cosker, D., Krumhuber, E., and Hilton, A. (2011). A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling.

Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362.

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

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(Apr):615–637.

Farkas, L. G. (1994). *Anthropometry of the Head and Face*. Raven Pr.

Fu, Y., Xu, Y., and Huang, T. S. (2007). Estimating human age by manifold analysis of face pictures and regression on aging features. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 1383–1386. IEEE.

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

Garrido, P., Zollhöfer, M., Wu, C., Bradley, D., Pérez, P., Beeler, T., and Theobalt, C. (2016). Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 35(6):219–1.

Geng, X., Zhou, Z.-H., and Smith-Miles, K. (2007). Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schönborn, S., and Vetter, T. (2018). Morphable face models-an open framework. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 75–82. IEEE.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

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

Guo, Y., Zhang, J., Cai, J., Jiang, B., and Zheng, J. (2017). 3dfacenet: real-time dense face reconstruction via synthesizing photo-realistic face images. *arXiv preprint arXiv:1708.00980*.

Hashimoto, K., Xiong, C., Tsuruoka, Y., and Socher, R. (2016). A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Hayashi, J.-i., Yasumoto, M., Ito, H., and Koshimizu, H. (2002). Age and gender estimation based on wrinkle texture and color of facial images. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 405–408. IEEE.

Heskes, T. (2000). Empirical bayes for learning to learn.

Hou, L., Yu, C.-P., and Samaras, D. (2016). Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

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

Huang, S., Quek, S., and Phoon, K. (2001). Convergence study of the truncated karhunen–loeve expansion for simulation of stochastic processes. *International journal for numerical methods in engineering*, 52(9):1029–1043.

Jacob, L., Vert, J.-p., and Bach, F. R. (2009). Clustered multi-task learning: A convex formulation. In *Advances in neural information processing systems*, pages 745–752.

Kang, Z., Grauman, K., and Sha, F. (2011). Learning with whom to share in multi-task feature learning. In *ICML*, pages 521–528.

Kemelmacher-Shlizerman, I. and Basri, R. (2011). 3d face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2):394–405.

Kendall, A., Gal, Y., and Cipolla, R. (2017). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *arXiv preprint arXiv:1705.07115*, 3.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., and Theobalt, C. (2018). Deep video portraits. *arXiv preprint arXiv:1805.11714*.

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

Kokkinos, I. (2017). Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, volume 2, page 8.

Kumar, A. and Daume III, H. (2012). Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

Kwon, Y. H. et al. (1994). Age classification from facial images. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 762–767. IEEE.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H., Hawk, S. T., and Van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388.

Lanitis, A. (2008). Comparative evaluation of automatic age-progression methodologies. *EURASIP Journal on Advances in Signal Processing*, 2008:101.

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

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–42.

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

Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M., and Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system.

Long, M. and Wang, J. (2015). Learning multiple tasks with deep relationship networks. *CoRR, abs/1506.02117*, 3.

Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., and Feris, R. S. (2017). Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *CVPR*, volume 1, page 6.

Lüthi, M., Gerig, T., Jud, C., and Vetter, T. (2017). Gaussian process morphable models. *IEEE transactions on pattern analysis and machine intelligence*.

McDonagh, S., Klaudiny, M., Bradley, D., Beeler, T., Matthews, I., and Mitchell, K. (2016). Synthetic prior design for real-time face tracking. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 639–648. IEEE.

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

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5.

Müller, C. (1966). Spherical harmonics, volume 17 of lecture notes in mathematics.

Negahban, S. and Wainwright, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of l 1, inf-regularization. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1161–1168. Curran Associates Inc.

Ng, C. C. (2015). *Face age estimation using wrinkle patterns*. PhD thesis, Manchester Metropolitan University.

Patel, A. and Smith, W. A. (2009). 3d morphable face models revisited. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1327–1334. IEEE.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. Ieee.

Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104.

Ramanathan, N. and Chellappa, R. (2006). Modeling age progression in young faces. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 387–394. IEEE.

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

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ruder12, S., Bingel, J., Augenstein, I., and Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *stat*, 1050:23.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., and Akarun, L. (2008). Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer.

Sela, M., Richardson, E., and Kimmel, R. (2017). Unrestricted facial geometry reconstruction using image-to-image translation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1585–1594. IEEE.

Sengupta, S., Kanazawa, A., Castillo, C. D., and Jacobs, D. W. (2018). Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *Computer Vision and Pattern Regognition (CVPR)*.

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

Sykes, L., Bhayat, A., and Bernitz, H. (2017). The effects of the refugee crisis on age estimation analysis over the past 10 years: A 16-country survey. *International journal of environmental research and public health*, 14(6):630.

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

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

Todd, J. T., Mark, L. S., Shaw, R. E., and Pittenger, J. B. (1980). The perception of human growth. *Scientific american*, 242(2):132–145.

Tran, A. T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., and Medioni, G. (2017). Extreme 3d face reconstruction: Looking past occlusions. *arXiv preprint arXiv:1712.05083*.

Turk, M. and Pentland, A. P. (1999). Face recognition system. US Patent App. 08/340,615.

Vlasic, D., Brand, M., Pfister, H., and Popović, J. (2005). Face transfer with multilinear models. *ACM transactions on graphics (TOG)*, 24(3):426–433.

Wang, C., Shi, F., Xia, S., and Chai, J. (2016). Realtime 3d eye gaze animation using a single rgb camera. *ACM Transactions on Graphics (TOG)*, 35(4):118.

Wang, M., Shu, Z., Panagakis, Y., Samaras, D., and Zafeiriou, S. (2017). An adversarial neuro-tensorial approach for learning disentangled representations. *arXiv preprint arXiv:1711.10402*.

Wang, X., Guo, R., and Kambhamettu, C. (2015). Deeply-learned feature for age estimation. In *2015 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 534–541. IEEE.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

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

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

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. J. (2006). A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Zhang, C.-H., Huang, J., et al. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594.

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

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

Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer.

Zheng, T., Deng, W., and Hu, J. (2017). Age estimation guided convolutional neural network for age-invariant face recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 12–16.

Zhuang, Z., Landsittel, D., Benson, S., Roberge, R., and Shaffer, R. (2010). Facial anthropometric differences among gender, ethnicity, and age groups. *Annals of occupational hygiene*, 54(4):391–402.

Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., and Theobalt, C. (2018). State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library.