Dermatologically Inspired Deep Face Analytic

Newcastle University

School of Computing

PhD's Thesis

submitted by

Conor Stuart Charles Turner

Newcastle, September 2023

Abstract

This PhD thesis explores the use of convolutional neural networks (CNNs) for the extraction of age related health information from face images. CNNs are well suited to face image analysis thanks to their ability to learn robust features from large diverse datasets. Our research focuses on illuminating the intersection between deep learning for face image analysis and healthy ageing, leveraging the key methods supporting state-of-the-art age estimation. We also investigate the influence of image preprocessing on face analysis, emphasising the significance of face alignment algorithms and other more novel techniques. Finally, we validate the summation of our work using health outcomes, the gold standard endpoint for measures of ageing.

The first part of our study revolves around image preprocessing techniques, investigating the influence of face alignment algorithms in downstream face image analysis tasks and proposing our own system for eye based alignment. Additionally, we investigate the efficacy of semi-supervised methods to facilitate learning from a novel medical dataset containing 3D rendered faces with heavy confounding artefacts.

The second objective is to analyse the properties of different CNN based methods for perceived age estimation. We assess the strengths and weaknesses of transfer learning schemes, layer configurations and unsupervised learning approaches. Through extensive experimentation, we elucidate the optimal design choices and achieve a new state-of-the-art in CNN-based perceived age estimation.

Lastly we assess the usefulness of our perceived age models on their association with relevant health outcomes and genetic markers, providing objective measures of its efficacy, free from the bias of human perception. We demonstrate the potential of automated perceived age estimations linked to health outcomes, paving the way for accelerated biological research by removing the bottleneck of human assessors.

Acknowledgements

I would like to thank Newcastle University and the ICOS Research Group for the research framework and doctoral training provided, without this structure my thesis would not have been possible. Additional thanks is deserved by our industry partner Unilever, who provided much of financial and collaborative means needed for this project to reach its full potential.

Special thanks is reserved for my supervisory team, who provided guidance and criticism where it was most needed. In particular I would like to thank Jaume Bacardit for imparting his extensive knowledge of deep learning algorithms to me, showing me the high level procedures needed for impactful machine learning research, your sage advice proved priceless time and again throughout the process of the project. I would also like to thank David Gunn for his insights into the industrial and biological facets of our work, from which I sourced a great deal of motivation. I commend Ruediger Zillmer and Micheal Catt for their advice on some of the more esoteric facets of deep learning, as well as their poignant outlook on the future of the technology. The last academic I would like to thank is Luba Pardo Cortes for her insights into skin ageing and welcoming personality, I thoroughly enjoyed our collaboration.

Aside from my academic colleagues I would like to thank Christopher Haywood for his ongoing moral support, without your encouragement and comedic intervention this project would have felt insurmountable.

Finally I would like to thank my parents for their ongoing support and vehement belief in education. Without their promotion of self directed learning during my younger years it is unlikely I would not have been able to complete this PhD.

This thesis received funding from the EPSRC and the Industrial Case Scholarship Scheme under the grant EP/S513684/1, co-funded by Unilever research. A link to the project is available here: `https://gtr.ukri.org/projects?ref=studentship-2384812`

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context of the Study

The human perception of facial age has garnered significant attention in various domains, including dermatology [197], epidemiology [171] and computer vision [32]. While the task of guessing how old someone looks from their face appears to be trivial, the factors involved [48] and potential applications in healthcare [244] and epidemiology are extensive. Research on perceived age (**PA** - how old you look) developed rapidly in the early millennia, setting out guidelines for data collection [108] and modelling [45]. Perceived age is a highly subjective measure, which can be influenced by the observer [17], observation environment [218] as well as the underlying ageing of the subject [109]. Challenges associated with the subjective nature of perceived age have been addressed by using the consensus of judgments by multiple observers. Recent works have shown that the use of 10 or more annotators produces a highly stable mean average perceived age [108]. Perceived age is important to dermatological and gerontological science because it provides a low cost, non-invasive, biomarker which is easy to gather. In addition to being non-invasive, it can also be gathered by unskilled workers in a home or non-hospital care environment. These features alone make PA an interesting area of research in both biological and computational science, though we go on to justify its importance further. To the best of our knowledge the reliability of PA annotations has not thoroughly investigated, and agreeability between groups of annotators may not be high enough to support clinical applications. Furthermore, the time and cost associated with the group annotation approach limits the scale of derivative research as well as the practicality in a clinical setting, and hence there is a need for reliable automated PA annotation methods.

Deep Learning (DL) has been developing in parallel to perceived age research, showing some notable advancement in recent years [192, 90, 31, 203]. Deep Learnings success is largely thanks to its rich feature learning capabilities when compared to classical machine learning [24]. Deep Neural Networks (DNNs), the underlying algorithmic structure powering deep learning, are composed of many sequential layers of artificial neurons. A problem formulation which is highly efficient to compute on parallel processing devices such as graphics processing units (GPUs) [59]. This high efficiency allows for orders of magnitude more data to be processed per unit of time compared to classical machine learning algorithms, facilitating the training of large scale DNNs and subsequently improving their robustness.

When it comes to computer vision, image processing and face analysis, Convolutional Neural Networks (CNNs) are the overwhelming majority choice of DNN architecture. CNNs are able to efficiently learn features from images thanks to their significantly reduced number of parameters when compared to standard (fully connected) DNNs. Many different CNN formulations exist with various benefits and specific applications, in recent years CNNs have been challenged by vision transformers [71], but this has yet to become a mainstream choice. In this thesis we focus on formulating perceived age estimation with CNNs as an Ordinal Regression [214] problem. Perceived age can be formulated as both continuous and ordinal regression, depending on the resolution of the age labels. Integer labels inherently promote ordinal regression as individuals with the same label can be of different ages by design. Perceived age labels are often real numbers resulting from the mean average of a number of integers.

Deep learning is nothing without data, and in most cases there is a direct correlation between the size of the dataset and the generality of the model resulting from it (assuming the dataset is unbiased) [14, 22]. Several public datasets exist for chronological age estimation but only one for perceived age. As perceived and chronological age are highly correlated, state-of-the-art approaches make use of both types of label to train groundbreaking perceived age predictors [192]. Several other perceived age datasets have been produced to support epidemiological research, The University of Southern Denmark, Leiden University, Erasmus University and Unilever have all produced their own perceived age face datasets. These private accessible datasets come with extensive terms and conditions, designed to respect the privacy rights of the subjects and the intellectual property rights of the organisation. Such terms create a challenging research environment, often requiring physical presence in order to access them. However, the unique nature of lab grade datasets is equally as interesting as it is challenging, providing novel opportunities to train models which would be impossible with the current state of public data. A key novelty of the lab grade datasets is their inclusion of medical, and detailed demographic data, allowing research to extend to cross-sectional analysis of facial ageing and its association with health.

The application of deep convolutional neural networks to perceived age prediction is a non-trivial problem and requires careful consideration for data [154], preprocessing [149, 91, 31] and model design [32]. Allowances must be made for the real-world performance of such models as well as the quality of underlying ground truth labels. Under the correct configuration, deep learning is the best suited algorithmic approach to extracting perceived age from face images, but significantly more work is required to justify its use in a healthcare context.

In summary, deep learning is well suited to the task of perceived age prediction thanks to its robust feature extraction and large scale learning capabilities. Existing works in computer vision exploit general CNN advances to improve PA prediction, as well as proposing problem specific model formulations to reach and exceed human level accuracy. Biological research linking perceived age to health, genetics and mortality is limited by institutional resources available, who struggle to justify the recruitment of annotators. This thesis focuses on the intersection between deep learning based perceived age estimation and epidemiology, exploring various associations between morbidity, mortality, genetic markers and PA. In the next section we explore the challenges associated with deep learning based PA estimation, providing a motivation for our research aims and objectives.

1.2 Problem Statement

The progress of current biological research on perceived age is limited due to the bottleneck of human annotators, leading to a need for a faster and cheaper automated solution. Given that human annotated perceived age has been shown to be associated with several age related morbidities, as well as longevity, there is a clear motivation to research this endpoint in depth. Put simply, the problem we aim to solve is the accelerated, automated, annotation of perceived age in face images datasets. We identify 3 barriers to the expansion of research on PA and ultimately its application in a clinical setting: cost, latency and reproducibility.

The cost and latency barriers are driven by the same feature of perceived age, its requirement for multiple judgments to achieve a stable mean. Group based annotation involves hiring a group of 10 or more observers and presenting them with all face images in a given dataset. At presentation, observers must assign the face to a bracket, which can be a single year or up to 5 years wide. This process is not only labour-intensive, but requires high levels of focus and persistence as every annotator must view every image. For extremely large datasets it is possible to allow each annotator to view only a subset of images, but in this case more annotators must be hired to achieve a minimum of 10 annotations per image.

In a research setting this process is cumbersome and costly to say the least, but in a clinical setting the barrier manifests differently. In a clinical setting latency is a key factor, as for a single image to be graded it must be viewed by the same number of annotators. One option would be to have a team of annotators on standby to receive and rapidly annotate images before returning them to the contacting healthcare professional. While this setup would be suitable for a consistent rate of annotations, this does not reflect the reality of patient admissions and screening requirements. For this reason PA annotation must be low cost and low latency to be valuable in a clinical setting, where it may be useful for screening geriatrics [107] or evaluating the success of plastic surgery [244].

The reproducibility of perceived age is more of a scientific concern than a practical limitation, given that existing research shows both significant and reproducible results regarding the association of PA with health [171], mortality [110] and genetics [155]. We call attention to the interaction between the subjective nature of perceived age annotations and the diversity in sampling of annotators. The subjectivity of human perception for any endpoint has been studied extensively, with a small subset of this work being focussed on the subjectivity associated with making judgments from another human's face [108, 176, 219, 58, 120, 95, 18]. Many factors external to the observer can bias the judgement of PA: is the subject wearing makeup, is the environment well lit, is the lighting colour such that it hides wrinkles and blemishes? Aside from external factors, the observer's demographic plays a significant role in their perception of age. In general observers can more accurately predict the ages of faces who are closer to their own age. Ethnic background also plays a key role, where observers are better at estimating the age of faces from the same origin as them. This ethnic bias is not genetic, as individuals who spent their entire lives around one ethnic group are better at predicting their age than the one they are most closely genetically related to [58]. General cultural differences also exist, such as the affinity of Japanese people for accurate age prediction, which has been hypothesised to stem the cultural importance of age in their society [18]. Other biases exist and will be explored later in this thesis, but the given examples are sufficient to demonstrate the challenges of balanced sampling when selecting a group of observers for PA annotation.

Cost, latency and subjectivity can all be solved by automating the process of PA annotation using some model, in our case a deep CNN. Cost and latency are simple to solve by replacing a team of annotators with a single algorithm which can return predictions at sub-second latency on consumer hardware. Subjectivity is solved by distilling the opinions of a group into a single model. If this model does not involve any random process, the outputs for a given image will be identical every time it is queried, creating perfect reproducibility. While it is true that deep learning models produce highly reliable outputs, they are subject to learning their own biases and confounders. Furthermore, not all models are implemented using identical algorithmic structures, such that two different DNN architectures may learn entirely different features given the same dataset. Even if identical architectures and training processes are used, inaccuracies may be present due to different implementations of the model at training and inference time, which may be unavoidable.

In some cases deep PA prediction models are published without details of the exact preprocessing used during training (see Table 2.11). Community users must then implement their own preprocessing which may not exactly align with that of the original authors, leading to biases or inaccuracies. Another unavoidable reason for the difference in preprocessing is resource constraints, as some applications of PA prediction may require models to run on hardware that limits the implementers choice of preprocessing. We see preprocessing for PA prediction and in fact all other face analysis as a key gap in the research, limiting the comparability and applicability to current work.

One of the most significant challenges posed by algorithmic PA prediction is managing bias induced by imbalances in class labels. Class imbalance is a defining feature in both public and private age datasets, present not only in PA but also CA datasets [9]. Class imbalance is present in public datasets due to their sampling criteria, face images are scraped from internet sources such as IMDB, Wikipedia, Flickr and Google Images, all of which have a bias toward young/middle aged individuals. Even datasets not utilising web scraping are subject to class imbalance, such as the Morph2 dataset [189] which is composed of mug-shot images from the United States penal system, carrying the same biases as incarceration rates. Private lab-grade datasets also suffer from class imbalance to a lesser extent, stemming from natural imbalances in the age of the human population and sampling criteria. While the cause of imbalance is different between datasets, the impact on downstream models remains the same, poor quality predictions for minority classes. The poor performance on minority classes is a significant limitation to the application of such models in geriatric science, as minority classes are always extreme, leading to poor performance in the elderly.

1.3 Aim and Scope

From a high level this thesis aims to tackle the challenges associated with the application of deep perceived age prediction models to a medical setting. The overarching goal is to develop deep CNNs with the ability to extract relevant health and genetic information from face images via the proxy of perceived age. To reach this goal several concrete technological challenges must be addressed, research questions must be answered and qualitative judgments must be made. We break down this goal into objectives which are grouped into three phases, somewhat representing a pipeline: face preprocessing, modelling PA and analysing health associations.

1.3.1 Face Preprocessing

Preprocessing is a key component of any advanced machine learning pipeline, at the least normalising/standardising data points to some known range, at most exploiting the semantics of the domain to transform the inputs to some canonical state. Face image analysis is no different, with various degrees of preprocessing complexity seen across the literature [32, 115]. We see face image preprocessing as one of the most significant research gaps in the field, with far reaching effects on both academia and industry. Preprocessing limits reproducibility in academia, weakening the scientific value of works that push forward the state-of-the-art in face image analysis. In industry, preprocessing can often close the gap between academic findings and real world performance, aligning uncontrolled privately collected datasets with those available publicly for research purposes.

Our work on preprocessing focuses on one key facet, face alignment. Face alignment is a broadly accepted practice in face image analysis, aligning training and inference time faces such that the maximum overlap of facial features. As the positioning of a face in the image frame is irrelevant to the perceived age of the individual, we view the alignment process as an extended form of normalisation, e.g. minimising the meaningless differences between data samples. Face alignment is implemented in most if not all leading PA analysis pipelines but lacks consistency and reproducibility. We design experimental objectives to address these weaknesses:

1. Critically compare leading methods for face alignment.

2. Systematically evaluate the impact of mis-matching train and inference time face detection algorithms within face analysis pipelines.

3. Develop a framework for standardising the operation of face alignment algorithms, allowing for flexibility in underlying implementation while guaranteeing consistent face positioning.

4. Validate our framework against deep regression models under a comprehensive set of datasets.

In sum, these objectives allow us to get closer to a consistent and reliable perceived age annotation pipeline. They provide insight into possible advances in face image analysis without specific consideration for the modelling portion of the pipeline, making the findings applicable in a general context. We present the methods and results related to these objectives in Chapter 3.

1.3.2 Modelling Perceived Age

Perceived age prediction using CNNs operates identically to CA prediction in most cases, such that both tasks exploit the same architectures and loss functions. Most recent advances, particularly those making a significant improvement on age and perceived age prediction benchmarks, present either a new loss formulation or a new network architecture. The other most notable focus in perceived age modelling is transfer learning (TL), which was first applied to age estimation early in 2015 [159, 192] and has been a foundational component of all SOTA methods since.

We define objectives to explore the formulation of transfer learning in the context of PA estimation. Lack of data, class imbalance and efficiency are the main facets of transfer learning we aim to address. We also assess a novel technique for reusing features from fixed CNN models, showing that training via gradient descent based algorithms is not the only valid approach for face analysis.

1. Identify the optimal transfer learning strategy for PA prediction.
2. Explore the power of fixed CNN feature extractors in the context of perceived age.
3. Investigate the impact of class imbalance when training deep CNNs for PA prediction.

Addressed in Chapter 4, these objectives answer many research questions regarding PA modelling and drive toward more accurate and less biassed predictions. Improvements made in the modelling stage can be combined with those in preprocessing to propose models suitable in a biomedical setting.

1.3.3 Health Associations

Given our foundation of work in preprocessing and modelling, the final body of work required to validate our proposed models is to apply them to lab-grade datasets with associated health and genetic labels. This can be summarised as aiming to show that deep learning predicted perceived age has the same associative power with health as group annotated human PA does. This aim is important as the subjective nature of PA makes it difficult to reason about scientifically. However, when interrogated in the context of healthy ageing, PA becomes a measurable factor for several accepted scientific endpoints. As the clinical dataset we access contains images from a non-standard 3D camera, we introduce a novel semi-supervised learning approach to improve the consistency of predictions on unseen data. We breakdown our approach into the following objectives:

1. Develop an automated rendering, preprocessing and modelling pipeline for non-standard 3D face images.

2. Apply a novel semi-supervised learning technique developed using public PA data to a lab-grade dataset, evaluating its effectiveness in this unusual setting.

3. Compare the associations between deep learning predicted perceived age and 5 well known morbidities.

4. Compare the association between deep learning predicted perceived age and the MC1R gene, replicated in a new dataset.

Each of these objectives is designed to answer a pertinent research question regarding application of automated PA estimation in a health and wellbeing setting, and are addressed in Chapter 5.

1.4 Significance of the Study

To introduce the significance of this study we first broadly review the context and aims; highlight the relevance of deep learning to ageing and medicine; identify gaps in the literature and review the broader impact of the work from applications and ethics perspectives.

1.4.1 Broad Review of Context

Perceived age is an interesting biometric with far reaching potential associations with health and genetics, but current research is limited in scale and scope due to the cumbersome nature of gathering annotations. Deep learning is a strong candidate to replace the annotation process for PA as it has been shown in recent years to reach and even exceed human accuracy in age estimation. Many facets of deep learning can influence the accuracy of PA prediction models, including preprocessing, model architecture, loss formulation and training regime; all of which are worth investigation. Previous works applying DL to PA prediction validate their models using standard accuracy metrics, such as measuring the MAE of predictions compared to the human ground truth in a validation and test set. We see a need for an objective validation metric, which accounts for the subjectivity associated with human perception. Several age related morbidities and one specific gene have been shown in biological studies to be significantly associated with human PA. Given morbidity and genetics are entirely non-subjective measures, we use their association with deep learning predicted PA as an objective measure of its validity in a health and wellbeing setting. Given an algorithm for the extraction of biomedically viable perceived ages, we see the potential for broad impact in both epidemiology and dermatology.

1.4.2 Relevance of Deep Learning

The extraction of age and perceived age from face images is a highly complex task to solve with machine learning, especially when compared to many other popular applications. Facial ageing comprises a complex and highly variant set of features, presenting differently in different genders, ethnicities and cultures. Perceived age judgments cannot be made using simple colour, texture or shape features, as many object recognition tasks can; but instead requires a combination of features which may present uniquely in each individual. DNNs, and specifically CNNs are well suited to learning complex combinations of image features thanks to their hierarchical structure and large parameter capacity. Architecturally CNNs have been suited to PA estimation since early in their development, but in recent years they have become particularly suitable thanks to improvements in their efficiency and the computational power of the devices on which they run. CNNs have reached a point in their development where high performance models can be run for inference tasks using consumer grade CPUs. This is relevant in a medical context where privacy is the utmost concern because patient data cannot be transferred to a high powered off-site machine. Medical analyses must take place within secure networks to avoid the chance of data being leaked to untrusted individuals. Small deep learning models could even be installed on patients' own mobile phones, allowing for 'at-home' analysis of their own face. We hope that the findings of this thesis will motivate researchers to further investigate deep learning as a potential tool for accelerating healthcare and biomedical research. Given the relevance of the topic, we look for gaps in the existing literature that may need to be filled to allow to propagation of deep learning technology in healthcare.

1.4.3 Research Gaps

As this thesis is interdisciplinary by nature, it is somewhat designed to create its own gap in the literature. To the best of our knowledge no existing research attempts to apply DNNs for PA estimation in a health or wellbeing setting. It is our opinion that this gap is driven by a simple lack of data, such that the correct collaborations have not yet been established between machine learning and biological researchers. To refine the previous point, no works have attempted to test if deep learning estimated PA holds the same links with morbidity, mortality and genetics as seen with human PA [171]. Nor has any work proposed a pipeline for reproducible PA estimation using 3D face captures.

Many works have developed algorithms for deep learning based CA/PA prediction, with development continuing on CA prediction to the present day. The broader advancement of deep learning for computer vision continues to impact PA predictions models with more efficient and accurate CNN architectures. Self-supervised learning has been shown to significantly improve some approaches for age prediction such as deep regression forests, however we see a distinct lack of semi-supervised approaches. Semi-supervised methods learning has been shown to improve the accuracy on widely used benchmarks such as Imagenet and CIFAR10 [224], none of which have ever been applied to face age prediction. The application of techniques such as Unsupervised Data Augmentation to PA prediction as a clear gap in the research which we aim to address.

Following a comprehensive review of the formulations of PA prediction in the literature, we see face alignment preprocessing as a clear gap to be addressed. This is not to say that there is a lack of face alignment, but instead we see rampant inconsistency. This inconsistency makes it difficult to critically compare the contributions of works, as well as making them difficult to replicate in industry. Another feature implemented inconsistently in leading works is transfer learning, which has shown to be possibly the most crucial step in training PA estimation models. Transfer learning can be formulated in several ways, such as general-to-specific or specific-to-specific [159]. Transfer learning effectively allows the model being trained to repurpose features from a larger more diverse dataset, providing more robust prior knowledge of the domain. As both pre-training and fine-tuning datasets for perceived age prediction are sampled from roughly the same class distribution, their labels share the common feature in having a large gap between majority and minority classes. This gap may reduce the efficacy of transfer learning to alleviate the biases caused by class-imbalance.

Straying away from the mainstream thinking around PA estimation we see an opportunity to exploit fixed CNN feature extractors, which have produced remarkable results in gender [208] and facial beauty prediction [226]. CNNs trained for face recognition learn to extract features from datasets which are three or four orders of magnitude larger than those used for perceived age, giving them significantly more robust capabilities. We term any approach using a fixed CNN feature extractor to train classical ML models as 'deep feature transfer' (**DFT**), which to the best of our knowledge has never been applied to age or perceived age prediction. The naive application of DFT to PA prediction is a clear gap in the research, with more nuanced extensions such as experimenting with image resolution and principal component analysis during the feature extraction stage. Such a wide range of classical ML methodologies can be applied to the DFT approach that we see both opportunities for our work as well as many for future work.

1.4.4 Broader Impact

We see both significant short term implications of the work presented in this thesis as well as the potential for extensive long term impact. In the short term many of the current limitations of PA annotation are addressed, facilitating epidemiological research and lowering the barrier to studying intrinsic ageing. At a longer timescale we see the potential for this technology to improve screening in healthcare and identify specific instances where facial analysis may lead to better patient outcomes.

Short-term

Our research aims are designed to specifically target the current limitations surrounding the application of PA in a health and wellbeing context. As such, the potential impact of any developments made is clear. Reliable automated PA prediction allows for larger and more comprehensive cross-sectional studies to be carried out with the only requirement being facial photographs. Removing the need for groups of human annotators makes longitudinal studies on ageing easier and more reliable by capturing patterns of ageing in a single immutable model. Our work on the validation of computationally predicted perceived age against health and genetic measures provides a strong foundation for further research, highlighting the possible benefits of interdisciplinary projects combining deep learning with medical disciplines. From a purely computer science perspective, our development of preprocessing and transfer learning methods can be applied to CA prediction or other face analysis tasks, supporting further research and applications. In summary, the short-term impact of this thesis is largely academic, calling attention to new avenues of research and supporting further work in this area.

Long-term

Over a longer time-scale the clinical application of automated PA estimation becomes significantly more viable. As larger datasets are produced containing more diverse samples of the global population, CNNs for PA prediction will become less biassed and more trustworthy. With the addition of explainability and interpretability techniques, researchers may be able to understand exactly which facial ageing features are associated with most prevalent age related diseases, further supporting the level of trust placed upon the technology. Under the assumption that PA prediction is trustworthy and reliable in a given population, its integration into primary and secondary care facilities becomes a viable approach to screening patients for age related degeneration. Such screening may have a significant impact on patient outcomes if morbidities are detected earlier than at present. One key morbidity is Osteoporosis, which is often discovered in an emergency setting when a patient experiences an unusual fracture. Early detection of Osteoporosis may prevent the need for emergency care and may allow elderly individuals to avoid the quality of life cost associated with serious fractures at old age. Preventing hospitalisation is just one example of the benefit of early screening, it is likely that many more will be uncovered as the body of research develops.

While the link between PA and systemic ageing is arguably the most impactful application of our work, cosmetic and superficial skin analysis are also noteworthy applications. Many individuals assign significant value to the youthfulness of their appearance, investing in both preemptive and reactionary approaches to pause or reverse their appearance of age. Deep learning based PA assessment could be used in an ongoing fashion to measure the rate of ageing in healthy individuals, giving a second order metric for the efficacy of any preventative measures used to slow ageing. The same technology could be applied to measure the impact of rejuvenation procedures like plastic surgery or botox, giving the patient a measure of success more closely tied to the desired outcome.

1.5 Overview of the Study

In writing this thesis we aim to cover 3 main areas of the perceived age estimation pipeline: preprocessing, modelling and links with health. In Chapter 2 we introduce terminology, address the existing literature for deep learning based age estimation and provide background to help the reader better understand the motivation behind the project. We then move onto the first part of the pipeline in Chapter 3, exploring the relevance of well design face alignment and its impact on the performance of face analysis models. Building on this work in preprocessing, we reuse our findings to support a range of CNN training experiments in Chapters 4 and 5. We first clarify unknowns surrounding the optimal transfer learning formulation for PA in Chapter 4, going on to explore the benefits of semi-supervised learning in a novel lab-grade dataset in Chapter 5. Finally we present our discussion and future work in Chapters 6 and 7 respectively.

Each of the methodological chapters (3-5) begins with a brief introduction and ends with a summary. The summary provides a recapitulation of the work completed and acts as a brief discussion. We pull out the most interesting discussion points from each chapter and present them together in Chapter 6, which stands to motivate the future work in the final chapter. An indexed overview of our specific contributions is shown in Table 1.1. We also include a table of the publications generated during the writing of this thesis in Table 2.12 along with their various statuses.

Table 1.1: Indexed overview of contributions in this thesis.

Chapter	Location	Contribution
3	Section 3.2	A unified framework for face alignment.
	Table 3.8	Remarkable performance of the iResNet18 backbone.
	Table 3.7	Identification of MTCNN as the most robust face alignment backend.
4	Section 4.3	A novel approach for exploiting CNN features at high resolution.
	Section 4.6.1	The first work toward optimising PA estimation in the elderly.
	Section 4.6.1	State-of-the-art performance in the APPA-REAL dataset.
5	Section 5.2.2	Automated 3D Face Preprocessing Pipeline.
	Section 5.3	Novel Semi-supervised Learning Method for Age Estimation.
	Section 5.21	Replication of the MC1R association with PA in a new dataset.

Table 1.2: Publications generated during the writing of this thesis.

Title	Chapter	Status
Look Me in the Eyes: The Importance of Canonical Face Positioning in Face Analysis Pipelines	3	In Transfer
A Validated Deep Learning based Perceived Age for Accelerated Cohort Analysis	5	In Review
Transfer Learning for Age and Apparent Age Estimation	4	Further Work

Chapter 2

Background

In this chapter we provide the necessary background information needed to fully understand the context of the project. We first introduce abbreviations and clear up ambiguous terminology, then review the technical and biological material supporting this work. We provide a historical review of the development of deep learning as it pertains to computer vision, describing the current landscape of hardware and software supporting it. We then introduce the links between facial ageing and health, presenting existing evidence and exploring some of the more esoteric facets of face age. Building on this background, we review the seminal works in automatic age estimation, including a brief review of classical approaches as well as a more in depth perspective on deep learning approaches. We review two age estimation taxonomies presented in the literature, drawing links between each category and relevant works we study. Later in the thesis we present a chapter focussing on face image preprocessing, which is supported by deep leaning models for face and face keypoint detection. We briefly review the relevant works supporting our face alignment framework, and quantify inconsistencies in the age estimation literature regarding pre-processing. Finally we review two advanced learning methodologies: transfer learning which is a fundamental component in age estimation and a semi-supervised learning technique which has yet to applied in this context. Finally, we summarise and review the numerous dataset available for age estimation, discussing their limitations.

2.1 Terminology

In this section we first introduce a list of commonly used abbreviations in Table 2.1, then clarify two points of contentious terminology.

Table 2.1: Abbreviation Table

Abbreviation	Meaning
OR	Ordinal Regression
DL	Deep Learning
DNN	Deep Neural Network
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
HOG	Histogram of Oriented Gradients
BB	Bounding Box
KP	Keypoint(s)
KL	Kullback–Leiber
MAE	Mean Absolute Error
UDA	Unsupervised Data Augmentation
DFT	Deep Feature Transfer
SSL	Semi-Supervised Learning
CA	Chronological Age
PA	Perceived Age
FR	Face Recognition
DEX	Deep Expectation
DLDL	Deep Label Distribution Learning
DRF	Deep Regression Forests
CV	Computer Vision

2.1.1 Prediction vs Estimation

This section is included to address the ongoing debate regarding the correct use of the words prediction and estimation in the context of deep learning based modelling. It has become standard across the age and perceived age literature to refer to the process of calculating a final value as 'age estimation', a standard we challenge.

In statistical reasoning the difference between prediction and estimation is unarguably clear. Estimation defines the process of calculating some underlying parameter which describes a given set of data, such that one estimates the coefficient of OLS regression. Prediction is the process of using some predefined model to infer the value of some random variable, such as predicting the Y value of a datapoint given the X and coefficient. [97].

In many fields machine learning has adopted these definitions, using prediction in most cases to describe the inference phase, and using the term estimation to describe the process of parameter fitting. If we view deep learning as an extension of machine learning, and machine learning as an extension of statistical inference, the term estimation should remain assigned to its original meaning.

We do however see some inconsistency in the literature, especially when considering interdisciplinary works. The original definition of "predict" according to the Oxford English dictionary is "say or estimate that (a specified thing) will happen in the future or will be a consequence of something". This definition muddies the water with regard to its use in machine learning, where it is most often used to describe the 'prediction' of some already known ground truth label from a set of random variables. Furthermore, when it comes to the extraction of an age value from face images, we mostly see the phrase 'face age estimation' used. It is our belief that estimation has become the standard term for this process because of the classical english dictionary definition "roughly calculate or judge the value, number, quantity, or extent of", which does somewhat accurately describe the process of calculating a rough estimate of the age of a face. If we reconsidered this naming convention in the context of perceived age, the argument for the use of prediction becomes even stronger, as PA is indeed not an absolute value, but instead a true distribution sampled from subjective observers. An individual is associated with a given PA via a probability, not a point value.

We postulate that "estimation" should be reserved for the fitting of parameters, "prediction" should be used to describe the process of inference from a set of random variables and "forecasting" should be used to describe the prediction of a future event. With this being said, we continue to follow the standard in the literature in this work, referring to the primary process at hand as "face age estimation". We do however believe firmly that this is not correct, and that a terminology bifurcation has occurred in the literature which will need to be addressed eventually.

2.1.2 Apparent Age vs Perceived Age

For the most part apparent and perceived age are used interchangeably to mean the same thing, the age of a face when viewed by a group of human assessors. Apparent age connotes the state of facial ageing without the known biases of human perception, whereas perceived age includes the perceptive biases. We argue that in reality there is no escape from human biases, and indeed there is no measure of apparent age which is not perceived by human assessors. For this reason we choose to predominantly use perceived age in this thesis, but may use apparent age when referring to papers in which it is used.

2.2 Deep Learning for Computer Vision - A Broad Review

In this section we broadly introduce deep learning, covering the foundational components of the high level methodology which are relevant to our work.

2.2.1 History

Deep learning broadly refers to the use of deep artificial neural networks (**ANNs**) to extract information from data. ANNs are constructed from artificial neurons [170] which are organised in layers, each layer passing forward information forward to the next. Deep neural networks (**DNNs**) contain significantly more layers than standard ANNs, hence their name. Both shallow ANNs and the more recent DNNs are trained using a process called back-propagation [133] [193], in which the whole neural network is treated as a differentiable function. First, a forward pass is carried out through the network which outputs a final value, this value is compared to the ground truth label using a given 'loss function'. A loss function is simply a measurement of how close the output of the NN is to the directed output e.g. the difference between the age of a face and the estimated age. The partial derivative of the loss value is calculated with respect to each artificial neuron. These derivates are useful as they allow the backpropagation algorithm to calculate what effect changing the value of a neuron would have on the final prediction, indicating how it should be updated for the final output value to be closer to the desired one. ANNs and back-propagation are two fundamental components of modern deep learning, but since their invention new developments have allowed NNs to become applicable to complex tasks such as computer vision.

Computer Vision (**CV**) is the study of the computer based extraction of information from images and video. The use of ANNs in CV was originally proposed in the late 1970s [86], but it was not until the late 1980s that it showed promising applications in real-world tasks [146]. In 1989 LeCun et al proposed the Convolutional Neural Network (CNN) in its most basic form, containing only 5 layers and operating on 28px by 28px images. At this point in time CNNs remained shallow for two reasons, firstly because computational power of the time made scaling difficult, and secondly because an unknown ANN constraint which appeared to limit the ability for deeper networks to learn effectively. This constraint was characterised by a student of Jurgen Schmidhuber in 1991, Sepp Hochreiter, who formally showed that deeper networks were prone to exploding or vanishing gradients [123]. Between 1991 and 2012 computer power improved allowing for larger and more complex CNN architectures to be produced, culminating in the release of AlexNet [140], which many credit as the first 'deep' CNN. AlexNet was designed for the task of image classification in the recently released ImageNet [59] dataset, and made use of graphical processing unit (**GPU**) compute power to accelerate training as was first proposed in [187]. Following the success of deep

CNNs for image classification, research efforts intensified to find new and deeper configurations which may bring better performance in computer vision tasks.

The next notable development was the release of VGGNets [204] in 2015, which were at least twice as deep as AlexNet. VGG16 and VGG19 were proposed, containing 13 and 16 convolutional layers respectively, with 3 fully connected layers. VGGNets represent an indicative limit to the depth of CNNs without additional architectural features due to the aforementioned vanishing and exploding gradients problem. To solve this He et al proposed Residual Networks [36], which incorporate 'skip-connections' into the standard CNN layer configuration. Skip connections concatenate the input to a convolutional layer with the output, as the name indicates, skipping out the operation. This additional parameter free flow of information allows gradients to propagate through the layers more effectively, reducing the prevalence of gradient issues. At this point in time, CNNs had largely matured into the highly effective feature extraction tools they remain today. In most cases modern CNNs are largely unchanged from the original ResNet architecture, with most algorithmic developments occurring in the processes used to train these models. In this thesis we focus on the CNN as the primary deep learning architecture for our experiments, however it is worth noting that in recent years a new architecture is beginning to show better performance than the those based on convolution.

In 2020 Dosovitskiy et al proposed the Vision Transformer (**ViT**) [72], which as the name suggests uses the recently popularised transformer architecture [217] to perform the same visual feature extraction tasks as is classically done by CNNs. Transformers are based on attention mechanisms which were first proposed for text translation in [21], but have also been used in addition to standard CNN architectures across a broad range of tasks including age estimation [154, 186]. For an overview of the various implementations of attention in computer vision see [232].

In this section we reviewed history of deep learning as it relates to computer vision, chronicling its development from the artificial neuron through to current bleeding edge approaches for highly complex tasks. We briefly mention the relevance of GPUs to accelerating the training of DNNs, but do not give much detail on the current landscape. In the Sections 2.2.2 and 2.2.3 we describe the current landscape of technology used to support deep learning from both a hardware and software perspective.

2.2.2 Hardware

At the time of writing the benefits of GPUs for training DNNs needs little justification, offering remarkable speedups at both training and inference time. GPU based deep learning is overwhelmingly based on Nvidia GPUs and makes use of the CUDA [118] runtime. CUDA is a software package allowing general purpose algorithms to be run on graphics processing hardware, atop CUDA is an additional library used for deep learning specific algorithms known as CUDNN [42]. Aside from Nvidia, there are two other GPU hardware configurations we see fit to make note of: Apple M-series chips and AMD GPUs, both of which have their own supporting software. AMD GPUs use the ROCm library in place of CUDA and MIOpen [134] in place of CUDNN, both of which are less mature than their Nvidia based counterparts. Apple M-series chips use the 'metal performance shaders' (**MPS**) library to expose their GPU compute to applications in much the same way CUDA and ROCm do. Higher level deep learning libraries interface with MPS, CUDNN and MIOpen to carry out deep learning computation on the GPU, which we discuss in more detail in Section 2.2.3.

The overwhelming popularity of GPUs in reality stems not only from their efficacy in deep learning but also their general applicability to other tasks, allowing them to be mass-produced and sold cheaply. However, this general applicability is a double edged sword, on one hand it reduces cost while on the other hand limits their performance in specific settings. For example, neural networks largely make use of matrix multiplication, for which GPUs are not the most efficient design. To combat this problem this problem NVIDIA introduced tensor cores [167] into their enterprise GPU lineup in 2017, these cores are specially to efficiently compute matrix-multiply-and-accumulate in a single clock cycle. This approach brings significant improvements in speed and power consumption when training DNNs, but is still falls behind the impressive performance application specific integrated circuits (**ASICs**).

The term ASIC refers generally to hardware designed to optimise some existing computational process, which in deep learning is the previously mentioned matrix multiplication. An impressive number of deep learning ASICs are currently on the market [165], all designed to improve some portion of the deep learning pipeline. Some ASICs are sold directly to the consumer as an on-premise hardware installation [165, 128] while others are rented from a cloud provider [1, 129]. Cloud based ASICs are the most accessible option for most researchers, allowing them to dynamically allocate budget for training experiments when large compute is needed. Conversely, on-premise ASICs make sense in an industrial setting where the hardware can be utilised 100% of the time.

Regardless of which hardware is used for the underlying computation, deep learning requires additional software layers to bridge the gap between computational primitives and architectural concepts.

2.2.3 Software

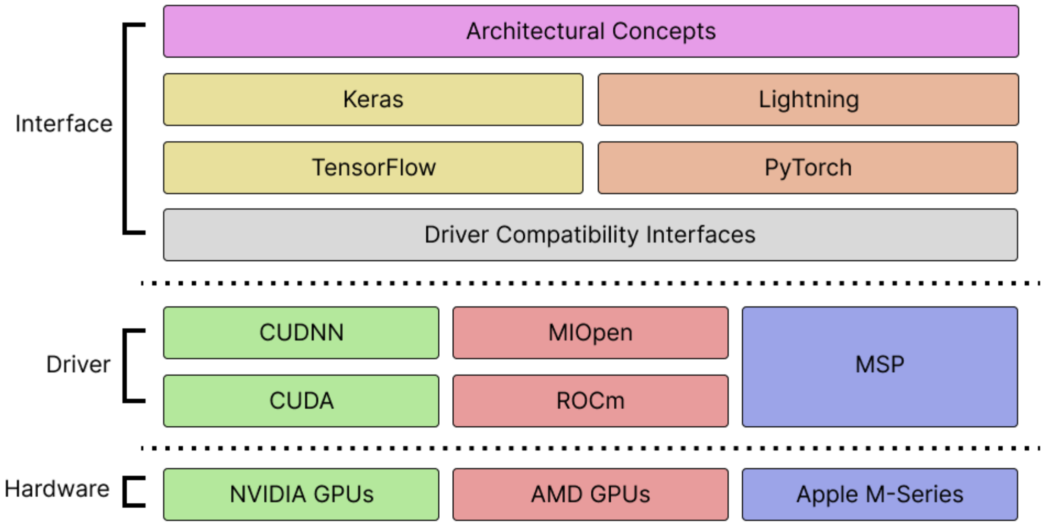
In this section we review the software stacks used to power modern deep learning, offering a pragmatic perspective on the tools needed for cutting edge research. We review primarily the specific software and libraries used in this thesis, comparing them occasionally to reasonable alternatives.

Training DNNs requires a non-trivial amount of compute power, such that basic consumer hardware is limiting in a research setting. For this reason most researchers develop a hybrid workflow involving both a personal computer and a remote server, on which most of the computation is done. This network based hybridisation is powered almost exclusively by the secure shell protocol (SSH) [236], which is widely supported and allows secure remote access to networked machines. SSH is often combined with an integrated development environment (IDE), which automatically carries out tasks such as file synchronisation and application debugging. The IDE we choose for this work is PyCharm [2] thanks to its professional feature set, but VSCode [3] is an extremely popular alternative.

In Section 2.2.2 we mention the need for an interface between deep learning concepts and underlying computational primitives. This gap in the deep learning stack is filled in two layers, a backend interface and high level API. At the time of writing two backend interfaces are noteworthy for their wide adoption and rich feature sets: Tensorflow [4] (**TF**) and PyTorch [181] (**PT**). TF and PT each have their own respective high level APIs (Keras [44] and Lightning [80]), though it is worth noting that PyTorch has a much richer functionality out of the box and is often used in applications without Lightning. Both TF and PT support all three GPU makers we mention previously, making them both highly portable with regard to hardware. A diagram of the complete deep learning stack for GPU based configurations is shown in Figure 2.2, only the packages, drivers and GPUs we mention in this section are included for demonstration sake.

In all our work we use PyTorch, simply based on preference for API design. Tensorflow is equally performant for most if not all tasks, making selecting between them largely a matter of personal preference at the time of writing, though this may change as feature development continues.

Figure 2.2: Diagram of Common Deep Learning Software Stacks



2.3 Linking Health and Perceived Age

For many years researchers have tried to understand the heterogeneous rate at which our bodies age, with the hope that with deeper understanding will come mechanisms for slowing or even reversing physical ageing. The age of an individual can be measured using one of 3 accepted endpoints: chronological age, biological age and perceived age; the definitions of which are not always consistent.

In this thesis chronological age represents number of years lived, biological age describes the relative state of decay of an individual's body when measured by some biophysical parameters, and perceived age means the visual measurement of face age by human or computer assessors. As the rate of ageing is relatively consistent throughout an individual's life, as is the presentation of this ageing on the face, CA/BA/PA are all highly correlated.

Biological age is a concept put forward by the biomedical community to summarise the state of ageing of an individual. It is useful to remove the noise associated with individual biomarkers, allowing for a single holistic measure of health. Specific biological ages are designed for specific purposes, such as cardiac health and mortality [206]; all sharing the common goal of gaining statistical insights into the health of individuals relative to their population [66]. Biological ages are often validated against known health outcomes to measure their efficacy as predictors for morbidity and mortality.

The key challenge with most biological ages is their invasive nature and associated cost of collection. Most biological age markers require blood or tissue samples [161, 206]; or other extensive tests such as frailty [98], cognition [78], eyesight [253] and hearing loss [141]. These measurements take a significant amount of time to retrieve and become cumbersome in the context of longitudinal studies, largely requiring regular visits to a medical or research centre. For this reason, perceived age has been proposed as a new geriatric biomarker, thanks to its potential for cheap acquisition, as well as its ability to capture factors associated with healthy or unhealthy ageing.

Perceived age is used in combination with chronological age to calculate the perceived age residual (PAΔ), which is the difference in years between an individual's PA and CA. A negative residual indicates an individual looks younger than their chronological age, and a positive residual indicates they look older. How old you look for your age is driven by a number intrinsic [225, 155] and extrinsic factors [103], many of which also relate directly to health. Extrinsic factors are numerous and well known, such as sun exposure [84], smoking status [179], diet [188] and stress levels [40]. Fewer intrinsic factors have been identified, but some genetic markers [155] do show a significant association with looking younger or older.

2.3.1 Measures of Ageing

In this section we introduce biological age in more detail to provide the reader with a high level view of the territory surrounding perceived age, as PAs most beneficial use in biomedicine is as a proxy for biological age. We then go on to discuss the link between PA and BA before delving into the factors of PA in more detail.

Biological Age

Simplistically one can say that biological age is the age at which a person's body has aged relative to the time they have been alive, and that perceived age is simply the answer you would get if you asked someone else how old that person looks. In order to confidently approach any problems involving 'perceived' age, we must first understand the intricacies surrounding the different definitions of ageing. In this section we first define the general concept of ageing and biological age, before narrowing focus to the perception of face ageing.

A more advanced measurement of age is 'Biological Age', which is useful to compare the rate at which an individual's body has deteriorated relative to their chronological age. Biological age is used to encapsulate the degenerative changes which occur within the human body over time, using biochemical and biophysical measurements. The study of biological ageing, or senescence, is of growing interest as the percentage of elderly people grows worldwide. Bourlière broadly defines senescence as a series of morphological and physiological changes, or simply the composition and function of an individual. He reviews an extensive set of tests which he states must all fulfil two key criteria [26, p. 32]. Firstly, variation in a measurement must be enough to show a significant change over time. Secondly, the tests should be non-invasive enough to allow for a pleasant experience, making longitudinal studies more possible. Both these points are still relevant at present.

The challenge of composing a value for biological age by combining a series of dimensions is first tackled by Furukawa et al using multiple regression [87]. They model the biological age using only health data from normotensive individuals, and apply the model to predict the biological age of hypertensive patients. They observe a significant increase in perceived age indicating that this is a good measure of biological age for blood pressure.

More recent works study biological age in terms of features such as telomere length [206], DNA methylation [161, 114] and frailty [29, 98]. One such work claims that frailty outperforms DNA methylation as a value for biological age [136]. A very recent paper compares 9 different biological age metrics in a single population of size 854, over 20 years [152]. Their work is of specific interest as its longitudinal nature allows them to show a non-linearity in some biological ages over time.

Perceived Age

When carrying out a dermatological analysis, clinicians have many options in their toolbox. A range of parameters can be taken ranging from biophysical and biochemical measures to visual clinical assessment. We define these categories based on work from 2009 which attempts to model perceived age using linear combinations of parameters which they place into the aforementioned 3 groups [65]. We take great interest in the group that they define as 'clinical', as these are all created using ordinal scales and human perception.

Ordinal rating scales are represented by a preset number of baseline images, each representing a different level of severity of the feature in questions. This parallels the concept of age perception as it uses human perception to categorise a continuous value into an ordinal class. It differs from age perception in that the baseline is abstract, clearly defined and reusable. An interesting benefit of their ordinal nature is that many of the recent technological advancements in neural networks for age estimation would also be applicable to any ordinal classification problem.

Many ordinal scales depicting the severity of skin conditions exist. Some focus on grading the severity of photo-damage [100, 46, 19]. Other photographic scales are used to quantify the appearance of wrinkling. One such scale focuses on the appearance of deep age related furrows [57]. A similar approach was taken to assess hyper-kinetic lines [135] and crows feet [33]. Notably, three papers that feature many of the same authors propose grading systems for the mid, lower and global face [34, 174, 194]. They include metrics for facial 'fullness', which is of specific interest to the current project because to the best of our knowledge this has not yet been solved computationally.

At present it is clear many grading scales exist for facial senescence, and that the process for developing such scales is simply a case of proposal and validation. Due to this lack of an objective truth several works replicate validation experiments, primarily calculating the kappa coefficient to evaluate the quality of a scale. Valet et al review 4 different severity scales for skin ageing [215]. Valet et al return to the topic 8 years later, carrying out a similar experiment again evaluating agreement [216]. In their later work they additionally review both intra- and inter- rater agreement, this is of interest to the current project as computational methods are also subject to the same reproducibility constraints.

A recent paper reviews the territory of photographic scales, stating "while many different photographic scales exist today, most of them have not been validated and/or are not available for general use" [127]. With this motivation they develop 12 different scales with an accompanying software application which assists observers in rating. One key limitation with this work, is the lack of data for male faces, as the focus only on Caucasian women. Further work is required to develop suitable systems for other demographics.

To the best of our knowledge the only work to compare observer based grading systems with digital ones comes from Hamer et al [116]. They compare measures over 7 different signs of ageing, showing moderate to excellent correlation between digital and photo-numeric systems.

Associating Perceived and Biological Age

First we covered biological age, stating any useful form of biological age should be tied to some real world outcome. We then explored how the age of faces is perceived and defined a basic model to formalise our research goals. In the previous section we reviewed many domain specific measures of face ageing, founded in relative states of senescence rather than ageing.

Given our general aim of understanding the facial features which contribute to perceived age, the next logical step is to investigate how different features can be combined to create a singular unified score for the aged appearance of a face. This could be presented as either a 'biological' age or simply a relative score. Some works construct skin ageing from only 'non-superficial' parameters [65, 104], for now we ignore these works and focus on that which can be observed with the naked eye.

Nkengne et al create a PLS model which combines some 19 different visually inspected skin parameters to estimate perceived age [176]. Their model shows that lip volume, skin colour and jawline are the three most significant contributors to perceived age. Nkengne revisited the subject recently, developing their methods by creating different PLS models based on the perspective of the graders age and gender [177].

One work proposes an abstract skin age score and does not attempt to estimate perceived age using it [178]. Taking inspiration from biological age models, we suppose that this score could simply be mapped to a hypothetical age delta to produce a final 'skin age'. A different tact used to understand the effect of facial features on perceived age is to control the subject images and analyse perceived age annotations. For example, Fink et al extract a 2d colour map from the face images of female subjects excluding wrinkles, spots and sagging [83]. They show a significant correlation between the age of the subjects and their perceived age, indicating that skin colouration is important for age estimation. Another approach involving image manipulation followed by perceived age annotation was taken by Porcheron et al, who accentuate/attenuate the prevalence of wrinkles, dark spots, dark circles and sagging [184].

2.3.2 Existing Evidence

In the previous section we review the general literature surrounding biological age and its links to the appearance of age on the face, in this section we focus more specifically on the links between human perceived face age and underlying health conditions.

The known links between the extrinsic factors of ageing and common age related morbidities makes the PA an interesting biomarker for health in the geriatric community. To the best of our knowledge the first study investigating the link between PA and health was the Baltimore Longitudinal study, where individuals (n=1086) were assessed for perceived age by a physician at their first visit. They found that PA is significantly associated with survival, and that men who died during the study looked older on average compared to those who survived. A more recent study on Danish twins replicated this finding (n=1826) showing that PA was significantly associated with survival, even when accounting for chronological age, sex and rearing environment.

Given the ultimate endpoint of mortality does not capture much information relevant to clinical decision making, follow up works aim to link PA to morbidities, which can also give useful insights into which types of disease relate most to facial skin ageing. A cross-sectional study on a population of long-lived individuals (n=) and their offspring showed associations between PA and the framingham cardiovascular disease (CVD) risk score, finding women in the lowest quartile of CVD risk looked more than 2y younger for their age than those in higher risk quartiles. They also find indications of the heritability of perceived age, showing male offspring of LLI looked 1.4 years younger than controls. This finding is supported by the genetic link found in [155] (n=2693), where the MC1R gene was identified as a significant factor in skin ageing. Under the same cohort as in [155], a 2023 work [171] uses additional morbidity data to test for associations with several known geriatric conditions. They show PA is significantly associated with ARHL, cognition, osteoporosis, cataracts and COPD.

2.3.3 Known Biases

We previously mentioned that PA is subject to the many observational biases present in human perception, ranging from demographic biasses to more esoteric factors. In this section we break down perceived age bias into six categories, not all of which are necessarily relevant in a healthcare context but should be considered nonetheless.

Gender Bias

Two works which collect and analyse first hand data on the effect of observer gender on age perception make differing claims. [108] finds a strong correlation between the estimation of both male and female assessors. In contrast to this [176] find a statistically significant difference between the genders, with female graders being more accurate than males. The out performance by women is substantiated in [219], who also find a statistically significant difference between male and female sales people. The reason for this difference is the method for measure of accuracy, where Gunn et al fit the observer as a random effect, compensating for observers who consistently over or under estimate. This method is valid as their goal is to develop a clinical method using multiple observers to gain a reliable mean perceived age. Such that they are not judging the accuracy relative to chronological age but rather judging it between groups of assessors.

Ethnicity Bias

Dehon and Brédart carried out a study to investigate the effect race has on age estimation between Caucasian and African subjects [58]. They showed 72 observers from each racial category colour images of faces ageing from 20 to 45 years old. They took the difference between the perceived age of a face and the real face as the error in perception. As in other works we review, this is likely subject to bias due to biological age differences that were not controlled, e.g. truly differential ageing between races. However, Dehon and Brédart do remove background and surroundings showing a well controlled scene bias. They show a significant increase in error when Caucasians perceive African age over Caucasians, and go on to show that this was mostly due to over estimation. They also find that in their experiments faces tended to be over estimated in general, which detracts from the strength of their results. It is interesting that African observers perform equally well on both races, which they explain by noting that the African observers lived in a predominantly Caucasian country, learning to perceive which faces better. For this reason, we propose that a better term for this bias is the ethnicity bias, reflecting its life experience based causation. The primary and uncontested takeaway from this research is that racial, or ethno-racial, bias is an interaction between subject and observer and is not a fixed observer bias. However, due to the inclusion of two races and a small sample size, further work should be done to strengthen this conclusion.

Age Bias

The age bias in age estimation has been extensively studied. Early work shows a general bias to overestimate the ages of younger individuals [120], which is exaggerated in the elderly [95]. Willner and Rowe find observers consistently overestimated the ages of the 13 and 16 year old's noting that the effect greater for the female subjects, and increased with the age of the observer [182].

Anzures et al directly investigate the effect of the own-age bias has on age perception [17]. This theory states that an observer is likely to be more accurate when estimating the age of face closer to their own. As with the own/other-race bias discussed previously, it is identified that is it likely not due to a specific demographic characteristic of the observer, but is due to their historical experience in face processing. They show 'passport' style photos to a groups of young adults in china, using a circular crop to eliminate hair and background from the images. After taking age estimations they carry out an analysis of variance between 3 subject age groups (child, young-adult, middle-aged), the 2 observer genders and each individual subject face. This showed a significant interaction between the age group of the face images and the individual faces. This interaction is explained by observing the variance in ratings for each group, where a much greater variance is found for young-adults. They believe this variance represents an ability to better differentiate ages. They also replicate their results with a second study in a Japanese population. Further work is required to show evidence that these effects are visible in different observer age groups and races.

As with the ethnicity bias discussed previously, the works of Gunn et al [108] and Nkengne et al [176] are not in complete agreement. Where the former again finds no significant correlation between observer age and estimation bias, the latter find younger graders (under 35 years) are more accurate than older (over 50 years). Nkengne et al use a linear regression model and show a significant correlation between observer age and estimation error. These differences are due to the slight difference in their measures of accuracy, where Nkengne et al use chronological age as their ground truth, Gunn et al use the mean perceived age from other observer groups.

Consequence Bias

We have previously cited the works of Vestlung et al [219] as their investigation into the age estimation abilities of alcohol sales people contains many insights into other biases. The bias discussed in this section is name consequence bias as it stems from the perceived risk associated with incorrect age predictions. Risk should theoretically create an increased learning rate as observers will take the task more seriously throughout their daily lives, meaning they are likely to develop the skill more than someone who cannot quantify the value of a correct prediction. An earlier study [182] also investigates the age estimation ability of alcohol sales-people in response to reports that underage drinking is 50% self facilitated. Both studies are not recent which makes them of greater interest as the effects were likely lessened by the check 21 and check 25 acts. These acts reduce the burden of correctness on sales-people by simplifying their task, instructing them to always request identification for individuals who look remotely young. Both studies attempted to directly estimate chronological age as legality is their key outcome, and neither study compensates for the scene effects of hair or clothing.

Vestlund et al compare their experts in age estimation with a control group and show a significant difference in bias. Both groups overestimated age in general but the control group did so to a greater extent. Willner unfortunately does not test a control group and does not publish a mean or SD for their data, making it difficult to compare with the work of Vestlung.

Anzures et al identify a link between social expectation and age perception abilities, noting that Japanese participants performed better in their study of 3 Asian populations [18]. These populations were Chinese, Japanese and Asian-Canadian. Their methods combined accuracy and speed measures to assess the perception of age, giving participants a maximum of 10 seconds to decide which of two faces were younger, as task which reflects introduction to a new social group. They found that Chinese and Asian-American observers were faster than Japanese. However, when scaling speed relative to accuracy, Japanese participants performed significantly better. They go on to hypothesise that this difference is due to the greater importance of respect for age in Japaneses society, which we would classify as a consequence.

2.4 Age Estimation from Face Images

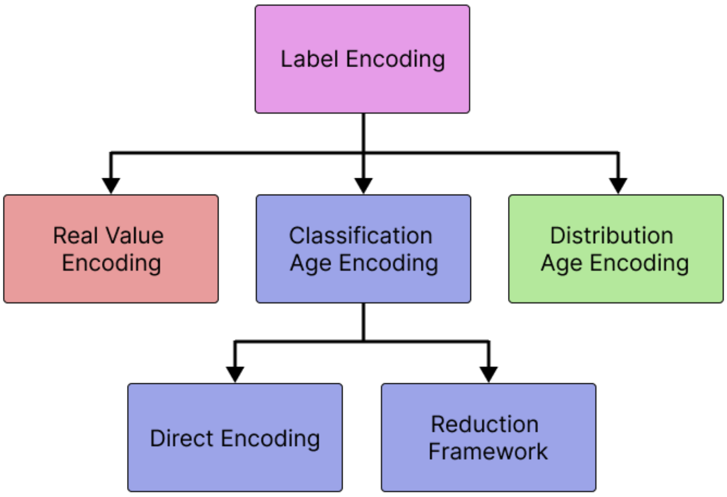
In the previous section we discusses the links between perceived age and health, commenting on the broad range of biases induced in PA estimation by the observer. Given the clear motivation for an automated, reliable and unbiased age estimator, we turn to a technical literature to review possible solutions. In this section we first review the high level problem formulation of age estimation using computational methods, presenting two taxonomies used to describe the algorithmic facets of the problem. We then review both classical and deep learning works which attempt to extract age values from face images, providing a foundation for our own work.

2.4.1 Problem Formulation

As with many machine learning problems, age estimation can be formulated in different ways depending on the perspective of the researcher and the desired inductive bias. There are 3 high level ways to view the problem: regression, ordinal regression (**OR**) and classification. Classification and OR share the property of treating age as a single integer value, differing only in that the OR formulation accounts for similarities between neighbouring classes. The plain regression formulation treats age as a continuous floating point value, assuming correctly that ages can exist between integers. We do however question the validity of tthis formulation, as the difference in facial appearance between an individual at two neighbouring ages is small if not indecipherable in many cases.

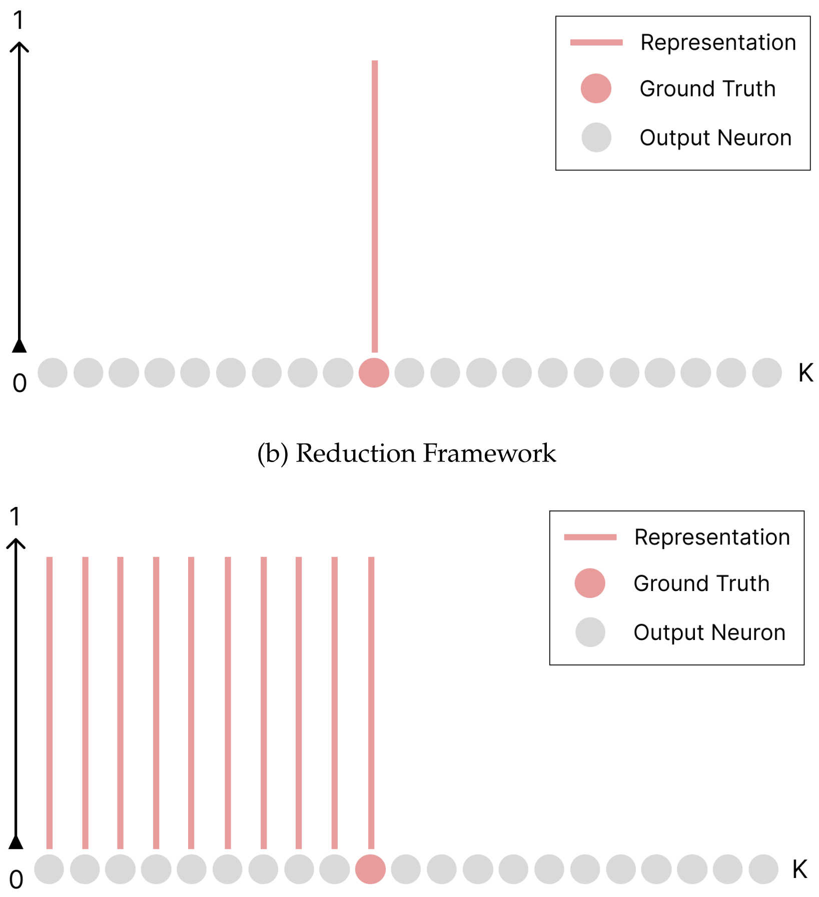
The formulation of age estimation is indeed far more complex than simply viewing it as a regression or classification problem, with complex taxonomies being proposed in recent years to make sense of the burgeoning literature on the subject. We call attention to two recent survey papers, the first [32] proposing a taxonomy for label encoding in age estimation and the second [8] proposing one for algorithmic design.

Figure 2.3: Taxonomy of Age Label Encodings from [32].



One of the key features differentiating approaches to age estimation is label encoding, which has developed significantly since its inception. Figure 2.3 shows the recent label taxonomy presented in [32], breaking down methods into 3 high level categories: real value encoding (**RVE**), classification age encoding (**CAE**) and distribution age encoding (**DAE**). RVE as the name suggests represents approaches where the label is encoded as a real floating point number, in contrast to CE and DE which encode the label as a vector of length K where K is the number of possible ages. CE has two sub-categories, direct encoding (CAE-DE) and reduction framework (CAE-RF). CAE-DE represents the labels as a one-hot encoding as if found in most classification problems, whereas CAE-RF represents the label as a series of binary classifiers indicating if the labels is greater than or equal to their value. To elaborate, CAE-RF uses a thermometer like encoding where the vector is filled with ones up to the age it represent, following which all elements are zero. DAE differs from CAE in that it assigns a value between 0 and 1 to each element in the vector, producing a gaussian-like representation centred around the label value. Each label encoding presents its own strengths and weaknesses which stem from the algorithmic designs they originate from, which we discuss later in this section. Visual representations of each encoding are shown in Figures 2.4, 2.5 and 2.6.

Figure 2.4: Classification Age Encoding.



(a) Direct Encoding.

(b) Reduction Framework

Figure 2.5: Real Value Encoding

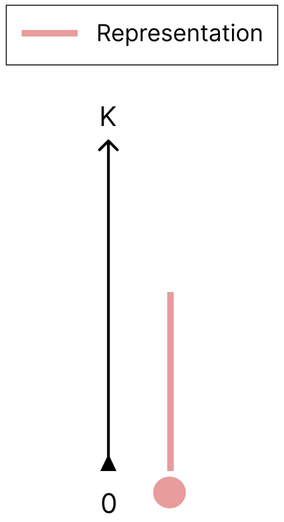
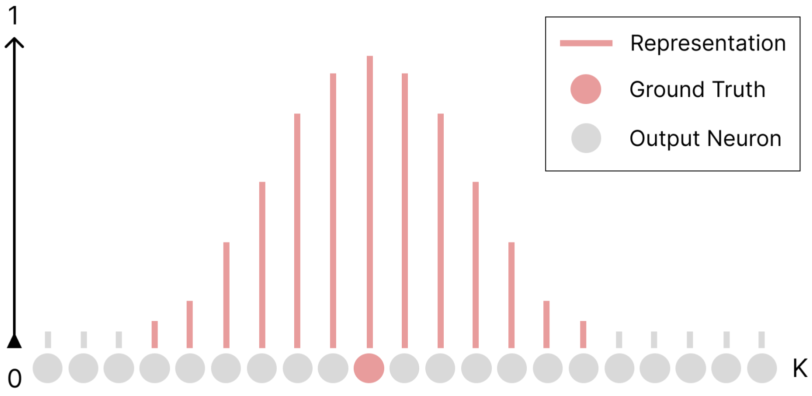
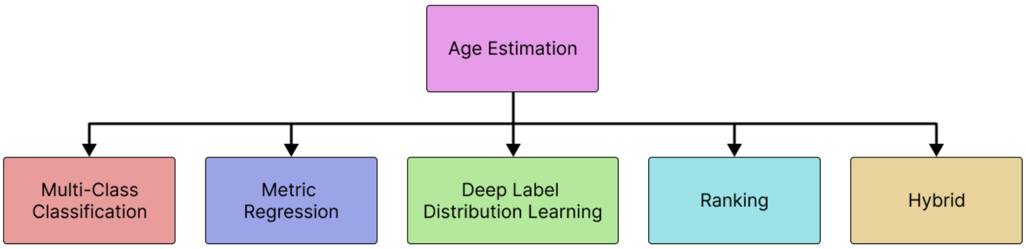


Figure 2.6: Distribution Age Encoding.



This label encoding taxonomy is well representative of what we see in the literature, capturing all the variation seen in both recent and historical approaches. The second taxonomy we review attempts to capture the variation in algorithmic implementations for age estimation, but in our opinion has a number of weaknesses. Figure 2.7 shows the breakdown of algorithmic approaches presented identically in both [8] and [9]. There are clear links between their algorithm classes and label encoding mentioned previously, such that multi-class classification (**MC**) links to CAE-DE, metric regression (**MR**) links to RVE and deep label distribution learning (**DLDL**) links to DAE. Ranking and Hybrid approaches are flexible with respect to the label encoding used, in some cases using multiple encoding within a single model.

Figure 2.7: Taxonomy of Age Estimation Algorithm Design from [8].



In this thesis we do not explore all algorithmic and label formulations in depth, instead choosing to work entirely with DAE thanks to its superior performance (discussed later in Section 2.4.3). We do however review the other approaches briefly for the sake of completeness, as well as to point out weaknesses in the current taxonomy.

Multi-Class Classification

MC is the general process of classifying an input into one of three or more classes, which in deep learning is most often represented using the one-hot encoding. In the context of age estimation this maps directly the the CAE-DE encoding, where the output neuron associated with the ground truth label is expected to have the highest value. A number of works use the MC approach [82] [166], showing it is suitable for the task of age estimation, it is however becoming less popular in recent years, giving way to more complex and tasks specific formulations.

Metric Regression

References to the term MR are few and far between in the age estimation literature, indicating it is not a particularly popular approach. The survey authors describe the MR process as finding the optimal mapping from the age-value space to the feature space using the appropriate regularisation, citing [247, 144, 112] as examples. MR uses the RVE label encoding and can be paired with the hyperbolic tangent activation function to speed up convergence as described in [91].

Deep Label Distribution Learning

Deep label distribution learning (**DLDL**) is a subset of label distribution learning (**LDL**) which includes DNN as the main algorithmic component. LDL for age estimation was originally proposed in [93] who describe the way in which a label distributions represent the degree to which each class represents a face. DLDL was first introduced in [233], which formulates an age estimation loss function using the KL-divergence between the label distribution and the CNN output. DLDL was developed further in [90] using pre-trained models and in [89] with the addition of several novel architectural features. As we make use of the DLDL formulation extensively in this thesis, we review these works in more detail in Section 2.4.3.

Ranking

Ranking approaches make use of the ordinal relationship between classes, often making use of loss formulations that take into account the relative ranks of faces to preserve ordering in predictions. One example of this is [156] which present an end-to-end topology preserving framework combining two different ranking approaches. Their implementation contrasts to that presented in [31], who use the CAE-RF label encoding with train a series of local binary classifiers with a novel weighting scheme to alleviate class-imbalance issues. It should also be noted that a recent state-of-the-art approach uses a comparative ranking scheme to reach previously unseen accuracies in several benchmark datasets [203].

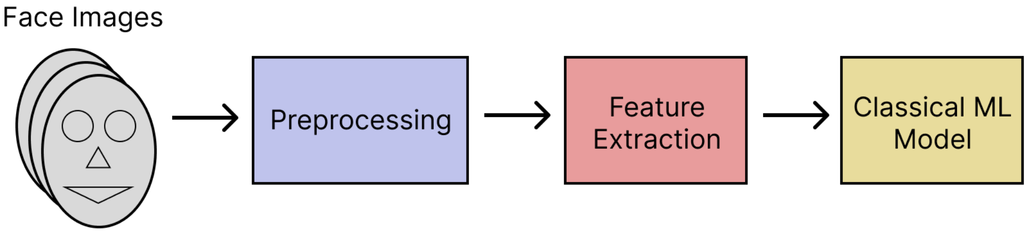
Hybrid

Hybrid approaches are not commonplace in recent literature, mostly being popular prior to the deep learning revolution in age estimation. Classical approaches ensemble two or more formulations to improve robustness such as in [112] and [43]. One recent deep learning based approach [63] achieved remarkable results combining features from DLDL with considerations for ordinal ranking, giving some credence to hybrid approaches in the current day.

2.4.2 Classical Approaches

The history of automatic age estimation from face images extends significantly further back in time than the methods which involve deep learning, with the first work being presented in 1999 [142]. These non-deep approaches, which we refer to as 'classical', are generally formulated as a three-stage pipeline. This pipeline first preprocesses the face, then carries out feature extraction on the preprocessed face, finally those features are used to train a face age estimator; as show in Figure 2.8.

Figure 2.8: General Formulation of Classical Approaches for Age Estimation



One of the main disadvantages of this classical formulation is its heavy dependence on preprocessing, requiring significant domain knowledge to transform the face into a format from which features can be extracted effectively. Many believe that deep learning removes the need for well designed preprocessing, an assumption we challenge in Chapter 3. Early preprocessing formulations involved processes such as feature localisation [142], mean-differencing [144] and face-region extraction [94]; all of which operate on grey-scale images.

Feature extraction is host to the most variance and complexity in the classic approaches we review, attempting to address the challenge of transforming an input face into some low-dimensional coding scheme. In [142] a variety of features are extract using facial landmarks to calculate facial distance ratios in addition to their own algorithm for wrinkle measurement. Following this seminal work a variety of approaches were proposed involving principal component analysis [144], biologically inspired features [111], gabor wavelets [124] and local binary patterns [105]. For a more comprehensive list of classical age feature extraction approaches see [117][85].

The final step in the process is to fit some classifier or regressor to the extracted features. Like in deep learning, the problem is formulated as either regression or classification, with different algorithms supporting each approach. As early as 2004 neural networks were applied for classification of age from age feature representations, but did not outperform other methods such as fitting simple quadratic functions [144]. Support vector machine based classifiers and regressors have shown great popularity in recent years [43], with research on their application to age estimation continuing to the present day thanks to their low computational cost [13].

In summary, a wide range of classical approach have been proposed for age estimation, mostly prior to 2014 at which point deep learning began to outperform all previous methods [234]. Much can be learned from the classical formulation of age estimation, with clear links between their original methodologies and modern formulations. Deep learning based approaches still require significantly more computational power than those based on hand-crafted features, such that older methods still remain relevant to the present day when resources are severely limited.

2.4.3 Deep Learning Approaches

Convolutional Neural Networks have become the de facto approach for age estimation since they were first applied in [234], though some may point to [230] which was announced in 2013 but was not published until 2015. Whether it is a CNN used purely for feature extraction or a complex highly task specific architecture, NN are involved in all record-breaking publications since its inception. Many datasets have been used for age estimation over the years, which are summarised in Table 2.12. The results in [117] indicate human annotators are able to estimate chronological age on the FG-NET [39] dataset with a mean absolute error of 4.7 years. The first work we find exceeding human level accuracy was published in 2017 [82]. However, it is likely this barrier may have been broken as early as 2015 with the creation of the IMDB-WIKI [192] dataset. Since then, the MAE reported on the FG-NET dataset has been reduced by over half to 2.23 [203], far exceeding any reasonable estimate of human abilities.

The rapid advancement in deep learning for age estimation is largely supported by the robust feature extraction capabilities of CNNs, but also in part by the ongoing novel reformulations of the problem. In the remainder of this section we review the key developments in age and perceived age estimation from face images, covering groundbreaking works, works presenting novel formulations and limitations of these works. We finally discuss the intuition behind the problem formulation we use in this thesis and outline the most current developments at the time of writing.

Seminal Works

Deep learning based age estimation begins in 2014 with the multi-scale CNN proposed in [234], who present a shallow architecture with multiple feature extractors designed to operate on different face patches. Patches are extracted using facial keypoints annotated with ASM [51]. Cropping face regions allows the CNN to operate at only 48x48px, requiring significantly less computational power than in the higher resolution CNNs of today. Even with such a low complexity model they were able to show an impressive MAE of 3.63 on the Morph2 dataset, far exceeding that of all previous classical approaches.

In 2015 the adoption of deep learning for computer vision was accelerating at a rapid pace, with advancements from research on object recognition making their way into the task of age estimation. This is also the year of the first Chalearn LAP Challenge on apparent (percevied) age estimation

[79], driving many researcher to switch focus to this task. A myriad of works attempted to train perceived age regressors in the small, challenging, competition dataset. We briefly review the top-3 entires [192, 159, 252] based on their test set performance and discuss how their innovations lead to the research landscape we see today.

All three top performing works made use of deep transfer learning, [159, 252] using face recognition [235] as their initial pre-training and [192] using object recognition. Following initial pre-training, all 3 works make use of chronological age datasets to learn strong priors for the final task of perceived age estimation. The only CA data used in [192] is the IMDB-WIKI (n= 500k) which they scrape themselves and remains to this day the largest publicly available CA dataset. [159] also scrape CA data from internet sources but at a significantly smaller scale (n= 10k), instead incorporating existing benchmark datasets [77, 172, 37, 189] into their transfer learning pipeline. All three approaches use deep CNNs, [192] using VGG-16 [205] and [159, 252] both using the GoogLeNet [209] architecture, though in [159] they add batch normalisation and remove all dropout layers.

Regarding label encoding and algorithmic class [159, 252] both use hybrid approaches, combining metric regression and multi-class classification, making use of the RVE and CAE-DE encodings respectively. The exact approaches used differ significantly, [159] combined MR and MC in an ensemble fashion, whereas [252] first classify a face into one of 10 age brackets, then use SVR to locally regress a final value. In contract, [192] uses a much simpler learning scheme, training the model only for MC using the CAE-DE encoding. The novelty of their approach stems from the introduction of the expected value of the output distribution at inference time. This approach allows features from neighbouring classes to contribute to the to the final predicted value, in much the same way as in the DLDL formulation. These three approaches clearly indicate that transfer learning is a key component of perceived age estimation, and that ensembling when used correctly can improve the robustness of predictions. The success of the approach presented in [192] in our opinion as based on the large chronological age pre-training they employ, not in fact related to any other feature of their formulation.

Since 2015 many works toward CA estimation have been published with fewer toward the task of PA estimation, likely because there are a range of CA benchmarks and only one for PA (the aforementioned Chalearn 2015 dataset). Each work generally aims to improve one facet of the estimation process, such as model size, accuracy on a given benchmark or training complexity.

Promising Avenues

There are several distinct promising research avenues in the age estimation literature, making use of entirely different algorithmic formulations. We call attention to comparative, distribution learning and deep regressions forest approaches, all of which attain state-of-the-art performance. Comparative approaches under our definition make use of exemplar faces to regress the age of a face by finding the interval in which it site, e.g. younger than X and older than Y. Distribution learning is one of the simplest and most effective approaches, training the CNN to output a predictive distribution of ages and using the expectation of that distribution to regress a final value. Regression forests, or more specifically deep regression forests use an unusual network architecture where the output of the CNN backbone is fed into a regression forest which is trained with gradient descent.

Deep regression forests (**DRFs**) were first presented in [201], who propose a scheme where the CNN and forest split nodes are trained together with back-propagation, then the leaf nodes are optimised in a separate step. They show remarkable performance for the time on the FG-NET and Morph2 datasets, achieving MAEs of 2.17 and 3.85 respectively. Later, the concept of self-paced learning (**SPL**) for DRFs was proposed in [11]. Self-pacing allows the model to learn from easier samples first, with harder samples being gradually introduced as training progresses. This feature further improved the MAE on FG-NET to 3.44 and on Morph2 to 1.99. This formulation was improved once again in [180] with the addition of consideration for under-represented samples. The authors note the basic SPL formulation leads to under-representation issues in the task of age estimation, where the distribution of classes is highly imbalanced. This imbalance leads to hard-under represented samples being ignored for most of the training process, producing bad solutions in minority classes. They propose a method that considers minority classes when ranking the difficulties of samples, ensuring that they are proportionality represented during training. The addition of consideration for under-represented samples reduces the quoted MAEs once again to 2.77 (FG-NET) and 1.91 (Morph2). The primary limitation these DRF implementations is their complexity, in terms of both computational requirements and the number of components. All the DRF based works require the full VGG-16 backbone to reach state-of-the-art accuracies, which to this day remains the most highly parameterised CNN used in age estimation. The need to optimise the two network portions independently slows down training, and the self-pacing scheme requires more steps can a standard supervised learning pipeline. From an implementation perspective, the need for regression forests makes the model less portable between deep learning libraries, making compatibility on less-common systems challenging. Overall we see the power behind the intuition presented in these works, but do not see their value as worth the complexity trade-off.

Comparative approaches operate on the intuition that it is easier to compare between two faces than it is to regress the age of a single face directly. Anecdotally we see this intuition as true in human perception, where human annotators are likely to find ordering a set of faces easier than accurately guessing an exact age, though we put this idea to the reader to consider. Works using the comparative approach are not common, but since a recent implementation showed ground breaking results in both CA and PA prediction, we see them as an extremely promising approach moving forward. To the best of our knowledge the first work using a comparative framework for age estimation is [147], who train a CNN using both a regression and comparative loss. Their model simply classifies if one face is older than another, and at inference time multiple forward passes are carried out over a range exemplar images until the bracket in which this input face resides is found. More recently [203] present their ground-breaking moving window regression approach, which operates under similar principals as [147] but breaks down the inference stage into local and global regressors, first finding a broad age bracket and then refining the ranking with a local regressor trained on only data from that age group. They achieve previously unseen MAEs on FG-NET (2.23) and Chalearn 2015 (2.95). The primary limitation of this class of approaches is the complexity associated with training and inference. At train time pairs or triplets of faces must be selected to train the comparative portion of the network, at inference time multiple iterative comparisons must be made involving multiple forward passes. This contrasts from the standard supervised learning formulation where only a single forward pass is needed to retrieve results.

The final formulation we review is in our opinion the current most promising approach to age estimation from face images when all factors are considered. DLDL was first proposed in [233] as an entry to the aforementioned Chalearn 2015 challenge, coming in fourth place by a small margin. This approach was developed further in [90] but reached its most mature state in [89]. The latest iteration of DLDL is named DLDLv2 by the original authors, and makes use of an extremely lightweight CNN with a novel pooling mechanism. The CNN they propose is a thinner and shallower VGG16 model, with fewer layers and fewer filters per layer. This reduction in size allows for a 36x reduction in parameters and 2.6x reduction in inference time compared to standard VGG-16, making their approach one of the most accessible from a computational standpoint. They also introduce a novel pooling formulation they name hybrid pooling (HP), which is the sequential combination of max-pooling and global average pooling. They claim this pooling strategy produces more discriminate features than either pooling method independently, showing ablation results to support this. In a later work they go on to show the generality of their approach from a task perspective [91], achieving strong results in facial beauty and head pose estimation as well as CA and PA.

2.5 Other Relevant Face Analysis Tasks

As this thesis does not have any objectives directly related to training models for other face analysis tasks, we review works which are involved in or closely related to aspects of age and perceived age prediction. This review acts largely as a foundation for out work in face preprocessing which is presented in Chapter 3.

2.5.1 Face Detection

Face detection algorithms localise human faces within 2D images outputting a series of bounding boxes, one for each face. In some cases these bounding boxes are near perfect squares, where others find the closest rectangle to the face in question. Machine learning based face detection has been studied for over two decades, notable early works include ASM [51], AAM [53] and the still popular Viola Jones approach based on boosted feature cascades or 'Haar Cascades' [220]. CNNs [169] have proven to be more robust than these classical methods for face detection, especially in more challenging settings. Recent works focus on predicting bounding boxes in a 'single-shot' fashion [245, 212, 158], improving the efficiency of deep approaches and allowing models to run entirely on accelerators such as GPUs.

We see a clear trend away from classical and hybrid algorithms toward single shot, end-to-end deep models for face detection. This trend extends to the associated task of face landmark detection, where recent advances focus on both detecting faces and keypoints with a single neural network [61].

2.5.2 Face Landmark Detection

Face landmark detection, also called fiduciary- or key- point detection, involves labelling geometric facial features using machine learning algorithms. Early methods for face detection can also be applied to landmark detection, [51] and [53] output landmarks as a side-effect of shape fitting. [220] does not directly predict facial landmarks but can be trained to detect facial semantics such as eyes, nose and mouth; from which landmarks can be inferred.

Most recent approaches utilise CNNs for face landmark detection, but we call attention to the use of Ensembles of Regression Trees [132], specifically the implementation in [138] which is still one of the most popular approaches thanks to its high speed and robust performance. We also make note of the CNN based method proposed in [242] as the most popular deep method in the literature we review. While [242] has held this title for many years, we see [61] coming to the forefront of the field thanks to its single stage design and highly efficient GPU based inference.

2.5.3 Face Alignment as Preprocessing

Face alignment has been long accepted as crucial preprocessing step in face analysis pipelines [221]. In the overwhelming majority of of cases face alignment is implemented as a 2D transformation calculated from facial landmarks. Some exceptions exist such as the detector response approach used in [192], where images are processed multiple times at different rotations, the rotation with the strongest detector response is considered best aligned. 3D face alignment has been shown to improve face recognition accuracy [126] under high pose, occlusion or challenging lighting. However, 2D Face alignment from landmarks proves sufficient when used as a preprocessing step for face age, perceived age and beauty analysis. Likely due to the nature of datasets in these tasks, where faces are singular, frontal pose, generally un-occluded and well lit. While landmark based alignment has been seen in the literature since prior to the deep learning revolution, it has not become commonplace until recently. Many papers still do not include any details of the procedure used for face alignment in their own pipelines, while those that do often do not include enough information to accurately reproduce their results. Some papers can be commended for including details of their alignment implementation while others release their code allowing for perfect reproducibility.

Table 2.9: Face detection and alignment components from a sample of 51 papers.

(a) Face Detectors

Face Detection	Papers
Unclear	17
MTCNN [242]	11
HeadHunter [168]	9
Dlib [138]	7
Haar Cascade [220]	5
Retinaface [61]	2

(b) Keypoint Detectors

Keypoint Detection	Papers
Unclear	19
MTCNN [242]	11
None	7
Dlib [138]	6
ASM [51]	3
Retinaface [61]	2
AAM [52]	1

(c) Alignment Procedures

Alignment Procedure	Papers
Unclear	17
No Mention	10
Box	8
Upright Pose	4
Eye	3
Orthogonal Procrustes	3
Similarity Transform	3
Other	3

In Table 2.9 are a quantifications of Table 2.11, which was generated by a manual review of 51 papers gathered largely from SOTA benchmark table on the 'papers with code' website. Additional papers were included from the Newcastle University library search using keywords such as 'age prediction' and 'beauty prediction'; publication date filtering was also used sample a balanced number of papers in recent years. In all tables 'Unclear' means the original authors mentioned the concept but do not include enough information to usefully categorise their approach. In Table 2.9b 'None' is used when the authors clearly state no keypoints were used. 'No Mention' in Table 2.9c represents papers where the authors do not give any indication of the alignment procedure used.

While not all details are always clear regarding the alignment procedure used in a given pipeline, three main veins exist in the literature we review: Box Based, Eye Based and Similarity Based.

Box Based alignment simply involves cropping the face using some pre-defined bounding box. This bounding box can come directly from a face detection algorithm or can be defined by the extremities of detected facial landmarks. In most case box based alignment calculates a square frame to match the input shape of the vast majority of CNNs.

Eye Based alignment focusses on the intra-ocular line, using the assumption that human eyes are both placed at equal vertical positions on the face, such that the line between them is parallel to the x-axis when the face is in an upright pose. While we cannot be sure due to lack of detail, we assume the term 'Upright Pose' in [90, 89, 238, 91] describes the process of aligning the face such that the eyes are level with the horizon, and as such can be classified as 'Eye Based'. Eye based alignment can be extended with further assumptions about the relationship between the image frame and the position of the eyes within a frontal face, this is commonly referred to as 'Canonical Positioning'. We follow chains of citations and ultimately uncover [20] which is to the best of our knowledge the earliest public implementation of canonical eye based alignment. The only recent paper we review which provides implementation details of their canonical procedure is [25], which includes parameters for vertical eye positioning but not for eye distance.

Similarity Based alignment minimises the distance between two sets of face landmarks using some optimisation algorithm. Orthogonal procrustes is a popular choice, aligning faces with standardisation followed by rotational correction with SVD. [150, 203, 38] describe their alignment process as a similarity transform whereas [149, 154, 81] share their underlying approach (Orthogonal Procrustes). We review code from those papers which share their implementation and note that they are based heavily on a non-academic article from 2015 [75].

The heterogeneity of face alignment across the literature makes it difficult to compare the results from different papers, while also increasing the difficulty with which future work can replicate results. We appreciate the need for individual works to focus on their unique methodological contributions but conversely see a need to standardise the face alignment preprocessing used in this field.

Table 2.10: Sample of papers training age, perceived age or beauty prediction models from face images, with associated alignment components.

Task	Paper	Face Detection	Keypoint Detection	Alignment Procedure	Year
PA	[192]	[168]	None	Detector Response	2015
PA	[233]	[168]	Unclear	Unclear	2015
Age	[6]	Unclear	Unclear	No Mention	2016
PA	[16]	[168]	None	Box	2016
PA	[166]	[138]	None	Box	2016
PA	[166]	[168]	None	Box	2016
Age	[175]	[220]	[52]	Nose Center	2016
Age	[90]	[168]	Unclear	Upright Pose	2017
Age	[119]	[168]	None	No Mention	2017
Age	[147]	Unclear	Unclear	No Mention	2017
Age	[207]	[168]	None	Box	2017
Age	[248]	Unclear	Unclear	No Mention	2017
Age	[89]	[242]	[242]	Upright Pose	2018
Age	[31]	[138]	None	Box	2019
Age	[31]	[138]	[138]	Eye	2019
Age	[125]	Unclear	Unclear	No Mention	2019
Age	[149]	[138]	[138]	Orthogonal Procrustes	2019
Age	[150]	[242]	[242]	Similarity Transform	2019
Age	[156]	[138]	[138]	Unclear	2019
Age	[210]	[168]	None	Box	2019
Age	[238]	[242]	[242]	Upright Pose	2019
Age	[241]	Unclear	Unclear	No Mention	2019

Multitask	[5]	Unclear	Unclear	No Mention	2020
Multitask	[7]	[168]	[132]	Unclear	2020
Age	[12]	[242]	[242]	No Mention	2020
Age	[15]	Unclear	Unclear	No Mention	2020
Age	[67]	Unclear	Unclear	No Mention	2020
Multitask	[91]	[242]	[242]	Upright Pose	2020
Age	[102]	Unclear	[132]	Unclear	2020
Age	[148]	[242]	[242]	Unclear	2020
Age	[157]	[242]	[242]	Unclear	2020
Age	[160]	Unclear	Unclear	Unclear	2020
Age	[180]	[242]	[242]	Unclear	2020
Age	[63]	[242]	[242]	Unclear	2021
Age	[121]	[61]	[61]	Unclear	2021
Age	[249]	[220]	Unclear	Unclear	2022
Age	[154]	[61]	[61]	Orthogonal Procrustes	2022
Age	[203]	[242]	[242]	Similarity Transform	2022
Beauty	[88]	Unclear	Unclear	Manual	2014
Beauty	[229]	[220]	Unclear	Box	2014
Beauty	[38]	[220]	[51]	Similarity Transform	2018
Beauty	[81]	Unclear	Unclear	Orthogonal Procrustes	2018
Beauty	[226]	[138]	Unclear	Box	2018
Beauty	[227]	[138]	Unclear	Box	2018
Beauty	[240]	[220]	[51]	Eye	2018
Beauty	[153]	Unclear	[51]	Unclear	2019
Beauty	[239]	Unclear	Unclear	Unclear	2019
Beauty	[68]	Unclear	Unclear	Unclear	2020
Beauty	[70]	Unclear	Unclear	Unclear	2020
Beauty	[25]	[138]	[138]	Eye	2022
Beauty	[69]	Unclear	Unclear	Unclear	2022
Beauty	[145]	[242]	[242]	Unclear	2023

Table 2.11: Sample of papers training age, perceived age or beauty prediction models from face images, with associated alignment components.

Task	Paper	Face Detection	Keypoint Detection	Alignment Procedure	Year
PA	[192]	[168]	None	Detector Response	2015
PA	[233]	[168]	Unclear	Unclear	2015
Age	[6]	Unclear	Unclear	No Mention	2016
PA	[16]	[168]	None	Box	2016
PA	[166]	[138]	None	Box	2016
PA	[166]	[168]	None	Box	2016
Age	[175]	[220]	[52]	Nose Center	2016
Age	[90]	[168]	Unclear	Upright Pose	2017
Age	[119]	[168]	None	No Mention	2017
Age	[147]	Unclear	Unclear	No Mention	2017
Age	[207]	[168]	None	Box	2017
Age	[248]	Unclear	Unclear	No Mention	2017
Age	[89]	[242]	[242]	Upright Pose	2018
Age	[31]	[138]	None	Box	2019
Age	[31]	[138]	[138]	Eye	2019
Age	[125]	Unclear	Unclear	No Mention	2019
Age	[149]	[138]	[138]	Orthogonal Procrustes	2019
Age	[150]	[242]	[242]	Similarity Transform	2019
Age	[156]	[138]	[138]	Unclear	2019
Age	[210]	[168]	None	Box	2019
Age	[238]	[242]	[242]	Upright Pose	2019
Age	[241]	Unclear	Unclear	No Mention	2019
Multitask	[5]	Unclear	Unclear	No Mention	2020
Multitask	[7]	[168]	[132]	Unclear	2020
Age	[12]	[242]	[242]	No Mention	2020
Age	[15]	Unclear	Unclear	No Mention	2020
Age	[67]	Unclear	Unclear	No Mention	2020
Multitask	[91]	[242]	[242]	Upright Pose	2020
Age	[102]	Unclear	[132]	Unclear	2020
Age	[148]	[242]	[242]	Unclear	2020
Age	[157]	[242]	[242]	Unclear	2020
Age	[160]	Unclear	Unclear	Unclear	2020
Age	[180]	[242]	[242]	Unclear	2020
Age	[63]	[242]	[242]	Unclear	2021
Age	[121]	[61]	[61]	Unclear	2021
Age	[249]	[220]	Unclear	Unclear	2022
Age	[154]	[61]	[61]	Orthogonal Procrustes	2022
Age	[203]	[242]	[242]	Similarity Transform	2022
Beauty	[88]	Unclear	Unclear	Manual	2014
Beauty	[229]	[220]	Unclear	Box	2014
Beauty	[38]	[220]	[51]	Similarity Transform	2018
Beauty	[81]	Unclear	Unclear	Orthogonal Procrustes	2018
Beauty	[226]	[138]	Unclear	Box	2018
Beauty	[227]	[138]	Unclear	Box	2018
Beauty	[240]	[220]	[51]	Eye	2018
Beauty	[153]	Unclear	[51]	Unclear	2019
Beauty	[239]	Unclear	Unclear	Unclear	2019
Beauty	[68]	Unclear	Unclear	Unclear	2020
Beauty	[70]	Unclear	Unclear	Unclear	2020
Beauty	[25]	[138]	[138]	Eye	2022
Beauty	[69]	Unclear	Unclear	Unclear	2022
Beauty	[145]	[242]	[242]	Unclear	2023

2.6 Advanced Deep Learning Methodologies

2.6.1 Transfer Learning

Transfer learning (**TL**) is an algorithmic process inspired by how humans learn, where knowledge is transfer from one task to assist in the learning of another. The exact formulations of TL are numerous, and it's prevalence in deep learning, especially perceived age estimation, is remarkable. In previous years many attributed the first application of transfer learning to neural networks to the NIPS conference in 1995 [185], however the work in reality began in the mid 1970s [27]. We are unable to access works from either claimed origin, and instead begin our review of TL as it is first applied to DNNs in [122]. Hinton and Salakhutdinov pre-trained an MLP autoencoder using the MNIST dataset before fine-tuning the encoder for classification. They show that this unsupervised form of transfer learning scheme improves classification accuracy without any additional data. As deep learning methods have developed this approach no longer bears fruit.

Instead, transfer learning from general to specific tasks has become popular, especially using CNNs pre-trained for image recognition, which has become the de facto initialisation for most computer vision tasks. The ImageNet dataset, released in 2009 [59], is by far the most commonly used pre-training set today, though in the field of age estimation it is debatable whether or not it has been overtaken by face recognition pre-training.

A number of early transfer learning works take features from a CNN pre-trained on ImageNet and use them to fit classifiers such as SVMs [35]. This formulation is powerful but is seldom seen in leading age estimation pipelines, instead we see the pre-train/fine-tune pattern used, where training is continued using the same model and optimiser, but different data. To the best of our knowledge the first paper using this formulation was [96], in which they first pre-train the CNN for image recognition using the ImageNet dataset, then fine-tune on their own dataset. They replace the 1000-neuron output layer needed for classification in ImageNet with a new randomly initialised 21 output layer, matching the number of classes in their dataset.

It was not long after this that transfer learning became a crucial part of the perceived age estimation literature, as we discuss in Section 2.4.3. Transfer learning for age estimation is generally formulated in one of two ways: CA or Face Recognition (**FR**). We also see some works including ImageNet pre-training but this is not of interest to us due to its prevalence across all the computer vision literature. Chronological age pre-training allows the model to learn features that are highly correlated with PA, whereas FR allows the model to learn features that effectively capture the variance of faces. It is interesting that it remains unclear which pre-training scheme is most effective for PA estimation, with conflicting statements found in the literature. A recent survey [9] concludes that CA pre-training is more effective than FR, which directly contrasts with the findings presented in [159] and [89], who find that FR is the better choice.

2.6.2 Semi-Supervised Learning

In many tasks the lack of labelled data poses an issue for deep learning, especially where labels are difficult or expensive to generate. In contrast, unlabelled data is often easy to access in large quantities, motivating researchers to find ways to incorporate unsupervised methods into the supervised pipeline. The history of semi-supervised learning (**SSL**) is long, and as it is only a small part of this thesis we do not explore it in depth. Semi-supervised learning is generally used when there are too few labelled images to produce a good classifier, or where a wealth of unlabelled images are used to further improve and already good one. A recent survey [231] breaks down deep semi-supervised learning methods into 5 high level groups: generative methods, consistency regularisation (**CR**) methods, graph based methods, pseudo labelling (**PL**) methods and hybrid methods. We find consistency regularisation and pseudo labelling methods of particular interest thanks to their relatively simple implementation and notable performance on mainstream benchmarks. From these two approaches, we select CR as the most promising approach for the problem of perceived age estimation. CR and PL operate in much the same way but under different assumptions, both techniques involve the use of predictions on unlabelled data to improve the networks response in the unlabelled domain. However, CR usually utilised data augmentation under the assumption that the networks output should remain consistent regardless of reasonable input perturbations. PL in contrast requires a highly performant teacher model to generate pseudo-labels on unseen data. PL is most suitable when the teacher-student pattern is used to distil knowledge from a large high quality model to a smaller one, with applications such as low latency or embedded processors in mind. The problem of perceived age estimation from faces current suffers from a general lack of performance, especially in minority classes, such that PL likely does not bear much fruit.

Consistency Regularisation

CR SSL is powered by reasonable assumptions about a models invariance to some kind of perturbation, which in most cases involves augmenting inputs, but also commonly uses other diversity techniques such as dropout and ensembling. One notable early work [196] designs a SSL framework for image recognition with CNNs. Their approach uses N augmented versions of the unlabelled images to calculate a consistency loss based on the MSE between pairs of final layer outputs. Noise is also injected into the model via dropout and random pooling, adding further variance for the model to learn to ignore. They combine the unsupervised loss with a supervised mutual exclusivity loss [196] to force the classifiers predictions to have only one non-zero element, stating that this naturally compliments the unsupervised loss.

A later work challenges the need for multiple forward passes on unlabelled data, showing that predictions from previous epochs can be aggregated to serve the same purpose [143]. They, like the previous approach, use the MSE loss for unlabelled samples, but instead of including multiple version of the same image in each batch, use the outputs on each image aggregated over a number of previous epochs. A weighting system is applied such that more recent predictions have the most significance in the loss, mentioning that they find a gradual warm up of the unsupervised loss component is key to prevent the network getting stuck.

A recent and popular work presents unsupervised data augmentation [224], a semi-supervised learning technique that challenges the need for model perturbations such as dropout. Their approaches leverages carefully designed augmentations to train consistency into DNNs using unlabelled data. Their approach differs from those mentioned previously as it only augments one of a pair unlabelled images, measuring loss between augmented and unaugmented versions. The KL-divergence between the outputs for each pair is taken as the unsupervised loss, while the supervised loss is a simple cross-entropy over the softmax normalised outputs of the model. While augmentations are key to their approach, making use of recent developments in augmentation for supervised learning [54, 55], they also implement several other techniques to improve the rate of convergence and avoid over-fitting. Namely, training signal annealing (**TSA**), sharpening and confidence based masking (**CBM**). Sharpening is a simple process used to reduce the entropy in the classifiers output [99], which involves dividing them by a small number (between 0 and 1), namely the temperature and passing them through a softmax. CBM is used to apply the consistency loss only to samples where the highest probability in the classifiers output is above a certain level. They find that this approach is dataset dependent, using a different threshold depending on the task at hand. Finally, TSA is used to prevent overfitting when there is a large gap between the sizes of labelled and unlabelled data. TSA defines a threshold as a function of training process which only calculates loss on labelled data with a maximum predicted probably lower than its value. This is to say that as training progresses, more labelled samples are gradually introduced, beginning with the samples the model is least confident about and finally the ones it finds easiest. This approach somewhat resembles self-paced learning as is used in [11, 180]. It is unclear if these additional features presented in the UDA approach are valid for age estimation, especially in our cases where the model is trained to output a predictive distribution.

2.7 Sources of Data and Their Limitations

In this section we review the face image datasets with accompanying age annotations that are available to use in this project. We first review the numerous face age datasets made accessible for academic research, then go on to discuss the data available to use privately through our industry collaborator, Unilever.

2.7.1 Public Data Sources

There are a wide range of publicly accessible datasets for face age estimation, with only two featuring perceived age [79, 10], the latter of which is simply an extension of the former. In Table 2.12 we provide and overview of the majority of face age datasets used in the literature to date. The majority of datasets were released between 2015 and 2017, a period that was kicked off by the rapid development of deep learning and interest generated by the Chalearn-LAP challenges. Early datasets such as FERET and Lifespan are seldom used in current research, being found mostly in papers using classical machine learning. FG-NET and Morph2, both released in 2004, are still the most commonly used age estimation benchmarks to date, likely due to their reasonably controlled environment and good range of demographics. Datasets without an age range specified in Table 2.12 ([183, 77, 250, 131]) only contain labels for age 'group' such that they are not useful for our work which requirers more accurate estimates. Image resolution is relatively similar in all datasets, ranging from 64px to at most 1000px. AFAD has many extremely low resolution images, making it an unusually challenging benchmark given its low age range. Label noise is also known to exist in web scraped datasets such as CACD and IMDB-WIKI, such that IMDB-Clean was released to fix incorrect labelling in the IMDB-WIKI dataset. The most common secondary labels in age datasets are identity and gender, likely because these attributes are easy to access during data acquisition.

Across all datasets image resolution is a key limitation to skin ageing research, which in many cases does not surface in the deep learning literature due to their shared low operating resolution. For example, most face age datasets do contain images of roughly 224px, which is the standard modern operating resolution of CNNs for classification and regression. It is clear however that at this resolution many fine-grained face age features cannot be seen, making it impossible for current CNNs to learn from them.

Regarding PA datasets, sample count is the main limitation. As we see datasets such as IMDB, AFAD and CACD containing well over 100k images, there remains fewer than 10k for perceived age in total. This feature of the data landscape has driven the advances in transfer learning for perceived age estimation, but there is still a significant need for more PA data, which is in part being addressed by research institutions.

Table 2.12: Public Sources of Face Images with CA/PA annotation.

Name	Year	Type	Reference	Age Range	Images	Additional Metadata
FERET	1998	CA	[183]	-	14,126	
LIFESPAN	2004	CA	[172]	18-93	1046	Expression
FG-NET	2004	CA	[39]	1-69	1002	
MORPH-2	2004	CA	[189]	16-77	55,134	Gender, Identity, Ethnicity
FACES	2010	CA	[76]	19-80	2052	Expression
WebFace	2012	CA	[250]	-	59,930	
Adience	2014	CA	[77]	-	19,370	Gender, Identity
CACD	2015	CA	[37]	16-62	163,446	Identity
LAP 2015	2015	PA	[79]	3-85	4699	CA
LAP 2016	2016	PA	[10]	0-95	7591	CA
IMDB-WIKI	2016	CA	[192]	1-100	523,051	Gender, Identity
AFAD	2016	CA	[175]	15-40	160,000	Gender
UTKFace	2017	CA	[246]	1-116	24,100	
AgeDB	2017	CA	[173]	1-101	16,458	
MegaAge	2017	CA	[248]	2-69	41,941	
All Age Faces	2019	CA	[41]	2-80	13,322	
FairFace	2021	CA	[131]	-	108.501	Gender, Ethnicity
IMDB-Clean	2022	CA	[154]	1-95	287,683	Gender, Identity

2.7.2 Restricted Data Sources

To address the lack of labelled perceived age data, four institutions have run their own data collection processes with the help of our common collaborator, Unilever. We have the privilege of accessing these datasets but each comes with its own limitation, in some cases making research extremely challenging. All datasets are taken in a highly controlled environment with a white background and good lighting, all images are very high resolution with cropped faces exceeding 500px.

The Unilever China and Spain datasets have not been published in any academic journal, instead used to support internal product research. They are comprised of mostly female subjects and were captured using an identical camera system to the Leiden data. We are able to access both datasets with relative easy and choose to combine them into a high resolution perceived age dataset with around 1000 samples.

The Danish Twins dataset is much larger than Leiden and the Unilever internal datasets, but comes with very restrictive terms of conditions, such that in order to access it we must transport our technology to their university where it is run by one of their scientists. We made initial attempts at this process without but failed to gain any interesting results, such that we choose not to mention it further in this thesis.

The final and most important restricted dataset we access is the Erasmus ERGO study, which was annotated using the same process as both Leiden and Unilever, but used a more advanced 3D camera system. The ERGO dataset is also the largest high resolution PA dataset we are able to access, making it highly promising with regard to extracting a rich feature space. This dataset also contains labels for morbidity and genetics which have previously been associated with human estimated perceived age, making for an exciting opportunity to apply algorithmic approaches in the same context.

Table 2.13: Restricted Access Datasets with Clinical Data

Name	Type	Images	Age Range	Additional Metadata
Unilever China and Spain	PA	250 + 247	25-77	Gender
Leiden Longevity Study [106]	PA	671	39-76	Gender, BMI
Erasmus ERGO Study [171]	PA	2282	49-86	MC1R, Morbidities, Phenotypes
Danish Twins Study [110]	PA	1826	63-90	Mortality

Chapter 3

Preprocessing for Facial Imagery

Deep learning methods have shown in recent years to be very effective at the broad area of face image analysis, such as age and beauty prediction. The task of automated analysis from facial images has reached human level accuracy, but requires a complex pipeline in order to function correctly. An often overlooked area of the face image analysis pipeline is face alignment. In this chapter we review and evaluate a series of face alignment methods in a range of face image analytics tasks, formulated as regression problems, solved using deep convolutional neural networks. Six face alignment methods are compared across five datasets and two convolutional neural network (CNN) backbones, quantifying the impact of face alignment on the overall performance of the pipeline. We empirically show the importance of aligning faces using the eyes as key landmarks, propose a flexible system for eye based alignment and validate our system in several relevant contexts. Our experiments thoroughly identify the strengths and weaknesses of different alignment implementations, as well as the impact when different alignment methods are used at training and inference time. When eye based alignment is used, mismatching in alignment components does not lead to the performance penalties seen with previous methods, allowing for significantly more flexibility for practitioners and consistency for researchers. The findings from this chapter are currently being transferred from their original target journal to another in their roster.

3.1 Introduction

Face image analytics broadly defines the use of technical procedures to extract information from images of the (human) face. These procedures often use machine learning but in recent years most commonly deep learning and convolutional neural networks [192, 31, 213, 92] thanks to their rich feature extraction capabilities. While there is a diversity of analysis pipelines and problem formulations, they all depend on face alignment preprocessing, such that it deserves careful consideration.

Given the breadth of facial analysis tasks, various different architectures, datasets, objective functions and training methods exist. These tasks fall into three broad categories: embedding, classification and regression. This work focusses on face age, perceived age and perceived beauty prediction, 3 well studied regression task in the field of deep learning, allowing for a diversity of datasets to support our findings.

In general CNN learning, an input image may have relevant features positioned at any location and scale within the frame. In some cases the goal is to train the model to output the same result regardless of the positioning of these features. In other cases the input image is aligned using some preprocessing algorithm to normalise the position of features, such that the model in training need not learn to find them anywhere in frame. In the case of face analysis the latter approach is used almost exclusively; face detectors are used to locate the face within the frame [56, 220], and face keypoint models are used to correct the fine grained positioning of the face [243, 30]. In some cases [61, 242, 23] both face and keypoint detection are done with a single algorithm. The influence of this crucial preprocessing step has never been thoroughly investigated in the context of face analysis.

Moreover, a problem that researchers and software developers face is the growing prevalence of pre-trained models without predefined inference pipelines. This leads to a limited capacity to reproduce published results and lower performance when using these models in other related datasets. Given the breadth of choices for face detection and alignment algorithms, the developers of downstream pipelines must often deal with undefined or inaccessible preprocessing code.

The focus of this paper is to perform a thorough and systematic assessment of the influence of face alignment methods on the train-time and inference-time performance of CNN-based face analysis models. We show the consequences of naive alignment algorithm selection using 6 leading alignment methods, interpret the difference between alignment outputs and finally validate the importance of using keypoint based alignment to induce alignment algorithm invariance in deep face analysis.

Overall, the contributions of this chapter are:

1. A critical comparison of leading methods for face alignment.

2. A systematic evaluation of the impact of mis-matching train and inference time face detection algorithms within face analysis pipelines.

3. A framework for standardising the operation of face alignment algorithms, allowing for flexibility in underlying implementation while guaranteeing consistent face positioning.

4. We validate our framework against two deep regression models under a comprehensive set of datasets.

3.2 Unified Face Alignment Framework

We present a unified face alignment framework combining our preferred approach for face alignment with several of the most popular algorithms for face keypoint detection, upon which all face alignment is built. This approach allows for near identical alignment regardless of the underlying algorithm chosen, which is distinct from previous approaches. All previous approaches are either limited to only one alignment backend or output heterogeneous face positions where multiple algorithms are available.

3.2.1 Alignment Backends

In this paper the phrase 'face alignment backend' is used to describe the underlying algorithms used to detect face boxes and keypoints in RGB images. This terminology is useful to make a distinction between backends and procedures, where a face alignment procedure operates above a backend to calculate and carry out image transformation.

In the open source community the DeepFace [200] library proves a good choice for both practitioners and researcher, offering a plethora of architectures and pre-trained models for face analysis. We focus specifically on their aggregation of face alignment backends, making six of the most popular approaches, shown in Table 3.1, easily accessible. An important limitation of their implementation is the alignment procedure chosen, which only corrects for rotation and does not apply any canonical positioning. Faces aligned with different backends consiquently contain small differences in alignment stemming from the inconsistent bounding boxes provided by each backend.

Our proposed method builds on the alignment backends provided in DeepFace by adding a flexible module for alignment procedure abstraction. This abstraction allows for both naive box based alignment and canonical eye based alignment.

Table 3.1: Alignment backend configuration taken from [199].

Backend	Face Detector	Eye Detector
RetinaFace [61]	Original CNN	
MTCNN [242]	Original CNN	
MediaPipe [163]	BlazeFace CNN [23]	
OpenCV [28]	Haar Cascade [220]	Haar Cascade
Dlib [138]	HOG [56] + SVM	Ensemble of Regression Trees [132]
SSD [158]	Original CNN	Haar Cascade

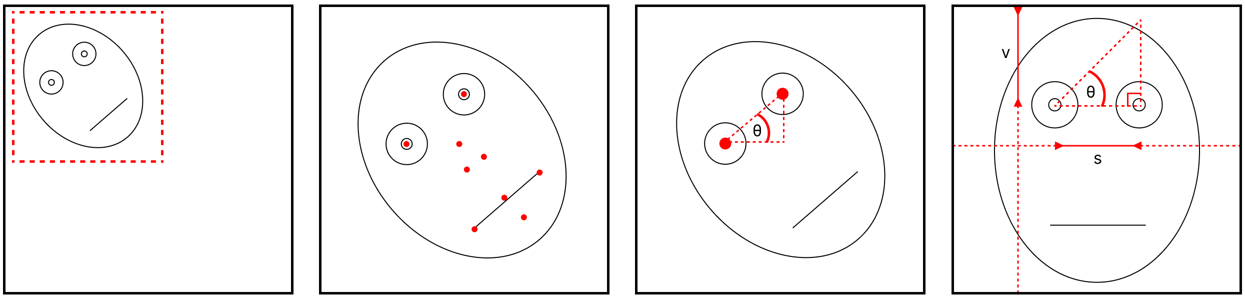
3.2.2 Alignment Procedure

We combine a modified version of the alignment procedure from [20] with the alignment backends from [199] to create a unified canonical face alignment framework. In [20], the alignment is parameterised by the (x,y) position of the left most eye in the frame. Under the assumption that the face should be symmetrical over the vertical line centrally dividing the frame, they are able to deduce the desired position of the right eye, and thus calculate an affine transform based on the current eye positions. We propose a different heuristic operating under S and V parameterisation;

representing 'Scale' and 'Vertical Position' respectively. Figure 3.2d shows the meaning of these parameters in situ. Scale (S) represents the ratio of eye distance to frame size, such that an S value of 1.0 positions each eyes on their respective frame extremities. An S value of 0.0 scales the face to an infinitely small size. Vertical Position (V) defines the vertical position of the eyes in the frame as a float between 0 and 1. A value of 1.0 place the eyes at the top of the frame, and 0.0 places them at the bottom. Using these two parameters alone the face can be positioned for tasks which require a tight crop including only face data, as well as looser crops including hair and background.

The remainder of the procedure operates in the same way as most advanced face alignment pipelines. First the face is localised in the images using bounding box detection (Figure 3.2a), following which facial keypoints are detected (Figure 3.2b). The penultimate step (Figure 3.2c) uses the eye keypoints to calculate the relative angle between the binocular line and the x-axis of the image, representing the rotation required to position the eyes at equal heights in the image.

Figure 3.2: Canonical positioning of face with S and V parameterisation.



(a) Bounding Box (b) Keypoint Detection (c) Eye Angle Calculation (d) Canonical Positioning

3.3 Experimental Design

To quantify the benefit of eye based alignment we compare it to the naive box based approach, where face are simply extracted based on the bounding box from a given face detector. Bounding box based face extraction is biassed to the bounding boxes in the dataset on which the original algorithms were trained. This differs from eye based alignment where there is far less ambiguity. We design experiments to compare the relative performance of eye and box based alignment in a variety of relevant settings, the most important of which being when two different alignment backends are used at training and inference time.

3.3.1 Datasets

We take the datasets used in [92] as our evaluation datasets and we used the dataset from [60] for pre-training (Table 3.3).

Table 3.3: Sizes of datasets and splits.

	Split	Train	Validation	Test
Task	Dataset			
Face Recognition	MS1MV2 [113] [60]	5,822,653		
Facial Beauty	CFD [164]	477	120	
	ScutFBP [223]	3300	2200	
Perceived Age	Chalearn [47]	4113	1500	1978
Chronological Age	Morph2 [189]	39764	4416	11044
	UTK [246]	19279	4821	

This selection of datasets provides a broad range of settings, resolutions and sample sizes to evaluate our methods. For the final regression tasks we include 2 facial beauty datasets, 2 age datasets and 1 perceived age.

Facial beauty and perceived age datasets are labelled with the average rating from a group of observers, giving each label statistical reliability. Chronological age labels are often scraped from the internet or calculated using some estimation based method. The Morph2 dataset is known to have very reliable labels, whereas UTK is accepted to have significant label noise.

We also include the MS1MV2 dataset, which is a filtered version of the original MS1M dataset created in [60]. We use this low noise identity dataset to pre-train our models in line with the pre-training approach from [92].

3.3.2 CNN Architecture

We select the DLDLv2 architecture as it represents a strong baseline for ordinal regression tasks, outperforming many other deep approaches while showing comparable results with lightweight backbones. The original paper presents a modified, thinner and shallower VGG-16 network which they train using both KL and L1 losses.

In addition to their original model we include iResNet18 [74] to ensure that both vanilla and residual CNNs are tested. Improved residual networks have been shown to marginally improve both efficiency and performance on image recognition tasks, but to the best of our knowledge have not yet been applied to ordinal regression.

3.3.3 System Configuration

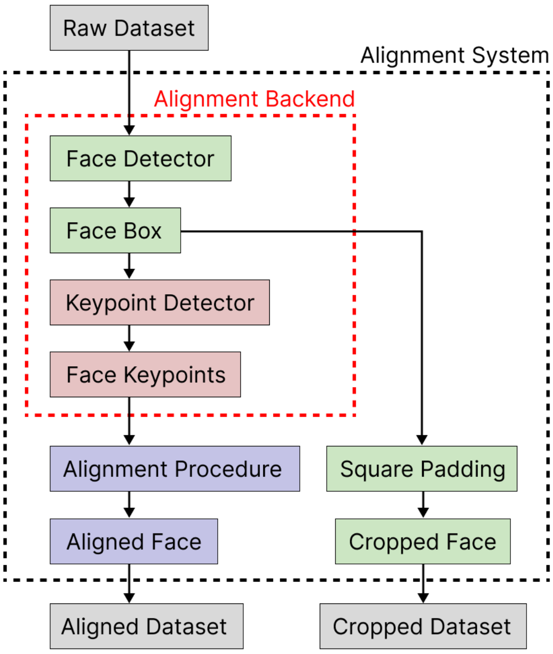
For all training experiments we used identical hyper-parameters, with the exception of CNN backbone resolution. For each backbone TinyAge and iResNet18, resolutions of 224px and 112px are used respectively. As with any hyper-parameter configuration further tuning can be done, we choose to not to optimise hyper-parameters for individual datasets, supporting the generality of our results.

All experiments are run using PyTorch 2.0.0 on an Nvidia RTX 8000 running CUDA 11.4. We implement a standard training loop running for 60 epochs across the entire randomised dataset. During training images undergo random augmentation (horizontal flip, 20 degree rotation and greyscale) following the protocol in DLDL. At inference time two forward passes are carried out with the original image and a horizontally flipped copy, from which a mean average prediction is calculated. A initial learning rate of 0.0005 is decayed by a factor of 10 at epoch 30 and a further 10 at epoch 45. All networks are optimised with ADAM using a batch size of 64.

3.3.4 Training and Evaluation

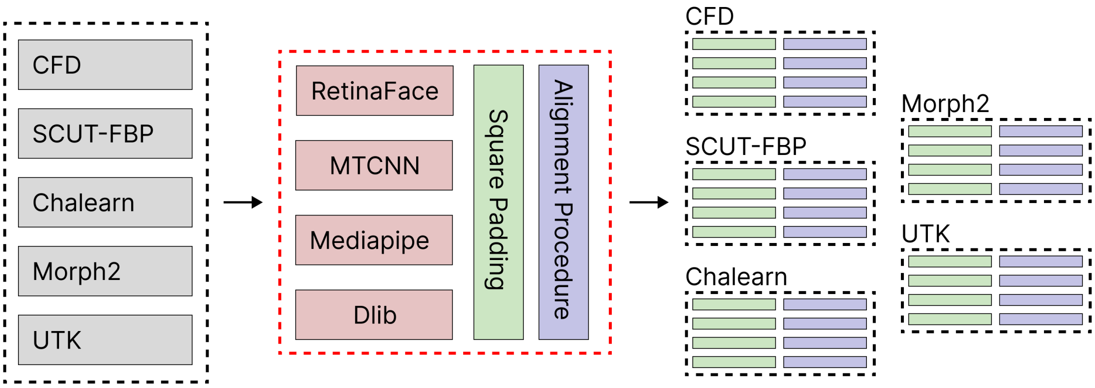
Note on the exclusion of two backends: we remove the OpenCV and SSD backends from the chapter herein. SSD only predicts bounding boxes and is paired with the OpenCV eye detector in the DeepFace implementation. This Haar Cascade based eye detector is unable to find the eyes in over 50% of images in some datasets, see Table 3.7. The next worse performer was Dlib with a less than 10% failure rate.

Figure 3.4: Face extraction and alignment pipeline.



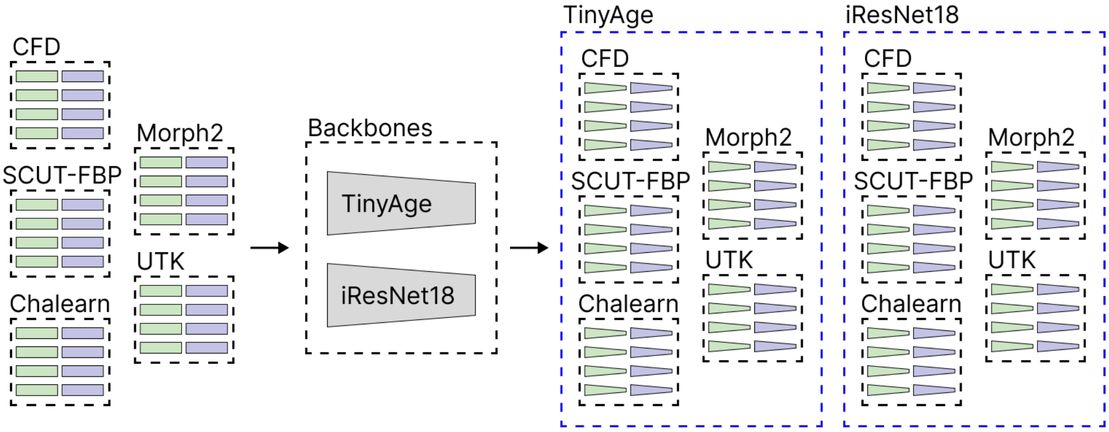
The alignment system defined in Section 3.2 takes a raw dataset as input and outputs two datasets, one aligned with our proposed procedure and one which is simply a square crop defined by the face detection component of the alignment backend. The alignment system as a whole can be seen in Figure 3.4. This diagram shows the distinction between alignment backend and alignment procedure to communicate which system components are changed within our validation experiments and which remain.

Figure 3.5: Extracting faces with combinations of alignment backend and procedure.



We repeat this extraction process across the five dataset defined in Table 3.3, for each one extracting faces using all of the six backends defined in Table 3.1. The output of the extraction process being 40 images sets, where 20 are aligned with our procedure and the remaining 20 are simple square crops. See Figure 3.5.

Figure 3.6: Model training across combinations of alignment and backbone.



Given these 40 datasets we train two different CNN architectures, both optimised with loss function presented in [92]. This results in 80 trained CNNs, half of which were trained with aligned faces and the other half trained with cropped face. Illustrated in Figure 3.6.

We then cross-evaluate each trained CNN on the other alignments of the dataset from which it was trained, within which box and eye based alignments are considered separately. More concretely we evaluate each trained model on the five differently aligned versions of the dataset from which it was trained. This is to say models trained with eye-aligned data are never evaluated on box data and vice versa. Results from these evaluations are recorded as the absolute error on a per sample basis, from which a global MAE can be calculated.

3.4 Results

3.4.1 Alignment Failures

Alignment failures are arguably the most important measurement for the quality of an alignment backend, as with no successful alignment, inference time systems must reject inputs. Alignment failures at train time may significantly reduce the size of the training set and the resultant models generality suffers a penalty. We present alignment failures as the percentage of images that failed to align due to lack of detection in either the bounding box or eye detection components (Table 3.7). As box based alignment is a precursor to eye alignment, failure rates are always the same or worse for eye based alignment.

We note that the CFD dataset is the highest resolution and least challenging dataset regarding occlusion and pose, providing a good benchmark for ‘easy’ alignment. Chalearn and UTK contain images with more challenging lighting, occlusion and pose; representing ‘challenging’ conditions. Our alignment failure results reflect these qualitative judgments precisely, showing CFD with the lowest rate of failures and Chalearn/UTK as the highest.

Box and eye detection is executed in a single stage for Mediapipe, MTCNN and Retinaface; explaining why failure rates are consistent between box and eye. Dlib, OpenCV and SSD use different algorithm components for the box and eye detection, which leads to inconsistencies in alignment failures in OpenCV and SSD. Regardless of this, Dlib achieves consistent results in both modes, indicating the two components are well tuned to operate together.

From a robustness perspective all algorithms other than OpenCV and SSD perform reasonably well across all datasets. Where eye based alignment is concerned, OpenCV and SSD perform remarkably poorly, failing to align around half of the images in some cases. MTCNN is consistently the most robust detector in all settings, showing remarkable performance even in challenging datasets. Retinaface also performs remarkably well, but is more affected by challenging settings than the other deep learning based approaches.

3.4.2 When Alignment Backends are Matched

Initially we review model accuracy (MAE) when alignment backends are matched, such that identical alignment approaches are used for both training and inference data. Matching alignment backends is the trivial solution to the problem of inference time performance degradation, however there are still some performance differences depending on which configuration is chosen.

Table 3.7: Alignment Failure Rates (lower is better)

	Dataset	cfd		chalearn		morph2		scutfbp		utk	
	Split	train	valid	train	valid	train	valid	train	valid	train	valid
Backend	Mode										
dlib	box	0.0	0.0	6.8	6.1	0.4	0.4	0.1	0.4	2.0	2.1
	eye	0.0	0.0	6.8	6.1	0.4	0.4	0.1	0.4	2.0	2.1
mediapipe	box	0.0	0.0	2.0	2.4	0.4	0.4	0.0	0.0	0.5	0.6
	eye	0.0	0.0	2.0	2.4	0.4	0.4	0.0	0.0	0.5	0.6
mtcnn	box	0.0	0.0	0.9	0.9	0.4	0.4	0.0	0.0	0.2	0.2
	eye	0.0	0.0	0.9	0.9	0.4	0.4	0.0	0.0	0.2	0.2
opencv	box	0.4	0.0	20.9	20.3	3.3	3.6	1.3	1.4	14.1	13.9
	eye	3.8	2.5	53.6	52.9	24.0	23.9	3.1	2.8	32.7	33.2
retinaface	box	0.0	0.0	1.9	2.7	1.8	2.2	0.8	0.7	2.2	2.1
	eye	0.0	0.0	1.9	2.7	1.8	2.2	0.8	0.7	2.2	2.1
ssd	box	0.0	0.0	8.7	8.5	0.6	0.8	0.9	0.5	4.5	5.2
	eye	3.8	2.5	47.9	49.1	29.2	29.0	23.9	23.8	28.5	29.2

Table 3.8 shows the results from these experiments broken down into two sub-tables to show the difference in performance in both residual and vanilla CNNs.

It is clear that iResNet18 (Table 3.8a) outperforms TinyAge (Table 3.8b) in the vast majority of tasks with roughly the same computational cost. While a comparison between CNN backbones was not part of our objectives, this is still a noteworthy observation. This performance difference may be thanks to the more advance feature extraction capabilities of the improved residual block or may stem from differences in pre-training, as both models were trained with identical hyper-parameters which may not be optimal. We do not explore these differences in this work.

Table 3.8: Average MAEs Grouped by Dataset, Mode and Alignment

(a) iresnet18

Dataset	cfd		chalearn		morph2		scutfbp		utk	
Mode	box	eye	box	eye	box	eye	box	eye	box	eye
Alignment										
dlib	6.17	**6.05**	4.07	**3.69**	2.49	**2.22**	4.73	**4.16**	4.72	**4.41**
mediapipe	6.12	**5.97**	4.14	**3.80**	2.40	**2.30**	4.40	**4.16**	4.55	**4.47**
mtcnn	6.19	**6.07**	3.85	**3.76**	2.34	**2.28**	4.29	**4.25**	4.52	**4.39**
retinaface	6.20	**5.76**	3.82	**3.66**	2.35	**2.33**	4.25	**4.15**	4.48	**4.40**

(b) TinyAge

Dataset	cfd		chalearn		morph2		scutfbp		utk	
Mode	box	eye	box	eye	box	eye	box	eye	box	eye
Alignment										
dlib	6.27	**6.15**	6.08	**5.51**	2.59	**2.38**	4.75	**4.25**	4.98	**4.49**
mediapipe	6.07	**5.78**	5.67	**5.61**	2.54	**2.44**	4.45	**4.21**	4.84	**4.57**
mtcnn	6.08	**6.00**	5.37	**5.24**	2.50	**2.42**	4.38	**4.34**	4.71	**4.54**
retinaface	6.35	**6.12**	5.54	**5.24**	2.47	**2.42**	4.34	**4.25**	4.67	**4.53**

3.4.3 When Alignment Backends are Unmatched

For many reasons, face analysis system implementors may choose to use different alignment backends at train and inference time. Perhaps there is a desire to run inference on lightweight embedded processors, while training can be run on a more substantial computer. Conversely, training may be run on a CPU base cluster while inference may use mobile GPUs. Mismatching backends can lead to degradation of inference time accuracy. We carry out a cross-comparison between 4 alignment backends within 5 different datasets, reporting both granular accuracy and average accuracy in terms of MAE (Table 3.9).

In every setting, including when alignment backends are matched, canonical eye based alignment our performs the bounding box approach.

In order to more concisely compare the two different approaches for alignment we calculate a weighted MAE for each backend, grouped by mode and matching. Due to the different sample counts in each dataset we first calculate the MAE per dataset and then a simple mean average across those values, assigning more weight to samples in smaller datasets. These results, presented in Table 3.10, show in bold the best MAE per algorithm, underlined is the best result per mode. As mentioned previously, eye based alignment outperforms box in all settings, with the retinaface backend as most optimal when backends are matched, and MTCNN most optimal when they are not.

The difference in accuracy between matched and unmatched shows the performance cost of using each alignment mode

3.5 Summary

In this chapter we integrated a long standing but underutilised face alignment procedure with a broad range of face and face keypoint detectors. We show that when faces are aligned by the eyes using our logic, mismatching face alignment backends is has a minimal negative effect on the accuracy of face analysis. Not only this, but we uncover the remarkable performance of the iResNet18 backbone for face analysis, reaching a new state-of-the-art in facial beauty prediction. We release our implementation on Github as a python module and encourage researchers and practitioners to begin following this framework as standard. During the evaluation of our framework, we discover that older face alignment backends such as OpenCV may not be suitable in modern datasets, where they fail to detect over 50% of faces under default settings. We find that MTCNN is by far the most robust face detector followed by RetinaFace and see a surprisingly strong performance from Dlib given its age. Our results allow further works to make well informed choices about their face alignment preprocessing, closing the gap between quoted and replicated results.

Table 3.9: Average MAE across both backbones; grouped by dataset, alignment backend and mode.

	Inference Align	dlib		mediapipe		mtcnn		retinaface	
	Mode	box	eye	box	eye	box	eye	box	eye
Dataset	Train Align								
cfd	dlib	6.22	**6.10**	6.36	**6.06**	6.57	**6.21**	6.67	**6.05**
	mediapipe	6.39	**6.36**	6.10	**5.88**	6.92	**6.24**	6.97	**6.23**
	mtcnn	6.65	**5.92**	6.15	**5.84**	6.14	**6.03**	6.05	**5.96**
	retinaface	6.82	**6.15**	6.23	**6.01**	6.38	**6.14**	6.28	**5.94**
chalearn	dlib	5.07	**4.60**	5.20	**4.84**	5.39	**4.77**	5.50	**4.76**
	mediapipe	5.20	**4.66**	4.90	**4.70**	5.37	**4.78**	5.50	**4.81**
	mtcnn	5.45	**4.43**	5.46	**4.55**	4.61	**4.50**	4.63	**4.54**
	retinaface	5.74	**4.36**	5.68	**4.55**	4.83	**4.45**	4.68	**4.45**
morph2	dlib	2.54	**2.30**	2.58	**2.41**	2.79	**2.37**	2.68	**2.41**
	mediapipe	2.69	**2.40**	2.47	**2.37**	3.01	**2.40**	2.87	**2.38**
	mtcnn	3.39	**2.38**	4.04	**2.39**	2.42	**2.35**	2.43	**2.39**
	retinaface	2.93	**2.40**	3.41	**2.39**	2.45	**2.41**	2.41	**2.37**
scutfbp	dlib	4.74	**4.21**	4.67	**4.31**	5.16	**4.32**	4.94	**4.29**
	mediapipe	5.11	**4.48**	4.42	**4.19**	5.43	**4.39**	5.15	**4.37**
	mtcnn	5.89	**4.31**	5.66	**4.35**	4.34	**4.29**	4.34	**4.31**
	retinaface	5.73	**4.28**	5.37	**4.22**	4.38	**4.26**	4.30	**4.20**
utk	dlib	4.85	**4.45**	4.96	**4.72**	5.24	**4.62**	5.19	**4.55**
	mediapipe	5.03	**4.54**	4.69	**4.52**	5.33	**4.62**	5.32	**4.54**
	mtcnn	5.78	**4.47**	5.76	**4.62**	4.62	**4.46**	4.63	**4.45**
	retinaface	5.66	**4.50**	5.74	**4.65**	4.70	**4.50**	4.58	**4.47**
Mean Average	dlib	4.68	**4.33**	4.75	**4.47**	5.03	**4.46**	5.00	**4.41**
	mediapipe	4.88	**4.49**	4.52	**4.33**	5.21	**4.49**	5.16	**4.47**
	mtcnn	5.43	**4.30**	5.42	**4.35**	4.42	**4.33**	4.42	**4.33**
	retinaface	5.38	**4.34**	5.28	**4.36**	4.55	**4.35**	4.45	**4.29**

Table 3.10: Weighted average MAE across all datasets when alignment backends are matched and unmatched.

Alignment Match	matched		unmatched	
Mode	box	eye	box	eye
Train Align				
dlib	4.68	**4.33**	4.93	**4.45**
mediapipe	4.52	**4.33**	5.09	**4.48**
mtcnn	4.42	**4.33**	5.09	**4.33**
retinaface	4.45	**4.29**	5.07	**4.35**

Chapter 4

Transfer Learning Driven Strategies for Age Estimation

4.1 Introduction

The impressive feature extraction capabilities of convolutional neural networks has given rise to their application across the vast majority of computer vision tasks. In most cases CNNs are trained using back propagation in combination with an optimiser such as SGD or ADAM. These optimisers work to minimise a given loss function which is calculated between the ground truth labels and the output of the neural network. Standard training has been superseded in task of PA estimation by transfer learning, a method for transferring knowledge between different tasks and datasets. Transfer learning can be formulated in a number of ways, two of which we investigate in this chapter. The most common form of transfer learning for PA prediction is the pre-train/fine-tune pattern, where training is initialised using network weights from a previous task and is continued used back-propagation. In this chapter we name this formulation simply transfer learning (**TL**), as it has become the standard for age and PA estimation. The second formulation we explore is deep feature transfer (**DFT**), which shares the same pre-training steps as TL, but instead training classical machine learning models on features extracted from the target dataset. This approach is more closely related to the feature-extractor-regressor pattern seen in classical machine learning approaches for age estimation [222].

In 2015 two seminal works were released showing the efficacy of TL in the context of age and perceived age estimation. For more details on TL see Section 2.6.1. Liu et al first introduce the concept of general-to-specific transfer learning in the context of PA prediction, showing significant improvements in the accuracy of predictions in when both FR (face recognition) and CA (chronological age) are used for pretraining [159]. Rothe et al introduced the IMDB dataset at a similar time, representing a strong baseline for CA pretraining and propelling their proposed deep PA estimation model to the top of the Chalearn LAP 2015 leaderboard [192].

Since then transfer learning has featured in the vast majority of works which improve the accuracy of CA and PA prediction on accepted benchmark datasets. To the best of our knowledge, no works have combined both CA and FR as pre-training for PA prediction since [159]. Nor has any work explored the formulation of transfer learning with the inclusion of recent developments in face age prediction. For example, [91] shows that combining FR pretraining with their lightweight VGG architecture and two-part loss function outperforms all previous works, including [159] and [192]. In its current state, the landscape of TL for perceived age estimation is neither clear nor is it consistent. A recent review paper concludes that CA pre-training is superior to FR pre-training, which directly contrasts with the findings in both [159] and the more recent [91]. Furthermore, the two leading approaches for PA estimation [91] and [203] use different pre-training strategies, making it difficult to compare their individual contributions.

In recent years work has been done to explore the possibility of repurposing CNN extracted features for classification and regression using classical ML models. To elaborate, the CNN is trained first in a standard fashion for some general or related endpoint, before being used to extract features from the training data. These features are stored on disk or in memory and are used as the training set for classical models ranging from simple linear regression to boosted trees. In some cases the secondary model is also deep learning based [137], but we do not explore such a formulation in our work. Aside from the various formulations of standard deep learning pre-training, we also investigate a more unusual approach where a classical machine learning regressor is used to fit activations from the deep neural network.

Several works have shown that face recognition is a strong pre-training approach for deep feature transfer, likely stemming from the large size of face recognition datasets and general nature of this task [159, 91]. Face recognition models are trained to produce a high dimensional embedding space where images containing the same identity are embedded closer under some distance function such as cosine or euclidean distance. This learned embedding space is a highly granular feature space containing enough information to define decision boundaries between individuals, and as such may capture other relevant information such as gender, ethnicity and age. One caveat to the feature space is the requirement to embed individuals at different ages closely together, making it challenging to capture variance associated with ageing.

Another caveat of face recognition models is their low operating resolution, which even in current research remains at 112px by 112px or lower [62], less than a quarter of the standard 224px used in most leading image recognition CNNs. At such low resolution many face ageing features are not visible, such as fine-grained wrinkles, pigmented spots and minor sun damage.

In this chapter we critically compare relevant formulations of transfer learning for age recognition, clarifying discrepancies in the literature. Furthermore, we go on to compare these popular formulations to our own more unusual approach, processing images at higher-than-intended resolution.

Our approach works on the assumption that CNNs for face recognition unintentionally extract features which are scale invariant when repurposed for other face analysis tasks. If this assumption holds true, broad swathes of research are motivated by the computational efficiency of this approach. End-to-end training of CNNs at high resolution is subject to an exponential increase in computational complexity, making works involving images above the resolution of 224px extremely uncommon. If successful, our approach allows face analysis researchers with access to high resolution training data to retest their hypotheses without the need for additional computational power. While we do not intend to investigate the implications of this approach in the broader setting of computer vision, it is also likely that derivatives of our methodology may be impactful on the training cost of other high resolution tasks.

Each approach has its own facets of interests and related research questions, but from a high level we aim to evaluate both TL and DFT in the context of PA prediction, measuring their accuracy and computational cost. We draw attention to the novel way in which DFT can efficiently extract age related features at high resolution, and the remarkable performance of DFT in the APPA-REAL validation set.

4.2 End-to-End Fine-Tuning

Two forms of transfer learning are used extensively in the PA and CA prediction literature, both operating under the principle that pre-trained CNNs learn more robust features which can be exploited for the final endpoint. In both cases the entire convolutional portion of the network is trained for either CA or FR, but in the case of FR some network modification is required before PA fine-tuning can be done. In this section we explain from a high level the 'deep learning only' transfer learning formulations we review.

4.2.1 General-to-Specific

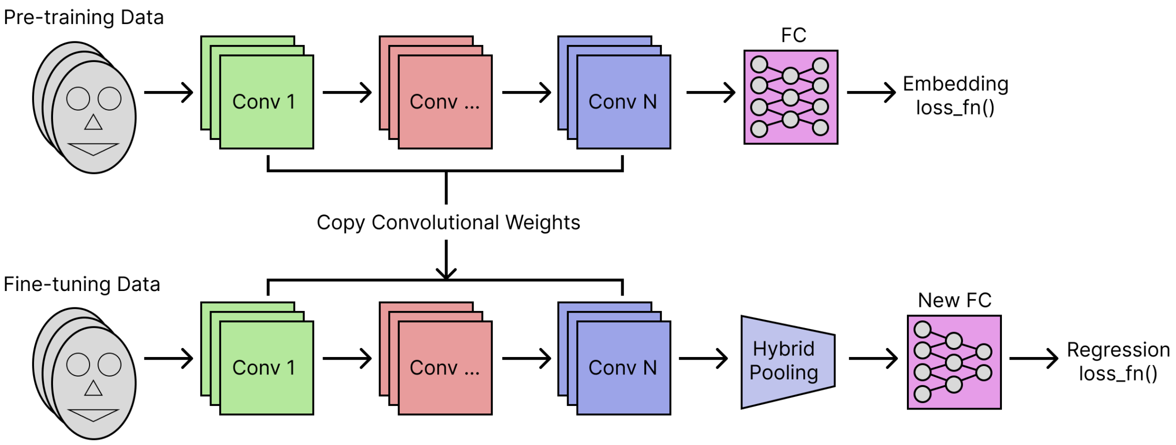
In the context of age prediction, general-to-specific transfer learning works on the assumption that CNNs trained for face recognition learn filters which are tailored for shapes and textures commonly found on the face. This is similar to how imagenet pre-training has become the de facto starting point for any computer vision task, but improves upon this approach further by learning to better represent the variance of human faces.

In our implementation, FR pre-training is done using the ARCFace loss function, which has been shown to outperform previous approaches for face recognition, and is the same pre-training as is used in [121]. We find it useful to refer to the convolutional portion of our model as the 'backbone', and the remaining portion as the 'head'. In the DLDLv2 architecture we use, the head performs the function of reducing the flattened convolutional representation to an output distribution using their hybrid-pooling layer. The DLDL head is not optimal for the loss function we use for face recognition training, such that we must modify the model between pre-training stages. To enable transferring between FR and CA/PA, we attach different heads to the model for each task as follows:

- FR: BatchNorm ->FC(512) ->BatchNorm
- CA/PA: BatchNorm ->HybridPooling ->FC(101) ->Softmax

We choose to replace the fully connected layers trained during the FR stage because the DLDLv2 method we follow for CA/PA prediction requires that a Hybrid Pooling layer be placed prior to the linear layers, invalidating the weights learned during FR which operate directly on the flattened CNN output without pooling. A diagram showing the transfer learning architecture is shown in Figure 4.1.

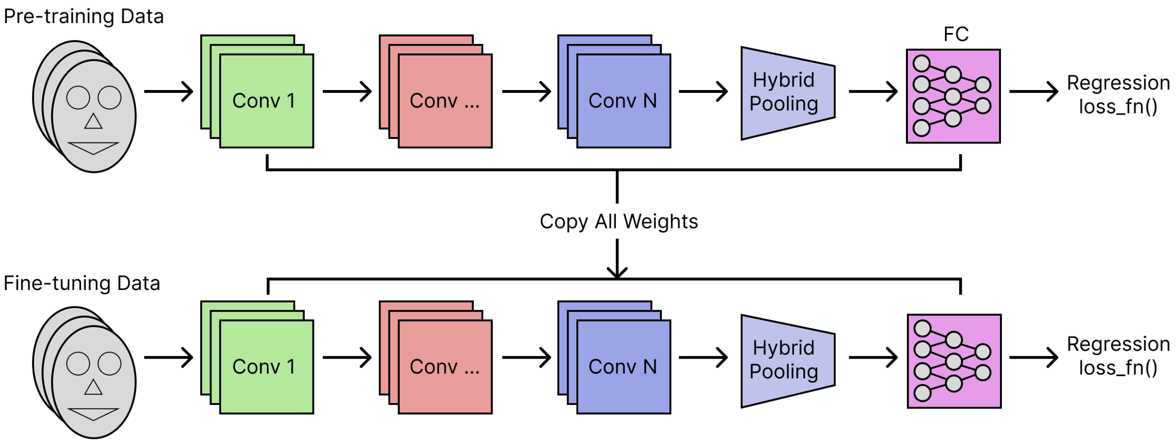
Figure 4.1: General to Specific Transfer Learning Pipeline



4.2.2 Correlated-to-Specific

Not only is chronological age highly correlated with perceived age, but it also boasts significantly more data in the public domain. Transfer learning between CA and PA is a simpler process than between FR and PA, requiring no modifications to the architecture or loss formulation. As CA and PA labels are highly correlated, it makes sense to copy the weights in the head of the model as well as the backbone. As shown in Figure 4.2, all learnable parameters are copied between the CA pre-training and PA fine-tuning stage.

Figure 4.2: Correlated to Specific Transfer Learning Pipeline



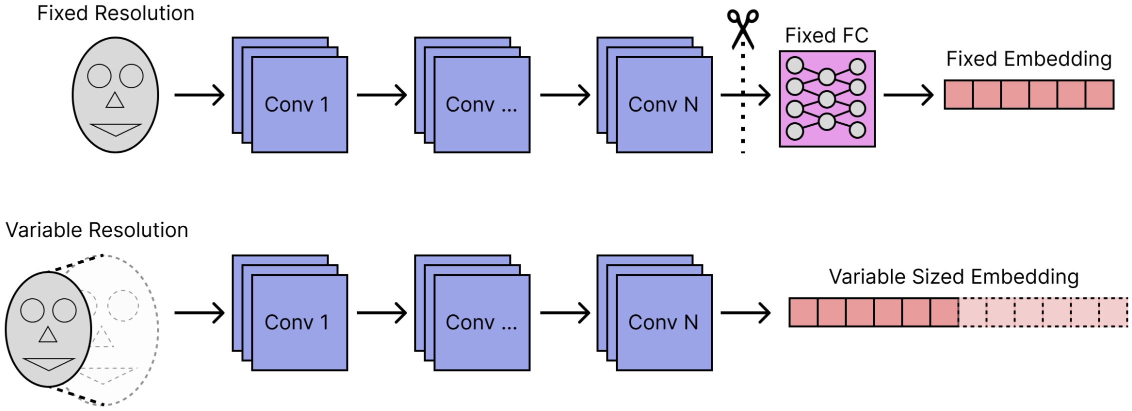
4.3 Deep Feature Transfer

Our approach is broken down into 4 key steps: pre-training, feature extraction, exploration and finally modelling. The pre-training stage makes use of the same face recognition training described in Section 4.2.1. Feature extraction requires a modified backbone, outputting facial embeddings which are stored on disk for further processing. Exploration and modelling are then done by reading embeddings from disk and processing them with a variety of classical machine learning algorithms. We include some exploratory analysis to demonstrate the feature representation capabilities of such models, before attempting to more specifically model age and perceived age.

4.3.1 Feature Extraction

Feature extraction in computer vision is the process of compressing high dimensional images to a lower dimensional feature space. Practically this involves passing batches of images through our CNN and extracting embedding space vectors from the final layer. We pre-train our CNN at an input resolution of 112px which in turn outputs a 512-d embedding or feature vector. When a higher resolution image is passed through the model all convolutional layers operate correctly, with the only effect being an increased spatial resolution following the pooling operation. This creates a problem however for the fully connected portion of the model, as input sizes to linear layers are fixed at train time. To solve this problem we remove the fully connected layers and simply add a flatten operation after the final pooling layer.

Figure 4.3: Network Modification Allowing for Increased Resolution



While the removal of the fully connected portion of the model would clearly degrade the accuracy with regard to face recognition, it may still be suitable in the context of general feature extraction.

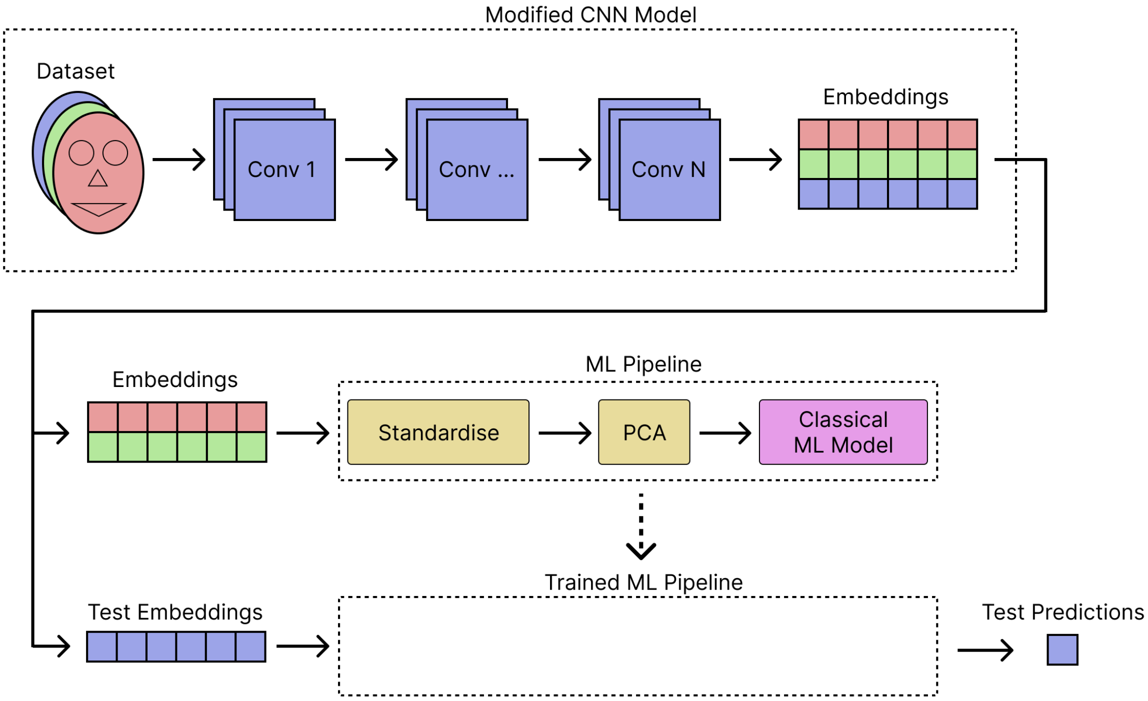
Given a high resolution capable CNN, features are extracted from batches of images loaded using the multi-threaded pytorch dataloader, allowing for full CPU utilisation and memory load efficiency within the GPU. Extracted features are copied from the GPU into numpy arrays in system memory, before being saved to disk in the binary npy format. We test both CSV and NPY format for saving feature volumes, ultimately finding npy to be far more efficient regarding disk space and read speed.

4.3.2 Modelling Pipeline

To explore the various configurations and parameters for training ML predictors on face embeddings we develop a modelling pipeline implemented with scikit-learn. First embeddings are extracted at the resolutions shown in Table 4.9 for every dataset. All embeddings are saved to disk using a filename which is indexed by dataset and resolution. This file structure allows the following stages to operate without the need to repeat feature extraction which is the most computationally intensive part of the process.

The machine learning pipeline we use is relatively standard, first standardising the features, then reducing dimensionality with PCA and finally regressing or classifying an output. We parametrise the pipeline with the following values: number of PCA components, input resolution and ML model. We also include the ability to skip PCA entirely, fitting the ML model directly to the standardised feature space. We define a number of experiments in which we make predictions using this pipeline under 5-fold cross-validation. The accuracy of the outputs is calculated using mean absolute error. An overview of the pipeline is shown in Figure 4.4.

Figure 4.4: Deep Feature Transfer Pipeline



4.4 Datasets

In our experiments we use a range of datasets supporting each stage of transfer learning, including two additional datasets designed specifically for learning to predict perceived age in the elderly, shown in Table 4.5. These datasets are labelled ECA and EPA representing elderly chronological age and elderly perceived age respectively, a full list of dataset abbreviations is shown in Table 1. The EPA dataset is created by filtering the APPA-REAL dataset for faces between 60 and 90 years old, resulting in an extremely small sample of 151 images, most of whom are between 65 and 75 years old.

The ECA dataset is produced by sampling all faces in the same age bracket (60-90) from four popular CA datasets, taking 11949 from the IMDB dataset, 3666 from UTK, 1134 from AAF and 248 from Morph2. It is our assumption that the number and distribution of ages in this dataset may allow the CNN to learn features that are more closely associated with the progression of ageing in the elderly. The remaining 3 datasets are common benchmark datasets seen in the literature. We select IMDB-clean for CA pre-training thanks to its strong performance in [192], the original IMDB dataset contained significant label noise, often completely mixing up identities. In [154], the authors filter the dataset and present results indicating the new version is not only smaller but leads to better accuracy. For these reasons we choose to use the cleaned version.

Following the theme of using pre-training data which is a filtered version of the original dataset, we access MS1M-V2 which is presented in its filtered form in the original ARCFace publication. This dataset contains 87K identities with 5.8M face images, allowing models to learn from an extremely diverse set of features, though we cannot comment on the diversity of demographics.

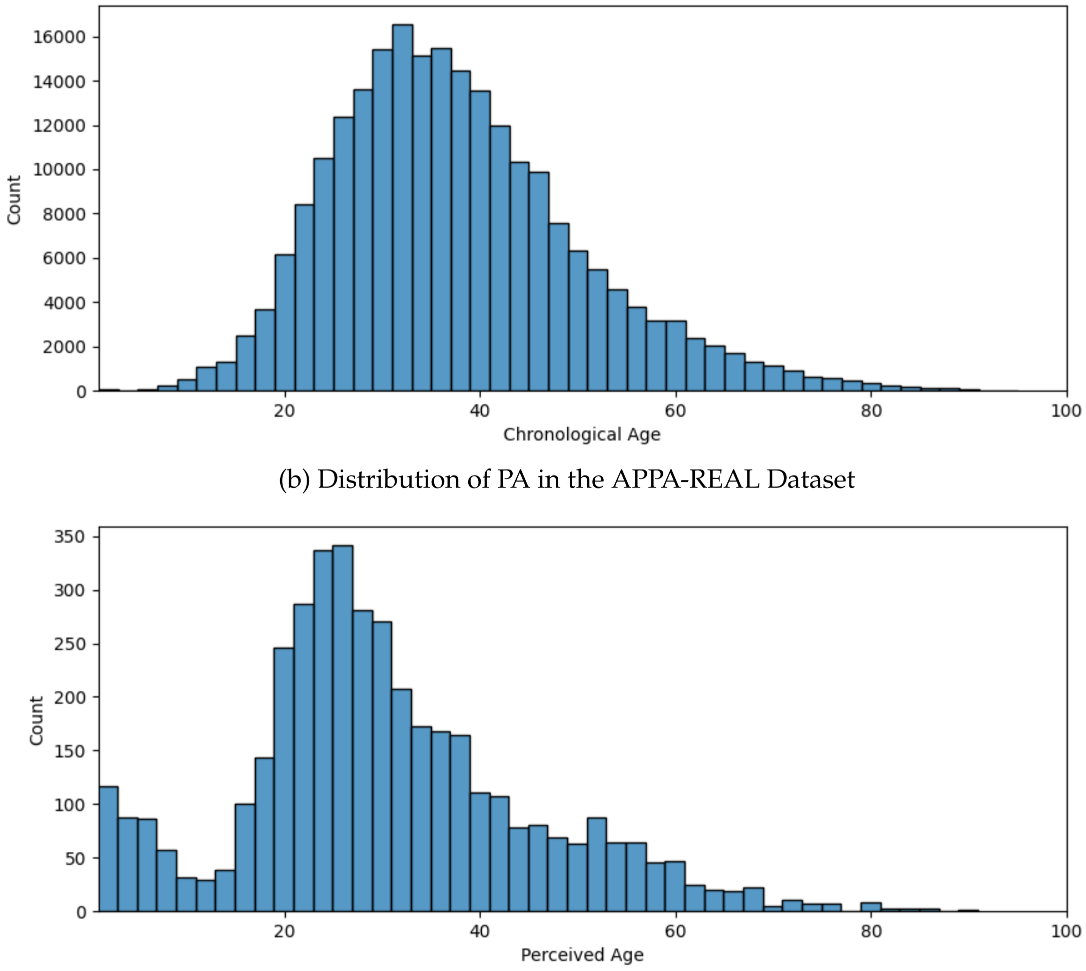
The final benchmark dataset is the unmodified APPA-REAL dataset, which is an extension of the well known Chalearn LAP 2015 challenge. It contains 4113 training images and 1500 validation images, on which we report all of our accuracy results in terms of MAE. All images are preprocessed using the framework developed in Chapter 3, using identical positioning parameters as used in that work. During training the images are standardised using the per-channel mean and standard deviation.

Table 4.5: Datasets used to support this chapter.

Abbreviation	Name	Task	Train Samples	Mean (Std)
FR	MS1M-V2	FR	5,822,653	-
CA	IMDB-Clean	CA	229263	36.9 (12.7)
PA	APPA-REAL	PA	4113	30.2 (14.7)
ECA	Elderly Faces	CA	16997	69.0 (7.0)
EPA	APPA-REAL Elderly	PA	151	67.6 (6.5)

Later in this chapter we present results showing the biases induced in models which may be associated with the imbalance in class distribution in PA and CA datasets. Figures 4.6a and 4.6b show the distributions of the two commonly used public datasets we access in this work, showing they are clearly not uniform. We also see clear long distribution tails in the oldest and youngest samples, making them the most challenging to learn.

Figure 4.6: Class Distributions in Two Key Datasets



(a) Distribution of CA in the IMDB-Clean Dataset

(b) Distribution of PA in the APPA-REAL Dataset

4.4.1 Additional Data

In addition to the transfer learning datasets specified at the start of this section, we include four more datasets with additional labels such as gender and ethnicity. These labels are valuable when interrogating the embedding space learned by the face recognition model, as that both gender and ethnicity play a key role in human identity. We include 2 other datasets to explore the ability of DFT to operate at higher-than-standard resolution. The first is a perceived age dataset made by combining images from the Leiden dataset with those from an internal Unilever study. While it is likely this dataset does not contain enough samples to train a reliable PA estimator, it is the only high resolution dataset with PA we are able to access during these experiments. The second dataset we access is the CFD dataset, a small sample of very high resolution face images with perceived beauty labels. This additional data, shown in Table 4.7, is exclusively in our DFT experiments.

We choose not to make direct comparisons between DFT and TL from an accuracy standpoint as previous works shows that TL outperforms fixed feature extractor approaches.

Table 4.7: Additional high resolution data with ethnicity, gender and beauty annotations.

Name	Samples	Labels	Face Resolution
CFD	597	Ethnicity, Gender, Beauty	~1000 px
Morph2	55,134	Ethnicity, Gender, CA	~200 px
APPA-REAL	7591	Gender, CA, PA	~250 px
Unilever Internal + Leiden	1168	Gender, CA, PA	~500 px

4.5 Experimental Design

To understand the relationship between various transfer learning stages we design a series of training experiments, from which results are analysed together.

4.5.1 Training Configuration

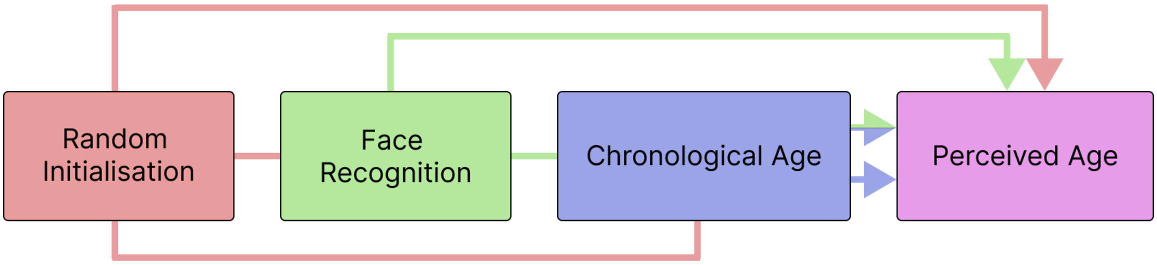
Other than FR training, all other training setups are identical. Following exactly the same protocol as in Chapter 3, use the iresnet18 backbone and DLDLv2 loss formulation, all models are trained for 60 epochs, the learning rate is set to 0.0005 and reduced by a factor of 10 at epochs 30 and 45. The deep learning library used is PyTorch, and all experiments are done on an NVIDIA RTX 8000 GPU running CUDA 11.4.

4.5.2 Multi-Stage Transfer Learning

Experiments are indexed by their component pre-training stages. For example, initial face recognition training is labelled FR, face recognition which is fine-tuned for PA is FR_PA and training involving all 3 stages is FR_CA_PA. Clear dependencies exist between our experiments, such that their ordering must be carefully considered. Figure 5 shows a flow diagram demonstrating these dependencies. Face recognition and random initialisation are done once, saving the weights as controlled starting points for the other experiments.

During training, we evaluate performance in a non-standard way. Instead of splitting the dataset into training and validation, we validate all models on their accuracy in predicting PA in the APPA-REAL validation set. This approach gives us an interesting view into the accuracy of each model with regard to perceived age prediction, even when it has seen no PA data. It also allows us to fairly compare between datasets without the need for additional CA data, allowing all CA data we access to be used for training.

Figure 4.8: Various Transfer Learning Configurations



4.5.3 DFT Parameter Search

To evaluate the DFT approach we broaden the range of tasks to test the generality of face recognition features. Our first round of experiments uses XGBoost with the GPU_HIST tree method. This approach allows us to work with significantly larger datasets such as Morph2, as well as completing training much faster than when fitting on the CPU. Both classification and regression with XGBoost are highly robust with regard to hyper-parameters, such that it is suitable for initial experiments to explore the feature space. Following these preliminary experiments we narrow our focus to regression tasks, testing a range of Scikit-learn regressors. As with XGBoost we do not modify the various hyper-parameters of each regressor, instead exploring the relevance of resolution and PCA. Table 4.9 shows an exhaustive list of the regressors, principal components and resolutions we test.

Table 4.9: Parameters search during DFT experiments.

(a)

Resolution (px)
112
224
320
448
512
640
800

(b)

Principal Components
No PCA
64
128
256
512
1024

(c)

Regressor
Linear SVR
Bayesian Ridge
Huber Regression
ARD Regression
Linear Regression
Elastic Net

4.6 Results

4.6.1 Demystifying Perceived Age Fine-Tuning

Even though the number of experiments we carry out is relatively small, data visualisation is challenging due to the irregular shape of our results and non-numeric nature of the parameters we explore. In this section we first show our results with regard to PA prediction across the entire age range, and secondly present the results regarding bias in the elderly and its mitigation. In addition to the standard MAE we also include a Weighted MAE, which is calculated first for each class and finally averaged across all classes, weighting all classes equally.

Optimal Configuration for General Performance

For the sake of comparability, work on perceived age prediction almost exclusively presents results on the validation set of one of the well known APPA-REAL apparent age datasets. In this section we present the results of our training experiments on the validation set of the APPA-REAL dataset, including all samples from all age ranges.

Table 4.10 shows the MAE and weighted MAE of each training run as taken at the final epoch. We select a subset of results in this table which make for the most interesting comparison.

Table 4.10: MAE and Weighted MAE Across All Ages.

Experiment	MAE	Weighted MAE
PA	6.37	9.77
CA	5.47	6.46
FR_CA	3.91	4.17
CA_PA	3.88	5.12
FR_PA	3.16	3.76
FR_CA_PA	2.84	3.45

With regard to the general accuracy of PA prediction, several interesting discoveries were made. The most obvious of which is that training from random initialisation with only CA data produces a more accurate PA classifier than training with PA alone, a gap which is even more significant when FR_CA is used. In contrast, FR_PA performs significantly better than FR_CA, indicating that FR pre-training has a bigger impact on PA than CA training. Both the best MAE and weighted MAE are produced when all 3 pre-training stages are used in FR_CA_PA, with FR_PA performing marginally worse. CA_PA outperforms FR_CA with regard to MAE, but performs worse regarding the weighted MAE, a pattern which is not seen between any other pair of experiments.

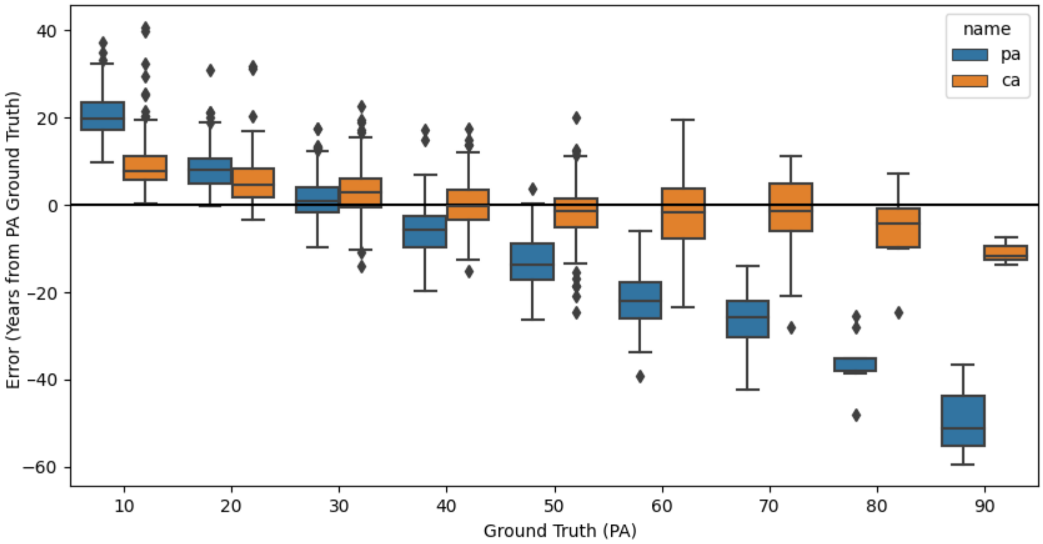
The Weighted MAE value in Table 4.10 gives some indication of the biases present through the various stages of transfer learning for PA prediction, however a more granular view can be taken by plotting the distribution of errors per-class label.

Class Distribution Induced Bias

Following each training experiment we extract a full set of predictions from the APPA-REAL validation set. In this section we show relative error instead of absolute error to indicate both the magnitude and direction of errors, broken down by age group.

Figure 4.11 shows the accuracy of perceived age predictions when training naively with CA and PA data from a random initialisation. Both models show the same characteristic bias associated with the class imbalance in each dataset, where the perceived age is generally under-predicted in the elderly and over-predicted in the youth. CA training with the IMDB-Clean dataset produces significantly less biassed results than training using the APPA-REAL dataset alone. Predictions from the CA experiment appear to be mostly biassed for individuals younger than 30 and older than 70. In contrast, predictions from the PA experiment are heavily biassed for anyone over the age of 50.

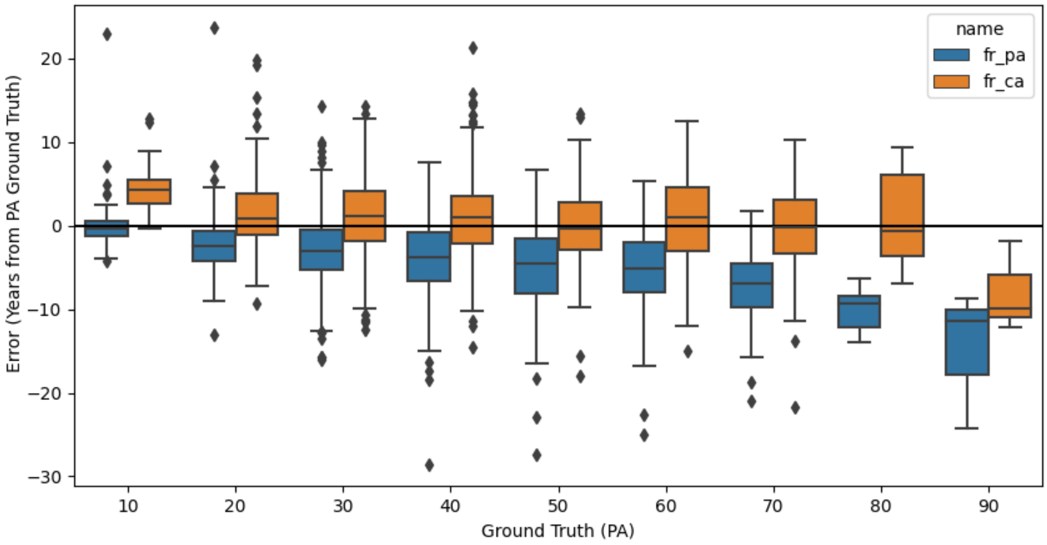
Figure 4.11: Per Class Error for PA and CA Trained from Random Initialisation.



With the addition of FR pre-training both CA and PA training produce more accurate results as shown in Table 3. Not only is the accuracy improved overall but so are the biases (see Figure 7), with FR_CA showing good performance between the ages of 20 and 80, an increase in the suitable age range of around 20 years. FR_PA also sees some benefit from the FR pre-training, but remains notably biassed across much the same age range. Crucially, we note that in the 80-90 age group the

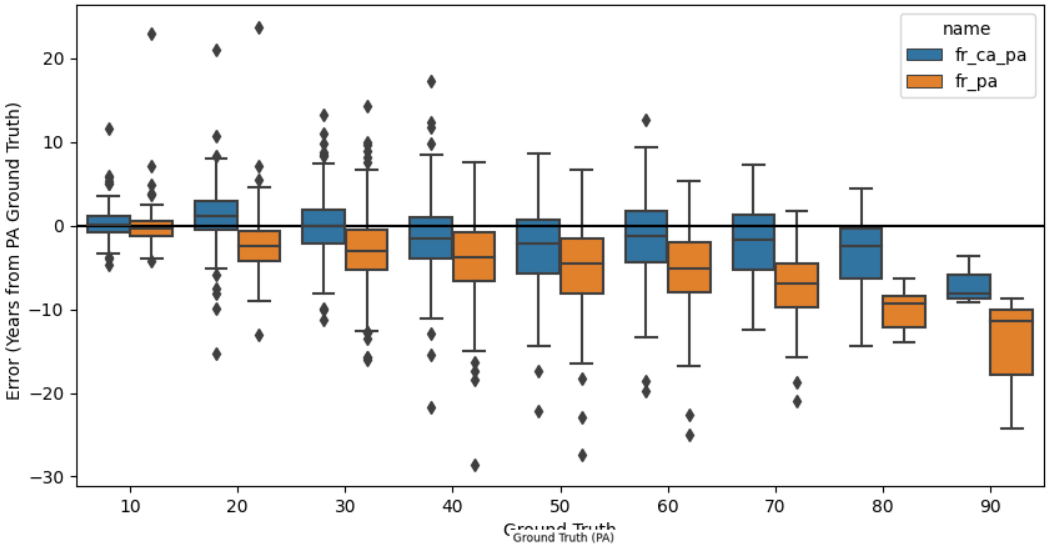
FR_PA model produces predictions which are on average younger than the 70-80 group, indicating that there remains a significant negative bias in the elderly.

Figure 4.12: Per Class Error for FR_PA and FR_CA.



In Figure 8 we show a comparison between FR_CA_PA and FR_PA, demonstrating the reduction in bias when the CA pre-training stage is included. FR_CA_PA produces predictions that are less biassed across the board than FR_PA, especially in the elderly. Both experiments still show a heavy bias in the 80-90 year old group, but FR_CA_PA is both more consistent and more accurate.

Figure 4.13: Per Class Error for FR_CA_PA and FR_PA.



Optimal Configuration for Prediction in the Elderly

The addition of the ECA and EPA datasets show an improvement in accuracy for the elderly with a remarkable reduction in accuracy in younger individuals. In Table 6 we show the top 8 results ordered by MAE in the elderly results. In most cases fine tuning on the EPA dataset outperforms fine-tuning on other datasets. FR_ECA outperforms FR_EPA in the elderly, in the same way FR_CA outperformed FR_PA across all ages (Section [ref previous results]). We note the remarkable accuracy of FR_CA_EPA for prediction in both the elderly, for which it is the most effective approach, but also in all ages. FR_CA_EPA does not suffer from the same elderly bias introduced in the next 4 best performing models in the elderly age group. We also draw attention to FR_CA_PA and FR_PA, which produce good results while still underperforming relative to elderly specialised configurations.

Table 4.14: MAE in the Elderly and Across All Ages.

Experiment	Elderly MAE	All MAE
FR_CA_PA	4.39	2.84
FR_PA	3.71	3.16
FR_EPA	3.45	24.07
FR_ECA	3.42	33.11
FR_CA_ECA_EPA	3.29	31.90
FR_ECA_EPA	3.06	32.33
FR_CA_EPA	2.82	4.76

4.6.2 Facets of Deep Feature Transfer

In this section we review the results of our DFT experiments, first showing the general applicability of the FR embedding space and then comparing various regression approaches and their links with PCA and image resolution.

Generality of Features

Our first set of experiments fit XGBoost models to the embeddings extracted from each dataset at the standard 112px operating resolution with no principal component analysis. For PA, CA and Beauty we use XGBoost regression and for ethnicity and gender we use the XGBoost classifier. The results of which are shown in Table 4.15.

In all datasets we see remarkable accuracy in gender prediction, reaching near state-of-the-art levels with no parameter tuning, indicating that FR training captures highly relevant features for this task. Facial Beauty in the CFD dataset significantly underperforms the results found previously in the literature, as is the case for CA and PA prediction in all 3 datasets. In the APPA-REAL dataset we see significantly more accurate results for PA prediction than CA, indicating that the DFT approach may be more suitable for extracting health related features than those related to natural ageing.

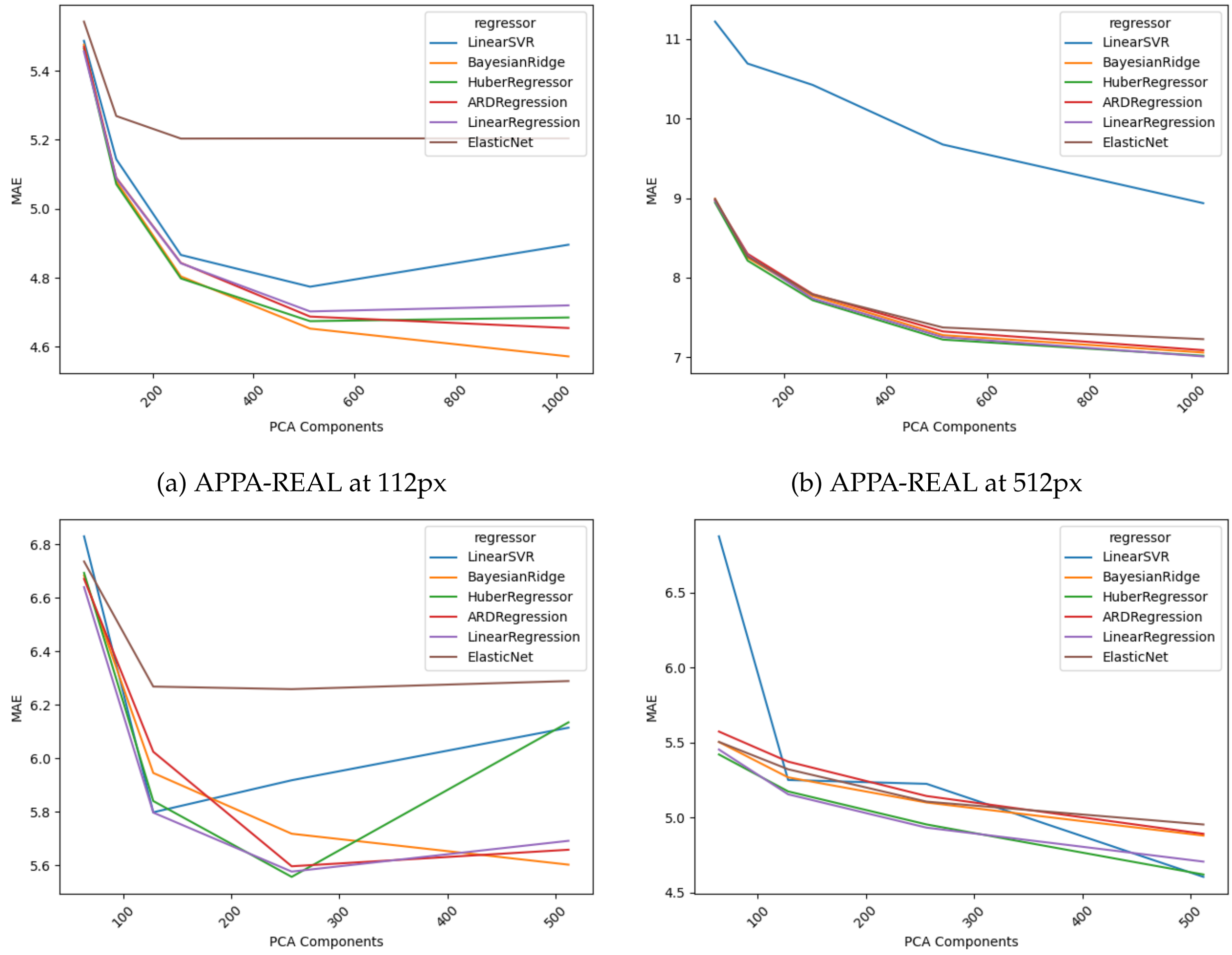
Table 4.15: MAE and Accuracy of Raw Face Recognition Features for Various Tasks

Metric	MAE			Accuracy	
Feature	PA	CA	Beauty	Ethnicity	Gender
CFD	-	-	0.542507	0.876047	0.974874
APPA-REAL	4.943088	6.503424	-	-	-
Morph2	-	4.915885	-	0.977037	0.984091
Unilever	7.156450	5.784407	-	-	0.971933

Principal Component Analysis

PCA plays a key role in DFT for two reasons. Firstly, it reduces the dimensionality and thus the size of features in memory, making the problem tractable even for large datasets. Secondly is the way in which it reduces the dimensionality. Because PCA removes covariance between features, the information lost likely does not represent a useful signal for regression. Furthermore, training a regression model with many highly correlated features can result in poor performance. In our second set of experiments we explore the influence of principal component analysis and image resolution on the MAE of PA and Beauty regression. We analyse our experimental results to measure the impact of the number of principal on the accuracy of each model, clarifying whether or not the aforementioned reduction in covariance is truly an improvement. Figure 4.16 shows the MAE resulting from various regressors against the number of principal components. For comparison we include plots for both APPA-REAL and Unilever datasets at both low and high resolutions. The maximum number if PCA components is limited by the number of samples in the Unilever dataset, whereas the arbitrarily select 1024 as the maximum for APPA-REAL. As the number of PCs increases the error generally decreases, a trend which is consistent between all regressors at high resolution (512px). This indicates that when the input image is large, a covarying features contain information which is useful for predicting age. At low resolution (112px) there appears to be an inflection point around 256 components, where additional PCs have a negative effect on accuracy. This trend is particularly strong in the Unilever dataset, where additional PCs hurt accuracy in every regressor other than Bayesian Ridge. The relationship between number of PCs, input resolution and accuracy is remarkable in this setting. It is clear that to exploit features from higher resolution images more principal components are required, thought it is unclear how this relationship continues outside the range of values we test.

Figure 4.16: PA MAE Against PCA Components for APPA-REAL and Unilever, coloured by Regressor.



(a) APPA-REAL at 112px

(b) APPA-REAL at 512px

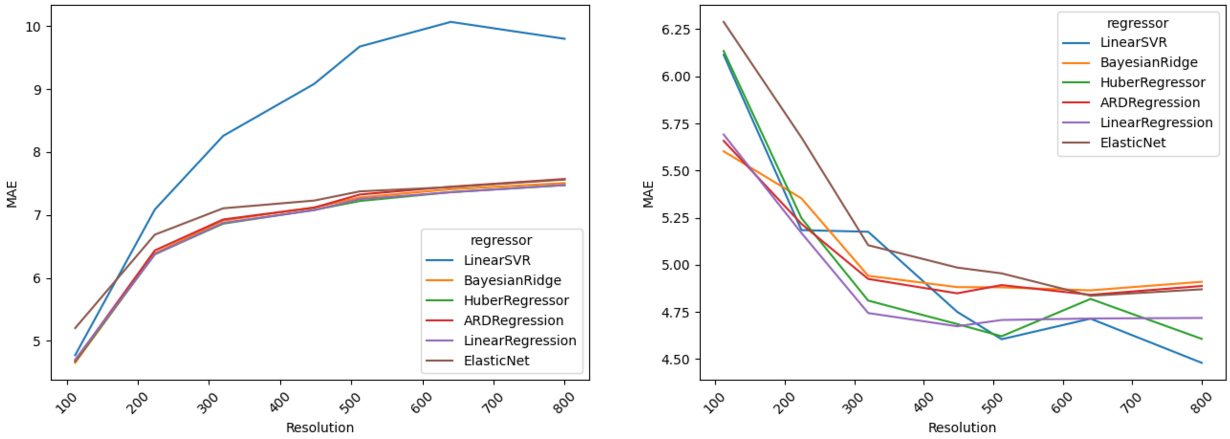
(c) Unilever at 112px

(d) Unilever at 512px

Increased Resolution

To better understand the link between input resolution and regression accuracy we fix the number of principal components at 512 and plot the MAE against resolution. Figure 4.17 shows opposite trends for the APPA-REAL and Unilever datasets. In the high resolution Unilever data MAE reduces as resolution increases, indicating that more relevant principal components are extracted at higher resolutions. In the APPA-REAL dataset the optimal operating resolution is 112px, with additional resolution worsening the MAE. This indicates that only when the source resolution is high can the DFT approach make use of higher resolution inputs.

Figure 4.17: PA MAE Against Resolution for APPA-REAL and Unilever coloured by Regressor at 512 PCs.

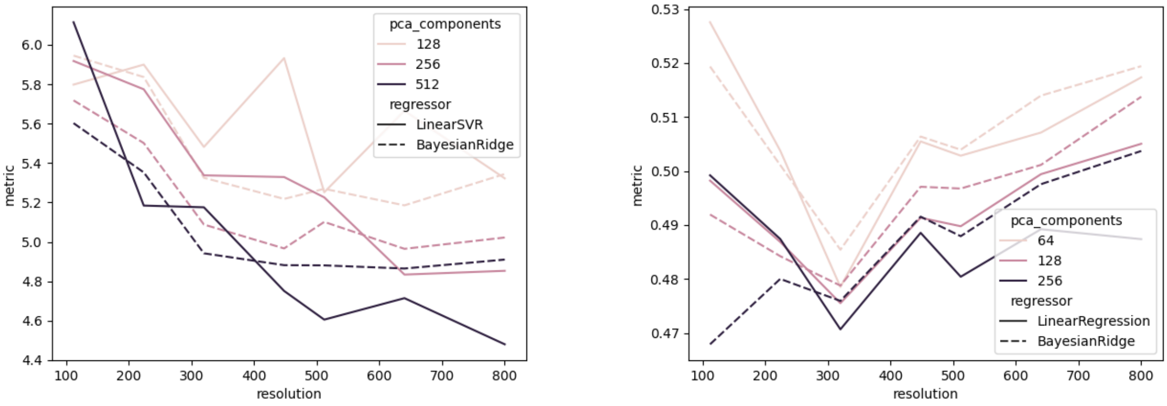


(a) APPA-REAL

(b) Unilever

To confirm the link between high resolution sources images and DFTs ability to leverage high resolution inputs, we explore the results of the CFD dataset. Figure 4.18 shows the link between MAE and resolution in our two high resolution datasets, including only the two best regressors for each dataset. Additionally we include number of PCA components to confirm visually the trend applies at all dimensionalities. In the unilever dataset we see the same trend of accuracy improving with resolution, but are surprised to see a more complex trend in the CFD dataset. MAE does improve initially in the CFD dataset, but begins to worsen again above 320px. The worsening seems less pronounced when higher numbers of PCs are used, but it should be noted that is finding goes against our hypothesis. In both datasets the best performance is found when Bayesian Ridge regression is used at the highest number of principal components, though at lower resolutions this relationship is less consistent.

Figure 4.18: PA MAE Against Resolution for CFD and Unilever, coloured by Regressor.



(a) Unilever

(b) CFD

4.6.3 Comparison to the State-of-the-Art

In this section we compare our best models from each approach with the leading state-of-the-art approaches for perceived age estimation. We include both models trained on the Chalearn 2015 [79] dataset as well as its newer APPA-REAL [10] iteration. For pretraining all top performing methods use the IMDB-WIKI [192], in place of which we use the latest IMDB-Clean [154] dataset. Our TL method outperforms all other methods by a large margin, possibly due to inclusion of both FR and CA pre-training with low noise datasets. Table 4.19 shows a summary of various SOTA approaches ordered by their MAE on one of the Chalearn validation sets.

Table 4.19: State-of-the-art models performance in the APPA-REAL dataset.

Paper	Validation Dataset	Pretraining	MAE
Ours (DFT)	APPA-REAL	MS1M-V2	4.94
Agbo-Ajala, O. and Viriri, S. [7]	APPA-REAL	IMDB-WIKI	3.809
Li et al. [151]	APPA-REAL	IMDB-WIKI	3.49
Gao et al. [92]	APPA-REAL	IMDB-WIKI	3.452
Rothe et al. [191]	Chalearn 2015	IMDB-WIKI	3.221
Gao et al. [92]	Chalearn 2015	IMDB-WIKI	3.135
Ours (TL)	APPA-REAL	IMDB-Clean	**2.84**

4.6.4 Computational Requirements

The primary benefit of the DFT approach is its power to leverage an already trained face recognition model, which can be accessed online with no training required. Table 4.20 shows the training time and inference speed of each approach under different configurations. We assume that the training cost of FR has already been paid in each case, as it is a common requirement for all approaches. Fine-tuning directly on the APPA-REAL dataset takes roughly 20 minutes, compared to training on both IMDB-Clean and APPA-REAL which takes around 6 hours. For DFT, we measure the training and inference speeds of Linear regression on the same dataset, finding that in all cases DFT is significantly faster.

At inference time DFT is only slightly slower than TL based methods when operating at the same resolution of 112px, this is likely due to the additional steps needed to copy the extracted features to the main system memory before processing the linear model. When operating at 800px the DFT approach is remarkably slow, this is because the feature extraction process takes significantly longer at this resolution.

Table 4.20: Comparison between the computational complexity of each approach on the APPA-REAL dataset.

Method	Configuration	Training Time (seconds)	Inference Speed (fps)
TL	FR_PA	1381	35
	FR_CA_PA	21,922	35
DFT 112px	128 PCs	63	34
	256 PCs	68	34
	512 PCs	81	34
DFT 800px	128 PCs	334	12
	256 PCs	361	12
	512 PCs	389	12

4.7 Summary

In this chapter we explored two different approaches for transferring knowledge from one task to another, TL and DFT. We demystify the the various stages of TL for PA estimation, showing that using the filtered versions of two pre-training datasets allows our approach to reach a new state-of-the-art in the APPA-REAL dataset. It is made clear that the use of a large CA dataset in pre-training significantly reduces the class-imbalance-related bias in PA estimation, even if the CA dataset contains an equal level of imbalance, indicating that a threshold for the minimum number of samples per class may exist.

We propose a new method for extracting face analysis features from high resolution images using a pre-trained face recognition model, finding that this approach only extracts better features in datasets with a high source resolution. We do not see the same trend in the two high resolution datasets used, facial beauty estimation accuracy does not improve at the most extreme resolutions, whereas in PA estimation it does.

Finally we compare the computational efficiency of the two contrasting approaches, finding that DFT is several orders of magnitude faster to train, but is slightly slower at inference time. When higher resolutions are used DFT is significantly slower, crossing below a 'real-time' rate. While our methods perform remarkably well in the datasets used, we note that very large task specific pretraining datasets are needed, limiting the efficacy of this method in other tasks. We do note that face recognition pretraining is an impressively general approach for other face related tasks, and should not be overlooked. The foremost contribution of this chapter is summarised well in Table 4.19, where we call attention to the marginal outperformance of our approach above all other competing methods. In the next chapter we take both TL and DFT as baselines for a novel semi-supervised learning approach to PA estimation.

Chapter 5

Association of Age Modelling with Health and Genetics

5.1 Introduction

This chapter is inspired by the remarkable link between human estimated perceived age and healthy ageing. PA is not only a measure of vanity or pure scientific interest, but it is a window into individual and population health. In 2016 Lui et al [155] found that compound variants of the MC1R gene were significantly associated with perceived age. The compound variant was a combination of four missense SNPs from the MC1R gene: rs1805005, rs1805007, rs1805008 and rs1805009. Risk alleles were categorically encoded to allow for statistical analysis with perceived age. While impactful, the nature of this genetic link is out of scope for our work, and is herein viewed simply as a categorical label. More recently, Mekic et al [171] used the same population to look for links between PA and several key morbidities. They found PA to be significantly linked to osteoporosis, COPD, cataracts, hearing loss and cognition. These findings make a strong case for further research into perceived age in larger and more diverse populations, however such future work is limited by the cost of human PA annotation.

We aim to replicate the findings of [155] and [171] but instead using PA as estimated by a deep learning, as opposed to human annotators. Computer predicted perceived age entirely removes the time and cost associated with classical perceived age annotation, which requires a minimum of 10 annotators per image to produce a reliable result. In comparison, recent deep learning [92] based approaches allow PA to be calculated at a rate of over 10 faces per second on standard computer hardware, such as laptops or mobile devices. In recent years the accuracy of computationally estimated chronological age has reached and exceeded human accuracy, giving way for the same technology to be used for PA. PA and CA prediction from digital face images operate in much the same way, regressing a final value using a series of neural network layers. PA is dependent only on visual facial queues, whereas CA is simply a measure of the time taken to reach a given biological state, indicating that prediction of PA should be easier for a vision-based algorithm than CA.

We combine several leading methods for training deep convolutional neural networks to improve the accuracy of PA prediction. In addition to this, we exploit the 3D information captured with the 3DMD [162] camera system to render frontal face images with highly controlled pose and lighting.

CNN based perceived age estimation is built on the back of chronological age estimation, in which standard datasets range in size from 10k to 500k samples, making it difficult to apply such methods to small lab grade datasets such as the Rotterdam ERGO study [101]. For this reason, we propose a new approach designed to extract age information from faces in a 'small data' setting. Our approach builds on the work on TL in Chapter 4 by adding a second unsupervised loss component to the DLDL architecture. The loss is taken from a recent work in semi-supervised learning for image classification, unsupervised data augmentation (**UDA**)[224]. We show the modifications needed to align UDA with the DLDLv2 loss formulation and DAE label encoding, finding that including a recent method for solving class imbalance allows UDA to improve significantly on standard transfer learning. This method, Balanced mini-batch sampling (**BMS**) [202], is used to ensure that every batch contains a roughly uniform distribution of CA or PA, allowing the model to learn from the UDA loss across a broader distribution of features.

In summary, we hypothesise that perceived age as predicted by deep learning will share the same links with health and genetics as found with human PA. To test this hypothesis several contributions are made:

1. A novel semi-supervised learning approach for PA estimation in 3D rendered images.
2. An automated rendering and preprocessing pipeline for 3D face captures.
3. Replication of Human PA associations with morbidity and genetics using deep learning PA.
4. Replication of PAs association with the MC1R gene in an unseen cohort.

5.2 Dataset

5.2.1 Data Collection

The high level focus of this chapter is to find links between CNN estimated perceived age and health labels known to associate with human perceived age. To support this work we access a dataset being collected at Erasmus University, namely the ERGO study. ERGO is a longitudinal epidemiological study focussed on healthy ageing, where participants have a range of physical and biometric tests. Most importantly for our research, high resolution face images of the subjects were taken allowing us to train image based estimators. The other key component of this dataset is the perceived age labels generated following the same procedure as in [108].

Participants in the ERGO study were asked to attend Erasmus University where their photos would be taken with a 3D camera system. They were also asked to avoid wearing makeup or skin products during the photo shoot, as well as maintaining neutral facial expression.

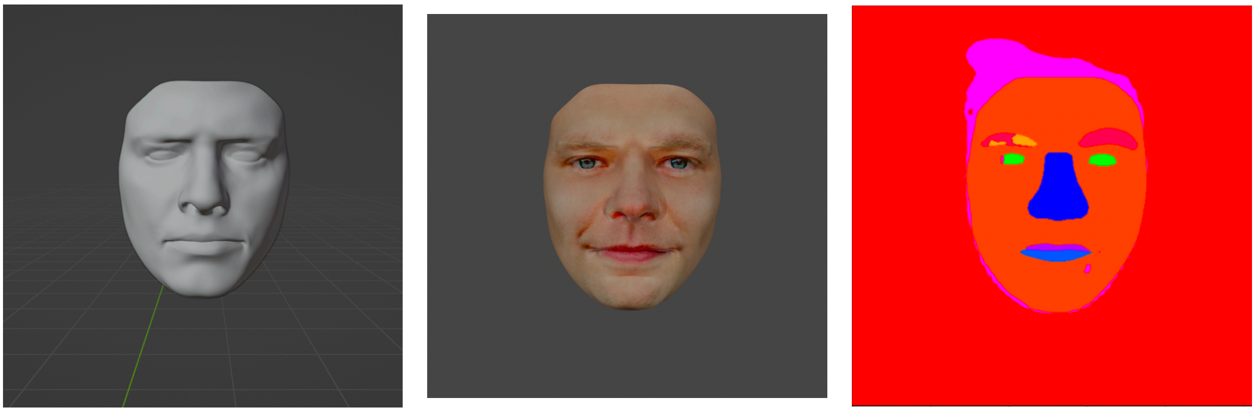
Following image capture, frontal and side images were rendered using the 3D imaging software 'Blender' [50] to match poses and lighting perfectly between samples. These pairs of rendered images were shown to a group of 10 annotators without prior knowledge of chronological age, following the protocol defined and validated in [108]. It is this set of perceived age annotations that were used for the initial genetic and morbidity analysis.

Perceived age annotations were taken for 2693, but since then almost twice as many more face images have been captured without annotation. We render these images using the same process as those shown to the human annotators, with the aim of incorporating them as part of the unsupervised training signal. In the previous study these 3D face captures were rendered manually requiring extensive human input. We take advantage of Blenders python API to automate the process, running it on the Erasmus HPC at greatly reduced time cost.

5.2.2 Preprocessing

The 3DMD setup used in this study was specifically designed to capture information about the face and facial skin, with less emphasis placed on the hair. For this reason, many of the images where individuals had curly or highly 3-dimensional hair contain significant artefacts such as missing or misplaced pixels. After rendering with Blender, we use a deep learning based facial semantic segmentation model (BiSeNet [237]) to identify relevant regions of the image. This segmentation allows us to computationally remove anything not classed as part of the face, namely the hair, neck and other background elements. Finally, we align the faces based on the location of the eyes to remove any minor variation in pose following our method outlined in chapter 3. A demonstration of the rendering and segmentation steps is shown in Figure 5.1.

Figure 5.1: 3D Face preprocessing pipeline.



(a) 3D Face

(b) Rendered Face

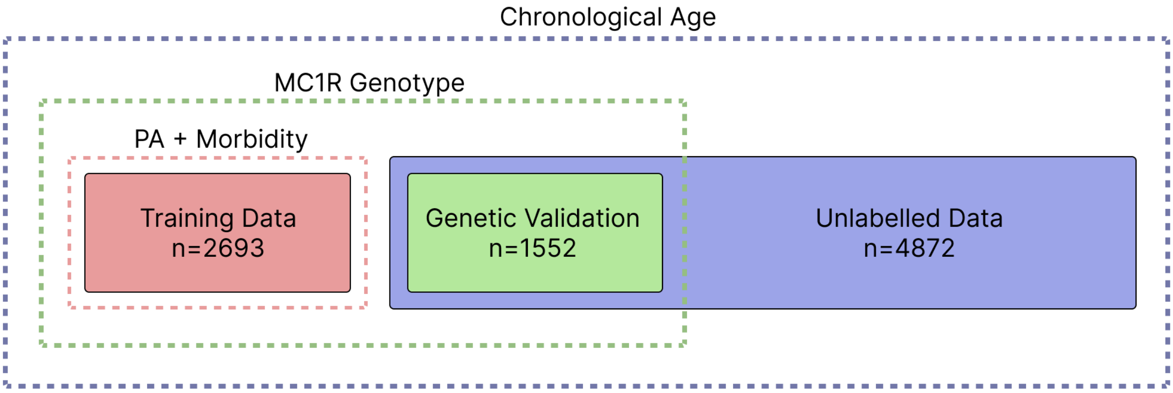
(c) Semantic Segmentation

5.2.3 Data Characteristics

The ERGO dataset has several novel features not found in any other perceived age dataset we access. The first key feature of this dataset is the presence of annotations for factors and outcomes of various age related diseases, in addition to genetic labelling for the recently discovered MC1R gene variants. These annotations allow us to complete an end-to-end validation of our perceived age predictions, ensuring that they are capturing more information related to the underlying ageing process than chronological age alone. As the Rotterdam Study is cohort based, images and annotations are produced in cohorts, such that several differently annotated subsets of data exits. As the underlying study is longitudinal and has collected data from multiple different cohorts, we must consider how to combine them into training and validation sets. So far a total of four cohorts have entered the study; RS1,2,3 and 4; where RS stands for 'Rotterdam Study'.

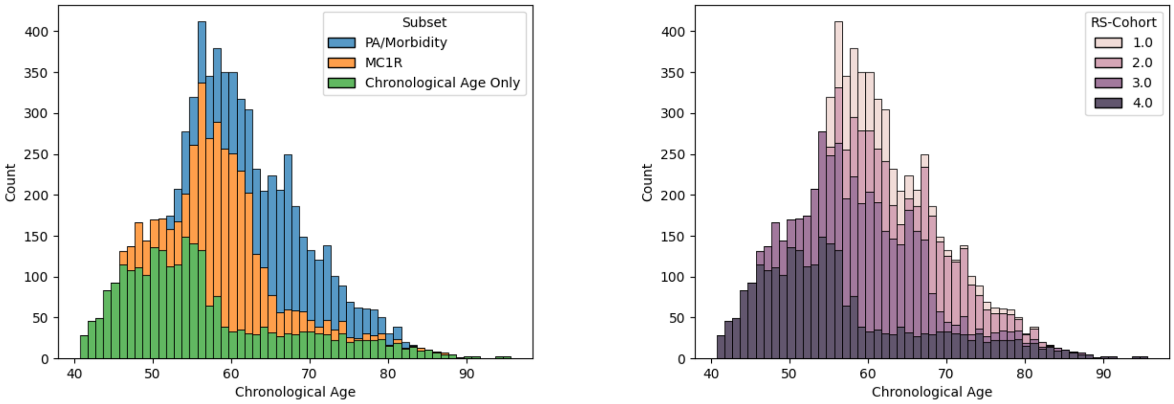
The images annotated with human PA were sampled from RS1,2 and 3; leaving the unlabelled set to contain the remaining images from RS1/2/3 with additional images from RS4. However, due to the significantly younger population sampled in RS4, we do not include them in the population used for MC1R association analysis. This leaves us with 3 distinct subsets of data: PA/Morbidity, MC1R and CA Only. PA/Morbidity is used for PA training and subsequent morbidity association analysis. MC1R and CA Only are used as unlabelled data for the unsupervised portion of the loss in our UDA model, with MC1R also being used to support the genetic association analysis. No other data was used to train or evaluate the final models proposed in this chapter, though other public data was used during method development (Section 5.4.1). To clarify the various subsets, we include a Venn diagram in Figure 5.2. Dashed outlines indicate the subsets of data with a given label type and filled rectangles represent each subset of data.

Figure 5.2: Venn Diagram Showing Labelling on the Erasmus Dataset.



To further illustrate the composition of the ERGO dataset, we include two histograms in Figure 5.3 showing the age of various subsets and cohorts present in the dataset.

Figure 5.3: Erasmus Subsets Chronological Age Distribution

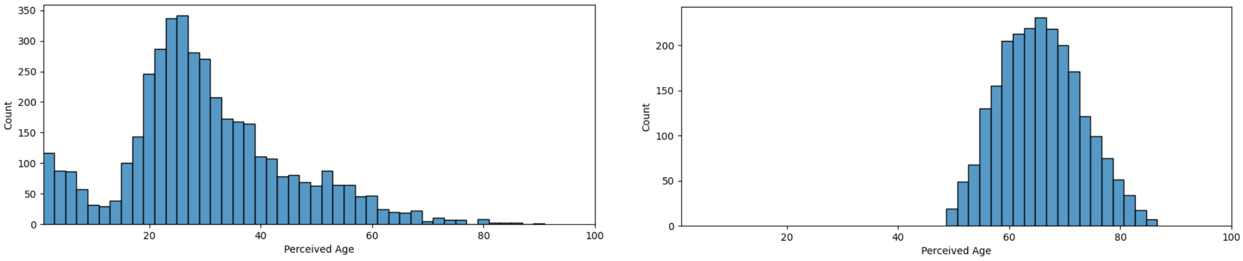


(a) Coloured by Our Subset Definitions

(b) Coloured by Cohort

The second feature of the Erasmus dataset is the significantly older population, with a mean PA of 65.6 (in the labelled subset) compared to 30.2 in the APPA-REAL [10] dataset. This difference in age distribution naturally alleviates the issue of poor prediction accuracy in the elderly by providing significantly more samples for the older age groups, even though there are fewer total samples. The distributions of perceived ages for both datasets are shown in Figure 5.4.

Figure 5.4: Class Distributions in Two PA Datasets



(a) APPA-REAL Training Set

(b) Erasmus Training Set

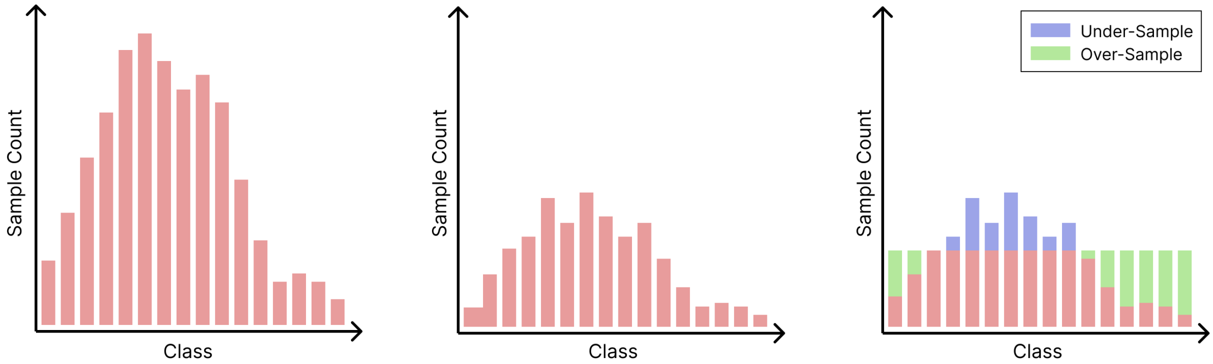
5.3 Semi-Supervised Learning

Deep neural networks suffer from overfitting in small datasets, leading to strong convergence in sample and poor generalisation out of sample. To alleviate this lack of consistency in unseen data researchers at google developed a new strategy, unsupervised data augmentation (UDA) [224]. This technique incorporates unlabelled data into the standard supervised training pipeline, allowing the CNN to learn to extract more robust features than with labelled images alone. Through each training iteration the model is presented with a batch of labelled samples, where loss is calculated between the ground truth and the model's output. In addition to this loss, pairs of unlabelled images are presented to the model, one image is the original sample and the other is an 'augmented' version; the unsupervised loss is calculated between the models output for the unaugmented and augmented sample. This approach effectively trains the model to produce the same response on unlabelled data regardless of the augmentation used, see Section 5.3.2

5.3.1 Balanced Mini-Batch Sampling

Classical UDA simply presents the unlabelled images in a random order with no consideration for sampling. We note that in the context of age estimation this does not produce satisfactory results, failing to improve on a standard supervised learning pipeline. For this reason, we add balanced mini-batch sampling (**BMS**) [202] to our training pipeline for both labelled and unlabelled samples. Labelled images are sampled uniformly by their perceived age using a combination of over and under sampling, such that on average the distribution of PA in each mini-batch tends toward uniform. Oversampling is done by duplicating samples, under-sampling randomly selects samples. Mini-batch uniformity is controlled by a sampling alpha parameter (**SA**), which varies between 0 and 1 where 1 forces complete uniformity. Unlabelled images are sampled using the same method but instead balancing the chronological age, as perceived age labels are not available. In this context, the aim of this method is to see if providing the model with a broad range of features in each mini-batch improves the UDA loss.

Figure 5.5: Illustration of Balanced Mini-Batch Sampling.



(a) Whole Dataset (b) Random Mini-Batch (c) Balanced Mini-Batch

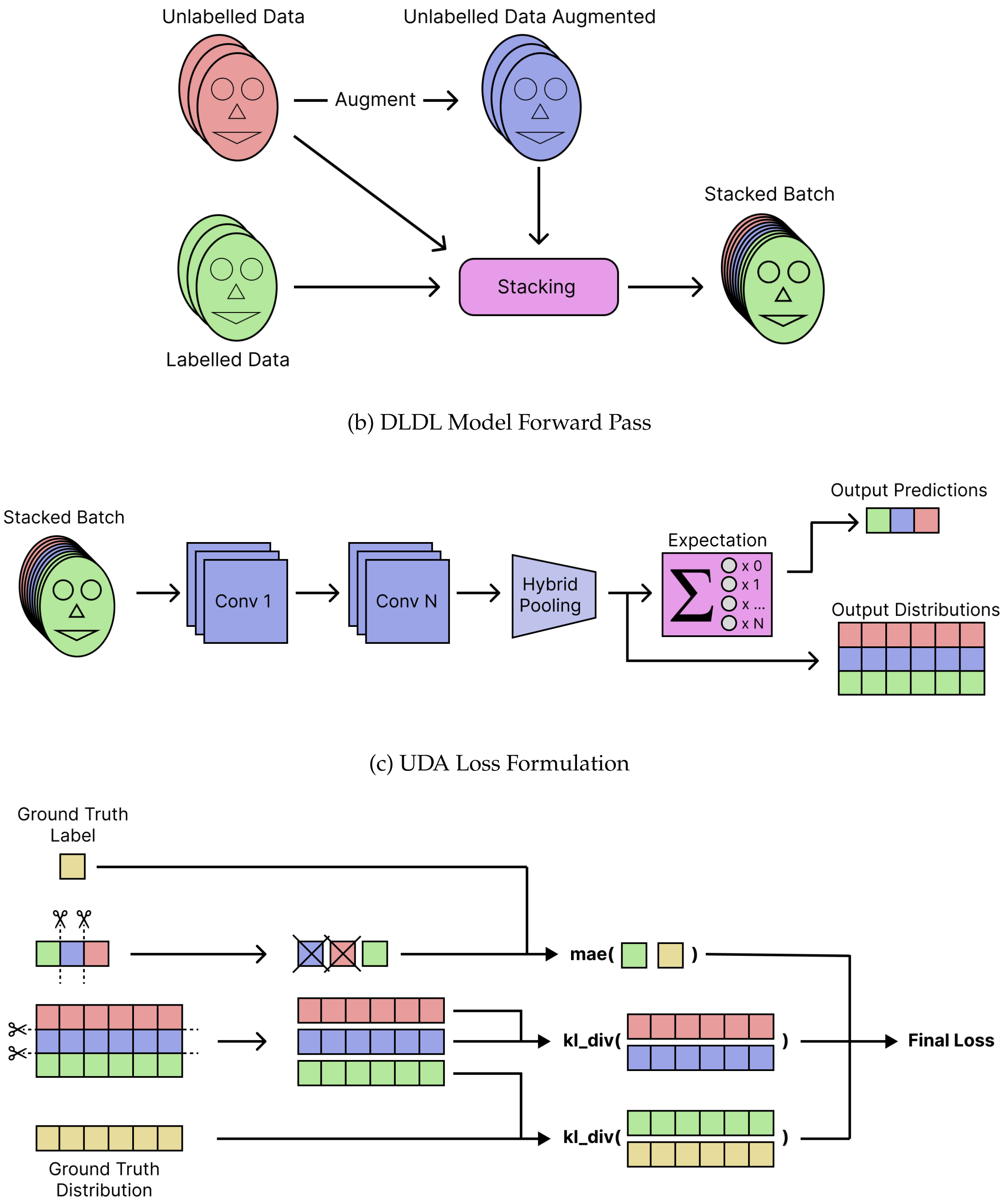
Unsupervised data augmentation can be used with any model which outputs a predictive distribution. We choose to use Deep Label Distribution Learning (DLDL) [92] as our supervised loss function, as it learns to output an age distribution and has recently been shown to outperform all other methods for perceived age prediction. Our final combination of these three methods is called 'Balanced Unsupervised Data Augmentation Deep Label Distribution Learning' or BUDA-DLDL.

5.3.2 UDA Implementation

Our UDA implementation follows exactly that presented in the original paper, with the only modifications being the addition of BMS and our own custom augmentation scheme, which we describe in Section 5.3.3. In this section we describe the overall UDA pipeline including the unsupervised and the DLDL supervised portions. Figure 5.6a shows how unlabelled images are first augmented then stacked into a batch of images including both version of the unlabelled and labelled images. Figure 5.6b shows a stacked batch of labelled, unlabelled and unlabelled-augmented images being passed through the deep learning model, with both predictive distributions and point estimates being returned as the expectation of said distributions. The outputs shown in Figure 5.6b are combined with the ground truth labels and passed to the loss function shown in Figure 5.6c. In all three diagrams red, blue and green are used to denote unlabelled, unlabelled augmented and labelled respectively. The final loss is the sum of the following three components:

1. The l1 distance between the labelled samples estimates and the ground truth.
2. The KL divergence between the ground truth distribution encoding and the predicted distribution for the labelled samples.
3. The KL divergence between the unlabelled and unlabelled augmented sample distributions.

Figure 5.6: High-Level UDA Pipeline



(a) UDA Augmented Batch Generation

(b) DLDL Model Forward Pass

(c) UDA Loss Formulation

5.3.3 Augmentation Scheme

The power of UDA and indeed most modern consistency based SSL approaches stems from well designed data augmentation. The correct augmentation in any given setting depends on semantic information about the task domain. For example, it is not useful to include vertical flip augmentation in the context of face analysis because facial positioning is controlled during preprocessing. Based on the assumption that face position is well controlled for, we choose not to include any translational augmentation in our SSL implementation, instead adopting the remaining 11 image augmentations presented in [55]. This scheme applies cutout [64] to every image in addition to a combination of the following augmentations with a probability between 0.2 and 0.8: AutoContrast, Equalize, Invert, Rotate, Posterize, Solarize, Color, Contrast, Brightness and Sharpness. We do not experiment with any of the magnitudes of these augmentations, simply using the same parameters as proposed for CIFAR-10 in [55].

5.3.4 Relevant Hyper-Parameters

In Chapter 2 (Section 2.6.2), we introduced UDA and its advanced features. In this section we review the hyper-parameters related to these features in addition to the sampling alpha parameter for mini-batch sampling described in Section 5.3.1.

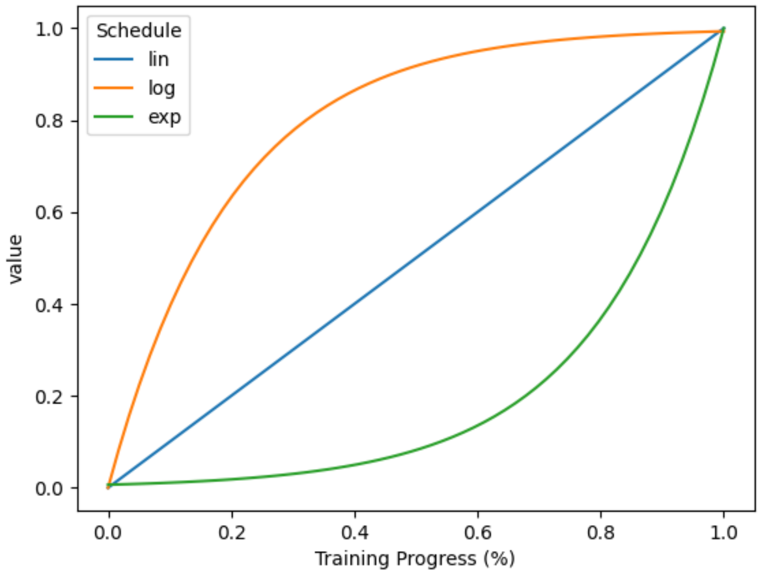
Sampling Alpha

The SA parameter controls to what degree of uniformity each mini-batch is sampled to. The possible range of values are any floating point number between 0 and 1. A value of zero samples randomly from the dataset, matching the original distribution. A value of one samples a perfectly uniform distribution of classes in each batch.

Training Signal Annealing

TSA is designed to reduce overfitting on easier labelled samples by setting their loss to zero and removing them from gradient calculations. Sample confidence is measured as the maximum value of any output neuron, with samples being classified as over-confident if the value is above a certain threshold. What is unique about UDA is the manner in which this confidence threshold varies as training progresses. The threshold is calculated as a function of the training progress in epochs, normalised to between 0 and 1. The original work proposes three different functions: linear (lin), logarithmic (log) and exponential (exp); visualisations of which can be seen in Figure 5.7.

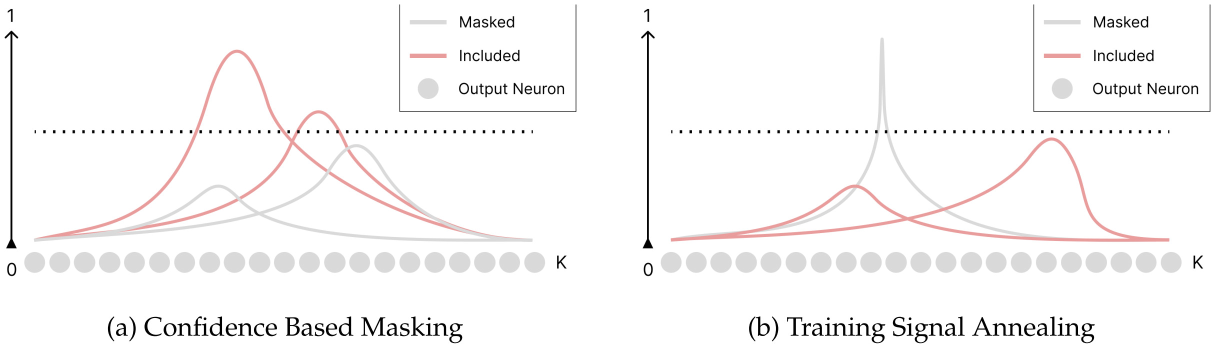
Figure 5.7: Training Signal Annealing Schedules



Confidence Based Masking

CBM acts as a threshold to prevent the model training on low-confidence unlabelled samples. This threshold unlike in TSA is fixed throughout the training process. Thus, the CBM hyper-parameter is a floating point value between 0 and 1 which is set at the start of training. In the case of age estimation via the distribution age encoding, the CBM threshold must be much lower than when a one-hot encoding is used. For this reason we see a value as low as 0.02 to be valid, compared to the 0.8 suggested in the original paper. To assist the reader in understanding the link between output values of neurons in our network and the TSA and CBM techniques, we include Figure 5.8.

Figure 5.8: Illustrations of CBM and TSA Thresholding



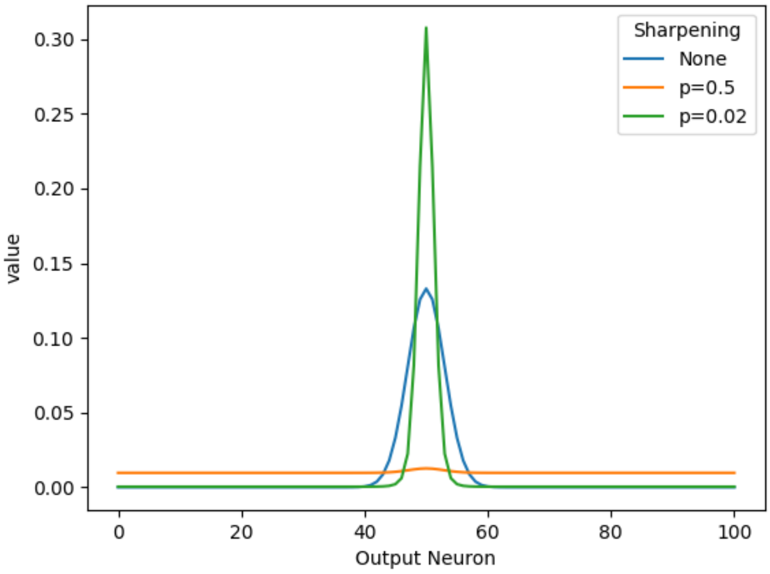
(a) Confidence Based Masking

(b) Training Signal Annealing

Sharpening

Distribution sharpen is applied to the models outputs for the unlabelled unaugmented samples, prior to the KL divergence loss between them and the augmented samples. Following the original authors we implement this as a softmax temperature value which occupies the real number space between 0 and 1. In the context of age estimation the softmax temperature must be considered differently to when it is used in classification. As sharpening re-applies the softmax function to the network outputs, it modifies the distribution shape even when the temperature is set to 1. For this reason extremely small STs are needed to actually have the effect of sharpening the distribution. The relationship between ST and output value is illustrated in Figure 5.13.

Figure 5.9: Output distribution for a 50 year old with a standard deviation of 3 years under various levels of sharpening.



5.4 Experimental Design

5.4.1 Method Development

Prior to evaluation on the Erasmus dataset we carry out a series of experiments to measure the impact of the proposed semi-supervised learning method on the APPA-REAL dataset, the same dataset used in Chapter 4. To make results as comparable with previous approaches as possible, we measure accuracy on the pre-defined validation set. Unlike in Chapter 4, we are not concerned about performance in the elderly, as the Erasmus datasets age distribution is not skewed away from our target age group.

In our developmental configuration we select the APPA-REAL validation set as our unlabelled dataset, replicating as closely as possible the segmentation of data in the ERGO images. For all experiments we use the identical standard hyper-parameters and preprocessing as in Chapter 4, including a batch size of 64, learning rate of 0.00005 and training epochs of 60. As stated in Section 5.3, our method has a number of additional hyper-parameters that must be tuned to optimise it for the task of perceived age prediction. We first optimise SA with TSA, CBM and ST disabled. We then fix the SA value and grid search over combinations of the remaining three parameters. The values used in the tuning experiments are shown in Table 5.10.

Table 5.10: Hyper-parameter Search Values

Hyper Parameter	Values		
	Option 1	Option 2	Option 3
SA	0.5	0.75	1.0
TSA	lin	log	exp
CBM	0.02	0.5	1
ST	0.01	0.02	0.03

Our final set of experiments are designed to ablate the performance contributions of UDA and BMS independently. We repeat 4 experimental configurations with 3 random seeds to compare the relative performance improvements of each feature of the proposed approach, shown in Table 5.11. In all experiments the CNN is initialised using the same weights pre-trained for face recognition, such that the varying seed values only impact stochasticity in the data augmentation and data sampling.

Table 5.11: Preliminary BUDA-DLDL Experiments

Experiment	Transfer Learning	UDA	Balanced Mini-Batch Sampling
Exp 1	Yes	No	No
Exp 2	Yes	Yes	No
Exp 3	Yes	No	Yes
Exp 4	Yes	Yes	Yes

5.4.2 Comparative Baselines

To fairly evaluate the impact of UDA in this context, we also train two baseline models following the methodologies presented in Chapter 4. The first is a standard DLDL model without UDA, which is directly comparable to the current state-of-the-art for perceived age prediction. The Erasmus dataset contains faces with resolutions over 500px, presenting an opportunity to further experiment with the DFT approach from Chapter 4. We previously found that the DFT approach has the unusual ability of being able to estimate ages with better accuracy when higher resolution input images were used. As it is difficult to find high resolution datasets with PA annotations, we seize the opportunity presented by the Erasmus data to test if this link holds true in a second PA dataset.

5.4.3 Validation in the Erasmus Dataset

Due to the small dataset size we choose to evaluate both methods using 10-fold cross validation. The dataset is split into 10 train/test sets with 90% and 10% of the images respectively. Each method is trained using the data in the larger subset and evaluated on the data in smaller subset. The predictions on the 10 test sets are combined to create a single set of deep learning-based predictions. We repeat this process with identical subsets of data for the Transfer Learning, UDA and DFT methods. We finally validate these 3 predicted PA signals using multiple regression analysis to confirm the same associations are found with morbidity as were found with the human PA ground truth.

5.5 Results

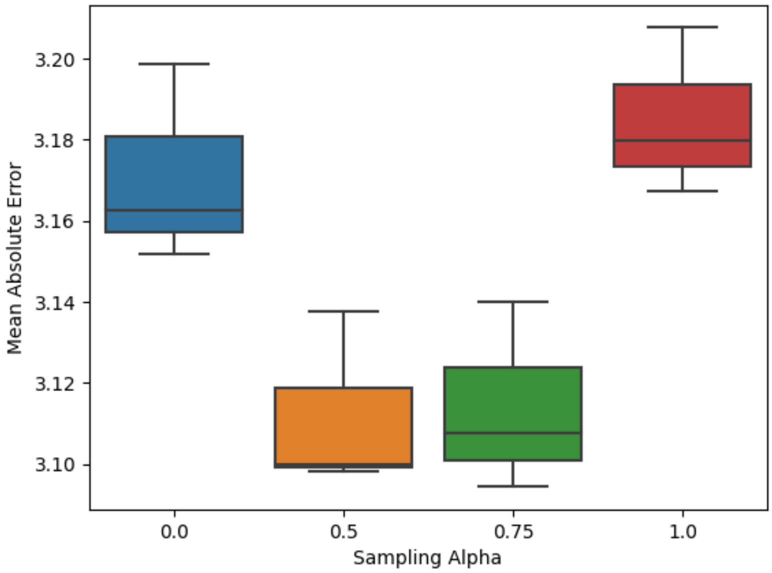
5.5.1 Semi-supervised Learning in the APPA-REAL Dataset

As access to the final medical validation datasets is limited, we develop methods first on the public Chalearn APPA-REAL dataset, which has been used throughout this thesis as the primary PA dataset. We first present the results used to assist in selecting optimal hyper-parameters for the SSL approach, before showing the independent contribution of UDA and BMS under a number of random seeds.

Hyper-Parameter Tuning

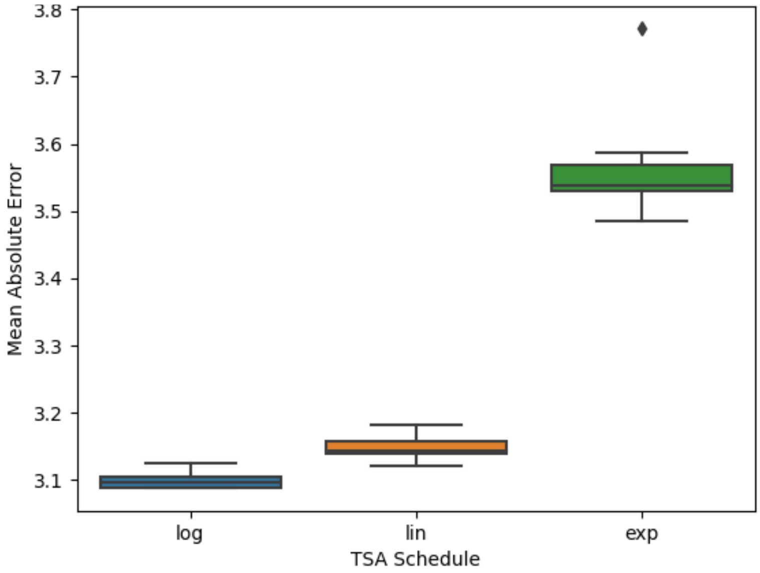
We first disable TSA, CBM and ST, optimising the sampling alpha value, for each sampling value we train 3 models under different random seeds. The results of these experiments are shown in Figure 5.12. Forcing a completely uniform distribution (SA=1) results in worse accuracy than when BMS is disabled. An SA of both 0.5 and 0.75 produce a similar improvement in MAE, though 0.75 reached a slightly better bottom end.

Figure 5.12: Sampling alpha value MAE under 3 random seeds.



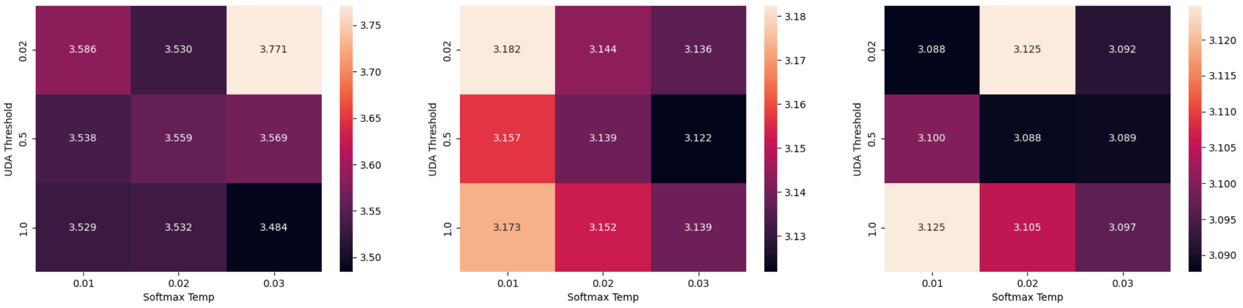
Following these experiments we fixed the SA at 0.75 and grid searched over 3 options for the remaining 3 parameters, resulting in 27 training runs and 27 resulting MAEs. We first plot the MAEs of all runs broken down by TSA schedule in Figure 5.13, from which it is clear to see that the exponential schedule is unsuitable but the linear and logarithmic schedules perform similarly.

Figure 5.13: Mean absolute error grouped by TSA Schedule across all other parameter settings.



We plot the remaining two hyper-parameters as heatmaps, with one heatmap per schedule to look for patterns. Each heatmap in Figure 5.14 has its own scale to allow for direction comparison of hot and cold spots. We see very little consistency between the optimal values for ST and CBM across the different TSA schedules, though in lin and log it does appear that a higher ST produces better results.

Figure 5.14: Heatmaps showing the MAE resulting from different CBM and ST values.



(a) Exponential (b) Linear (c) Logarithmic

Method Ablation

Finally, to better understand the individual contribution of BMS and UDA, we train the four different configurations defined in Table 5.11. We show a clear improvement in MAE with the addition of each component, and see that the combination of BMS and UDA outperform each independently. UDA alone outperforms TL but only by a small margin, both of which are outperformed by BMS alone. These findings suggest that our hyper-parameter tuning experiments improved accuracy by a reasonable degree, but it remains to be seen if this configuration will perform will in the elderly. We present the mean average MAE of three different random seeds in Table 5.16 and box plots of the same results in Figure 5.15.

Figure 5.15: Ablation of BUDA-DLDL Method on the Chalearn Dataset.

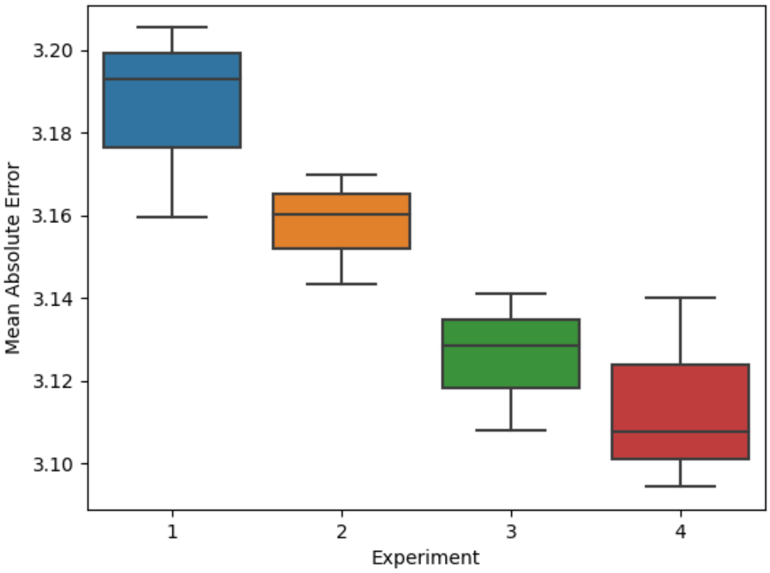


Table 5.16: Median MAE for each Ablation Experiment.

Experiment	Mean Absolute Error
Exp 1	3.19
Exp 2	3.16
Exp 3	3.13
Exp 4	3.11

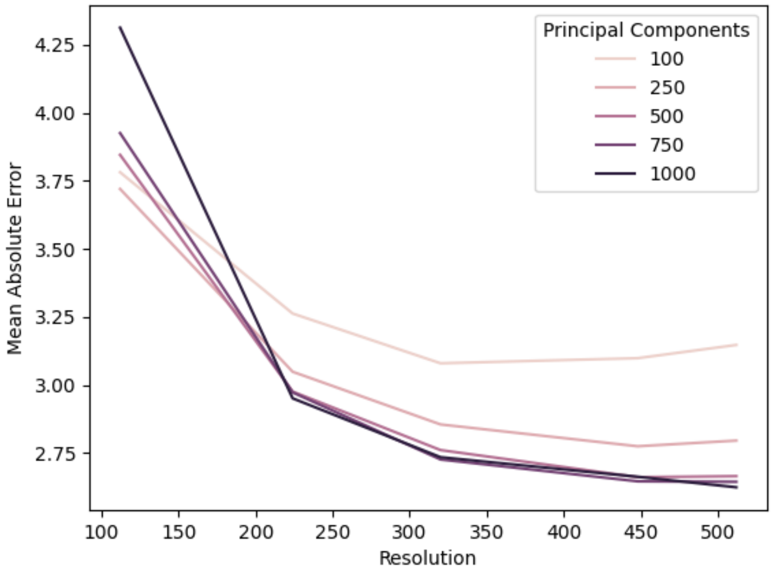
5.6 Application to the Erasmus Dataset

In this section we present results gathered from the Erasmus dataset, first evaluating our methods from a machine learning perspective using measures of MAE. We then go on to run the same statistical analysis as use previously to link our various estimates of PA with morbidity and genetics.

5.6.1 Confirmation of DFTs High-Resolution Properties

During our evaluation of DFT as a baseline for the SSL in the erasmus dataset, we carry out a similar set of experiments as in Chapter 4, this time only using Bayesian Ridge regression as it performed reasonably well in the 3 datasets used previously. We find a similar link between resolution and accuracy as seen previously, as well as confirming that a high number of PCs is required to reach a good result (Figure 5.17).

Figure 5.17: MAE of the DFT approach against resolution and PCA components.



5.6.2 Mean Absolute Error on the Erasmus Dataset

Perceived age predictions from both algorithms were compared to the mean perceived age as annotated by human accessors. Following the standard practice set out in the deep learning literature, we measure accuracy using mean absolute error.

As predictions are made using 10-fold cross validation, MAE values are calculated for each split as well as an overall value, shown in Table 5.18. Performance for all algorithms is relatively consistent across splits, indicating a fair distribution of features associated with each class. SSL is consistently marginally more accurate than DFT when compared to the Human annotated ground truth, which is reflected by an improvement in the overall MAE of 0.45 years.

Table 5.18: MAE per fold for each PA prediction method.

Fold	1	2	3	4	5	6	7	8	9	10	Mean
Method											
DFT	2.85	2.69	2.71	2.71	2.79	2.95	2.62	2.78	3.26	3.03	2.84
TL	2.22	2.16	2.46	2.59	2.53	2.55	2.35	2.42	2.57	2.41	2.42
SSL	2.23	2.16	2.43	2.56	2.49	2.43	2.22	2.38	2.52	2.45	2.39

5.6.3 Multiple Regression Analysis of Morbidity Associations

For automated perceived age prediction to be valuable in a medical setting, it's efficacy as a predictor for health must be validated. We carry out the same multiple regression analysis as in [171] using the algorithmic predicted PA in place of human PA. Two statistical models are created, Model 1 and Model 2, following the analysis done in [171]. Model 1 corrects for CA, sex and PA batch. Model 2 corrects for CA, sex and PA batch, smoking and BMI. As our analysis is not done with exactly same subset of data as previously, we reanalysis human perceived age under identical conditions.

Table 5.19 shows a breakdown of the associations between different PA estimates and the 5 morbidities previously found to associate with human PA. Table 5.19a shows the re-evaluated associations, showing that with our reduced sample size all associations except cataracts remain significant. Table 5.19b shows the associations of the DFT approach at 512px, the maximum face resolution were able to extract from the Erasmus data. All associations remain significant, even those which lost significance in human PA under this new modelling setting. We note that cataracts is only borderline significant under both models and Osteoporosis is only borderline significant under model 2. This contrast from the SSL approach in Table 5.19c, in which Osteoporosis is not significant under model two, but all other endpoints are significantly associated with PA under both models.

Table 5.19: Morbidity Associations for Human and Algorithmic PA in the Training Cohort

(a) Human

Model		Model 1		Model 2	
Outcome	n	Odds Ratio (95% CI)	P-value	Odds Ratio (95% CI)	P-value
Osteoporosis	2197	0.65 (0.53 - 0.8)	0.00	0.77 (0.62 - 0.96)	0.02
COPD	2282	0.67 (0.6 - 0.75)	0.00	0.76 (0.68 - 0.85)	0.00
Cataracts	2185	0.89 (0.77 - 1.02)	0.09	0.88 (0.77 - 1.02)	0.09

Outcome	n	B	P-value	B	P-value
ARHL	1945	-0.92 (-1.52 - -0.31)	0.00	-0.87 (-1.49 - -0.24)	0.01
Cognition	1864	0.07 (0.04 - 0.1)	0.00	0.06 (0.03 - 0.1)	0.00

(b) Deep Feature Transfer

Model		Model 1		Model 2	
Outcome	n	Odds Ratio (95% CI)	P-value	Odds Ratio (95% CI)	P-value
Osteoporosis	2197	0.64 (0.49 - 0.82)	0.00	0.76 (0.58 - 1.0)	0.05
COPD	2282	0.64 (0.56 - 0.73)	0.00	0.73 (0.64 - 0.84)	0.00
Cataracts	2185	0.84 (0.72 - 0.99)	0.04	0.84 (0.71 - 0.99)	0.04

Outcome	n	B	P-value	B	P-value
ARHL	1945	-0.97 (-1.69 - -0.25)	0.01	-0.94 (-1.68 - -0.19)	0.01
Cognition	1864	0.08 (0.05 - 0.12)	0.00	0.08 (0.04 - 0.11)	0.00

(c) Semi-Supervised Learning

Model		Model 1		Model 2	
Outcome	n	Odds Ratio (95% CI)	P-value	Odds Ratio (95% CI)	P-value
Osteoporosis	2197	0.71 (0.56 - 0.9)	0.01	0.86 (0.67 - 1.11)	0.24
COPD	2282	0.66 (0.58 - 0.74)	0.00	0.75 (0.66 - 0.85)	0.00
Cataracts	2185	0.83 (0.71 - 0.97)	0.02	0.82 (0.7 - 0.97)	0.02

Outcome	n	B	P-value	B	P-value
ARHL	1945	-1.02 (-1.7 - -0.34)	0.00	-0.99 (-1.69 - -0.28)	0.01
Cognition	1864	0.08 (0.04 - 0.11)	0.00	0.07 (0.03 - 0.11)	0.00

To allow for a more direct comparison between methods, we transpose our results and include PA estimations for both TL and DFT at low resolution in Table 5.20, 5.20a shows the effect size and 5.20b shows the significance values. Effect sizes are mostly the same across all methods, though UDA, TL and Human have slightly larger effects than the two DFT approaches. In Table 5.20b bold is used to represent insignificant values and 0.00 indicates a significance of less than 0.005. We call attention to the difference in significance of the Osteoporosis association with the DFT approach at different resolutions. Under Model 2, only human PA and DFT HQ are significantly associated, indicating that all other approaches are not capturing the ageing features relevant to this condition.

Table 5.20: Transposed Morbidity Associations Comparison.

(a) Effect Size (Odds Ratio/B-Value).

		Odds Ratio			B-Value	
Model	Feature	Osteoporosis	COPD	Cataracts	ARHL	Cognition
Model 1	UDA	0.71	0.66	0.83	-1.02	0.08
	TL	0.69	0.65	0.82	-1.06	0.08
	Human	0.65	0.67	0.89	-0.92	0.07
	DFT HQ	0.64	0.64	0.84	-0.97	0.08
	DFT LQ	0.66	0.75	0.84	-0.81	0.06
Model 2	UDA	0.86	0.75	0.82	-0.99	0.07
	TL	0.83	0.74	0.81	-1.02	0.07
	Human	0.77	0.76	0.88	-0.87	0.06
	DFT HQ	0.76	0.73	0.84	-0.94	0.08
	DFT LQ	0.83	0.85	0.83	-0.85	0.07

(b) Significance (p) - insignificant values shown in bold.

Model	Feature	Osteoporosis	COPD	Cataracts	ARHL	Cognition
Model 1	UDA	0.00	0.00	0.02	0.00	0.00
	TL	0.00	0.00	0.01	0.00	0.00
	Human	0.00	0.00	**0.09**	0.00	0.00
	DFT HQ	0.00	0.00	0.04	0.01	0.00
	DFT LQ	0.00	0.00	0.05	0.04	0.00
Model 2	UDA	**0.24**	0.00	0.02	0.01	0.00
	TL	**0.14**	0.00	0.01	0.00	0.00
	Human	0.02	0.00	**0.09**	0.01	0.00
	DFT HQ	0.05	0.00	0.04	0.01	0.00
	DFT LQ	**0.20**	0.02	0.05	0.03	0.00

5.6.4 Genetic Associations

Finally, we validate our predicted perceived ages against the known association between PA and variants of the MC1R gene [155]. Using exactly the same statistical analysis as previously, we compare the associations of our PA extraction methods with human PA in the same cohort as previously with the addition of the brand new validation cohort. Results for both the training (n=2693) and validation (n=1552) cohort are shown in Table 5.21. We see similar effect sizes in the training dataset for all methods, with only the DFT approach showing a slightly smaller effect than the rest. All approaches show a very significant association with gene in the training set but show reduced significance in the smaller validation set. All estimates remain significant in the validation set but with marginally lower effect size, indicating that our model does not have perfect generalisation, though without human annotations in the validation set this cannot be substantiated.

Table 5.21: Association between Perceived Ages and the MC1R Gene Composite

Perceived Age	Training (n=2692)			Validation (n=1552)		
	Estimate	P-value	Std	Estimate	P-value	Std
Human	-0.823	<0.001	0.235	-	-	-
DFT	-0.752	<0.001	0.230	-0.694	0.008	0.261
TL	-0.843	<0.001	0.210	-0.700	0.011	0.274
UDA	-0.835	<0.001	0.211	-0.713	0.009	0.274

5.7 Summary

In this chapter we apply the culmination of our research to a novel clinical-grade dataset with annotations for disease and a relevant genetic marker. We show that deep learning based PA estimation is a useful tool in an epidemiological research setting, and can be applied to images rendered from high resolution 3D face captures. We propose a new semi-supervised learning approach by combining the UDA loss with BMS and illuminate the parameter modifications needed to work well in the context of PA estimation. We use the TL and DFT approaches proposed in Chapter 4 as baselines and show that SSL outperforms them both in most settings. We seize the opportunity to exploit the high resolution nature of the Erasmus dataset to confirm the link between high resolution and accuracy in the DFT approach, reaching remarkable accuracy compared to previously, evening showing a stronger link with Osteoporosis than other deep learning approaches.

5.8 Conclusion

The results in this chapter show conclusively that deep learning PA captures the same links with health and genetics as human PA does. Our findings give us confidence that CNNs are well on their way to replacing human perception, even in subjective endpoints such as perceived age. It also gives us confidence in the validity of CNN PA in broader settings, showing that these models are not only predicting chronological age but are indeed capturing variance in the ageing process. In a context where images are cheaper to acquire than annotations, we show the value of deep unsupervised learning techniques for improving generalisation. Using additional unlabelled data is now not only an option, but often a requirement to reach state-of-the-art accuracy. We kindly thank Erasmus University for their collaboration and contribution of data to this project, we hope that with our technology their research will continue to make impactful discoveries in healthy ageing.

Chapter 6

Discussion

In this chapter we recapitulate our aims and objectives, analysing the strengths and limitations of the research presented in Chapters 3, 4 and 5. We outline our contributions to the field, calling attention to their significance in industry and academia. Flaws in the existing research are addressed, specifically pertaining to the lack of comparability, benchmark specificity and ablation. We structure the remainder of this chapter into four sections, the first three representing each of the methodological chapters, and the last containing general discussion. This structure is suitable given that each methodological chapter focuses on a distinct set of objectives, with very little overlap. What overlap does exist is addressed in the final section, covering a small number of interdependencies between our objectives as well as more general insights into the face age prediction research landscape.

6.1 Preprocessing for Facial Imagery

Face image preprocessing is unarguably a crucial component of age and perceived age prediction systems, implemented with a wide range of variability in the current literature. We set out to bring consistency to the existing preprocessing step of face alignment, as well as proposing a novel approach for confounder removal. These high level objectives were broken down into sub-objective, each of which we address in this section.

6.1.1 The Lack of Consistency in Face Alignment

The initial motivation to explore face alignment stems from preliminary results generated by a Newcastle University Masters student. They applied the pretrained model presented in DEX to a lab-grade dataset of images of elderly faces, noting a significant reduction in accuracy compared to that claimed in the publication. We carried out further work to explore their implementation and discovered that the face alignment used prior to feeding images into the CNN does not match that found in the original paper. This finding calls to attention a feature of face age prediction pipelines which is never noted in their respective publications: models are not invariant to face alignment, even when the alignment is completely reasonable (face fills the frame without losing any features). The implications of this simple finding are significant, limiting the comparability of results, lowering the impact of novel modelling and adding a barrier between academic results and industry applications. To make the situation worse, we undertake a broad literature search and compare the implementation details of face alignment in over 50 face analysis publications, finding that not only are implementations highly inconsistent, but often poorly or entirely undocumented. Given the importance of correct face alignment, and the lack of consistency in the literature, the impact of a reusable framework is clear.

6.1.2 A Framework for Flexible and Reproducible Face Alignment

In our paper "Look Me in the Eyes: The Importance of Canonical Face Positioning in Face Analysis Pipelines" we propose a framework to reduce the error inducing variance in face alignment while still allowing for flexibility in the underlying algorithmic implementation. Flexibility is key to ensure the accessibility of our contributions, ensuring that our framework is suitable regardless of the constraints of the downstream users. We also hold accuracy in high regard, as improvements in compatibility and accessibility will ultimately fall to the wayside in the wake of suboptimal results.

We give thanks to [200] for their well architected aggregation of face alignment backends, as it provided a starting point and strong foundation for our codebase. We in turn make our framework publicly available on github with a highly permissive licence, further strengthening our contribution. In order to develop a system which is able to align faces to near identical positions, regardless of the underlying componentry, one must first understand why different components produce different results. We align a number of different face datasets using the bounding boxes output by 5 different face detection algorithms and use ORB feature matching to calculate the affine transform between each. While some random effects were seen, which likely stem from the weaknesses of each algorithm, we see clear systematic biases associated with each approach. Without retraining every approach we cannot provide concrete evidence, but it is our belief that these biases come from the underlying bounding boxes in the original datasets on which they were trained. This is an uncontroversial assumption given different datasets are annotated by different annotators under different heuristics, some insist on defining square bounding boxes while others allow unconstrained rectangles.

Eye based alignment is a strong solution for the inconsistencies present in face box detection algorithms. This is because the same level of ambiguity is not present in facial keypoint algorithms, where eye keypoints are located in near identical locations in every FKP dataset. Our framework uses the position of the eyes in combination with two novel parameters to define a new face bounding box as well as an affine transformation to rotate the eyes to equal heights in the image. Under the assumption that the face should be horizontally centred, we add parameters for the scale and the vertical position of the face, making further assumptions about the default values of these parameters based on manual review. Powering our eye based alignment are the 6 alignment backends from the aforementioned deepface [200] library.

We validate our framework using 5 different datasets covering 3 different tasks, two CA, one PA and two FBP. Each dataset is aligned using every alignment backend, from which two different CNN architectures are trained. Finally, we cross-compare the accuracy of each trained CNN when tested on data aligned with a different backend. The validation experiments measure the impact of eye based alignment on the MAE of face regression models using bound box alignment as a baseline, showing improvements in both accuracy and compatibility in several settings designed to emulate real world constraints. We find that not only does eye based alignment almost entirely remove the incompatibility between alignment backends, but also improves the accuracy of downstream models when the same backend is used. The positive results of our validation indicate that our objectives as a whole have been met, such that we proudly present our framework with the hope that it will have a significant impact in academia and industry.

In the process of completing our primary objectives we also uncover several interesting secondary findings, mostly relevant to the practical application of face alignment. Our first and most significant unexpected finding is the poor performance of OpenCVs Haar Cascade based face and eye detectors. The Haar Cascade approach for object detection is still popular, which we assume is simply due to its extremely low computational cost and inclusion in the standard OpenCV package, which is most computer vision practitioners first introduction to the field. Regardless of its popularity, without fine-tuning detection settings it is by far the least reliable face detector in our experiments, failing to detect faces in over half of the images in some more challenging datasets. For comparison, all other algorithms achieved success rates of over 90%, with the best performing at over 99% in every setting. For this reason we recommend the exclusion of Haar Cascade based face detectors in future work on face analysis, unless the deployment criteria explicitly requires it.

Secondly, we find that MTCNN is the most reliable face and keypoint detector of those tested, while also inducing competitive accuracy levels in downstream models. It is perhaps for this reason, while not reported, that MTCNN has become the most popular alignment backend for works published in recent years (see Table 2.11). This does not indicate that all other alignment backends are useless as each may be more or less suitable for a specific deployment environment, but that indeed MTCNN makes a good choice in a research setting where computational resources are not limited and speed of alignment is not paramount.

Finally, we show that the recent CNN architectural development presented in [74] reaches previously unseen accuracies in both perceived age estimation and facial beauty estimation when combined with our eye based alignment. Improving on the MAEs presented in [91] by a significant margin with little to no additional computational cost. It was not our intention to make this discovery during these experiments, but it is nonetheless noteworthy and should be considered as a possible avenue for future work.

6.2 Transfer Learning Strategies for Perceived Age

In Chapter 4 we explore two very different approaches for transferring knowledge from face recognition to PA estimation. The first approach (TL) follows the protocols outlined in the literature, fine-tuning the CNN weights through a number of stages and datasets. The second approach (DFT) uses a CNN trained for face recognition to extract features which are then used to train a secondary classical ML model, an approach which is not well tested in the context of PA estimation. These two approaches share the same initial FR pre-training but diverge from there, having individual strengths and weaknesses stemming from their unique implementations. The strength of the DFT approach is not its absolute accuracy in PA estimation, but instead its generality and efficiency. We see DFT as a promising avenue for the exploration of new and unknown face analysis tasks, which may exist only in small datasets with few samples per class. Given a dataset with a number of facial annotations, FR features must only be extracted once, from which ML models can be trained rapidly to assess if the feature space contains a signal relevant to a given annotation. TL on the other hand has been studied in depth in the context of PA estimation, with many works proposing their own formulation. In our experiments TL reaches accuracies far exceeding that of DFT, making it unquestionably the best choice where absolute performance is concerned. TL however requires a complex pipeline with multiple pre-training stages, with additional domain specific data needed to reach state-of-the-art results. In the remainder of this section we discuss the most interesting findings from out modelling work.

6.2.1 Remarkable Performance in the APPA-REAL Dataset

We show that fine-tuning for PA estimation using the IMDB-Clean and MS1M-V2 for pre-training reaches a new state-of-the-art in the APPA-REAL validation set. We use the same training formulation and loss function as in the original DLDLv2 paper, changing only the CNN backbone and adding chronological age as an additional pre-training stage. Our formulation exceeds their results in the APPA-REAL validation set by 0.62 years, a margin which is surprising given the low operating resolution of 112px compared to their 224px. While the ARCFace loss function we use for pre-training is newer and more advanced than the softmax approach used in the original paper, it may not be the only factor contributing to our improved performance. As we found in Chapter 3, the iResNet architecture outperforms the original DLDL backbone (TinyAge) by a remarkable margin, showing that even when the same pre-training is used the backbone is still a performance bottleneck. It is interesting that our results challenge current thinking regarding backbone selection for age and perceived age estimation, such that we suggest further work reconsider the use of residual networks for this task.

6.2.2 Pre-training for Accuracy in the Elderly

In addition to developing a transfer learning scheme for PA prediction over the entire age range, we aimed to better understand how TL contributes to the accuracy of predictions in minority classes, specifically in the elderly. We used two datasets filtered to contain only elderly faces, one is a subset of the APPA-REAL dataset with very few samples and one we create ourselves as the aggregation of four other datasets. It is important to note that while our motivation is to improve estimation accuracy in the elderly, this approach may be applied to any regression task suffering from class imbalance. For example, it could be applied to estimation in children, or to a different task entirely such as facial beauty.

Regarding PA however, we make some interesting findings. Firstly, by plotting the estimation errors broken down by class we confirm the suspected bias induced due to class imbalance, with a general trend to under-estimate ages in the elderly. We find that this is most significant when chronological age pre-training is not used. This finding is more complex than it initially appears, as we see the a similar degree of imbalance in both our PA and CA datasets (Figure 4.6). As shown in Figure 4.11, training with CA alone produces a remarkably unbiased model, indicating that once a critical number of samples for each class is achieved the level of imbalance becomes less important.

We showed that fine-tuning with the FR_CA_PA configuration produces a reasonably unbiased classifier with an elderly MAE of 4.39 and overall MAE of 2.84. We do however significantly improve on this elderly MAE by fine-tuning on the elderly PA subset with the FR_CA_EPA configuration, reaching and elderly MAE of 2.82 with a worsened overall MAE of 4.76. While this reduction in overall MAE is not desirable, it is remarkably small compared to when we include the ECA dataset, which in all cases produces an overall MAE over 30. Our results clearly show that elderly specialisation should be induced during the final perceived age fine-tuning stage, and that the inclusion of elderly specialised CA data only detracts from performance.

6.2.3 State-of-the-Art Results at Very Low Resolution

When FR pre-training is needed we make use of the MS1M-V2 data, in which most images are extremely low resolution. Standard face recognition models operate at 112px or below, as such we follow the literature and train our two backbones (TinyAge and iResNet) at this low resolution. Following FR pre-training we are faced with two choices, continue with CA and PA training at this resolution, or to attempt to transfer these low resolution features to the higher resolution age datasets. Training at 112px is nearly 4x faster than a 224px due to number of computations required to convolve over a 2x large pixel space, motivating us to continue with low resolution training under the assumption it may still produce strong results, as were found in [31] at 120px.

Our results indeed indicate that strong PA estimation performance is not reserved for higher resolutions CNNs, with our lower resolution approach reaching and exceeding the current state-of-the-art. This finding goes against findings in other computer vision tasks, where increased resolution tends to improve performance [139, 195]. We do not make a direct comparison between optimised models at 112px and 224px, but simply suggest that the standard operating resolution of CNNs may be excessive in the context of PA estimation.

6.2.4 Face Recognition Feature Extractors in Controlled Datasets

Arguably the most surprising finding in this thesis is the relatively strong performance of our DFT approach at higher input resolution. For reference, high resolution in the context of DFT far exceeds the normal operating resolution of normal CNNs which most often do not exceed 224px. In contrast we show that DFT is cheap and efficient to train using image sizes ranging up to 800px, and may even be efficient at higher resolutions though we do not have the data to support this. One confounding effect we are not able to control for in these results is the setting in which our high resolution data is placed. All 3 high resolution datasets we explore (CFD, Unilever/Leiden, Erasmus) are taken in a lab setting with highly controlled lighting, pose and background. For this reason we cannot be sure if it is the high source resolution or the controlled nature of these images contributing to the improved accuracy.

At present we cannot give a good explanation for why CNN filters learned on faces at low resolution can efficiently extract information at higher resolution. It is possible that there is some inherent scale invariance provided by the convolutional prior, though this is not supported by the existing literature [228]. Another explanation is that the shapes and textures in both low and high resolution images are similar, a feature that the model may have learned through being exposed to some downscaled images during FR pre-training. One could argue that at different resolutions facial texture such as wrinkles appear totally differently, though this argument largely cannot be made for shape.

6.3 Health and Genetics

Regarding the application of deep learning to health and genetics, our aims were clear and narrowly focussed. While the modelling and preprocessing portions of our work were completed in parallel, we waited until the most relevant research questions were answered before applying our technology to health. Two models with strong performance on the publicly accessible PA dataset [chalearn] were evaluated using data from the Rotterdam cohort study. Each model was trained and tested using 10-fold cross-validation on their small (n=2693) face image dataset, no fine-tuning was done as to not bias the results, training using identical hyper-parameters to those shown to be effective for a range of tasks (Chapter 3). Following training and the extraction of PA predictions, we measured the association of deep learning PA with several morbidities and the MC1R gene. These association results were compared with those of the human annotations to make judgements about the efficacy of the technique for capturing features relevant to health and wellbeing.

The statistical methods used to measure the link between predicted PA values and health related endpoints are key to arguing the validity of PA in a medical context. This is due to the link between chronological age and health, when considered alongside the strong correlation between PA and CA. We look for evidence that predictions of PA from the face provide useful information which is more useful than chronological age alone. Our statistical analysis includes chronological age, gender and smoking status as covariates to PA, fitting logistic and linear regression models to binary and continuous morbidity endpoints respectively. We find that deep learning predicted PA is a significant component in all 5 of the morbidities previously linked with human annotated PA, with some nuances which we discuss in this section. We discuss the strengths and weaknesses of our two approaches, the challenges associated with 3D clinical grade data and finally the competencies and limitations of deep PA prediction as a replacement for human annotators.

6.3.1 Technical Comparison of SSL and DFT

In [ref chapter/section] we compare the associations of two distinct deep learning based methods for perceived age estimation, DFT and SSL. The advantage of the DFT approach is its ability to process relatively high-resolution images, ranging up to 640 pixels compared to the mere 112 pixels square accepted by the SSL model. We note that at a resolution as small as 112 pixels, the model is forced to ignore more granular features such as pigmentation, fine wrinkles and minor sagging. To better understand the benefits of using more granular features we compare the associations of the DFT model at both its maximum (optimal) resolution with the resolution at which the SSL method operates (reference table). Minor reductions in significance are seen across all morbidities, but most interestingly we see a clear loss in significance in the associations with osteoporosis. Of all methods we evaluate, only the high-resolution approach is able to capture features associated with osteoporosis, indicating that this condition may be visible only in more fine-grained details.

It is interesting that even though high resolution DFT is more significantly linked to osteoporosis, it is generally a less accurate approach to perceived age prediction. When measured by mean absolute error DFT is significantly less accurate than SSL in the Erasmus dataset, showing results of 2.84 and 2.39 respectively. Not only does DTF underperform compared to SSL, but it is also significantly worse than our baseline deep learning approach which reached an MAE of 2.42. A finding that is reflected to an even greater extent in our experiments applying DFT to the APPA-REAL dataset (Chapter 4). In fact, the strength of DFT in this setting compared to uncontrolled settings is worth consideration, as it is likely exploiting the highly controlled nature of this 3D rendered dataset. DFT uses features learned from a large uncontrolled dataset for face recognition, producing an highly dimensional embedding space where images of the same individual are clustered close together. It is unclear why such an embedding space would be able to extract more robust features for PA prediction when highly controlled images are used. One possible explanation is the removal of pixel variance due to pose and lighting, allowing the model to only capture information relevant to the morphology and texture of the face. This information is not densely represented in the models outputs, but instead requires PCA and a further regression model to extract a final estimation.

6.3.2 The Unique Nature of 3D Face Captures

While the methods we propose generalise to other datasets with entirely different distributions of features, the model trained in this work is limited to inference on images generated with our 3D imaging pipeline. Images shot with the 3DMD camera and preprocessed with semantic segmentation contain a unique set of computer vision features when compared to images shot with standard 2D cameras. For this reason, the replication of our model in other settings may require additional 2D data to be gathered and annotated by humans. We do not make any claims about the difference between highly controlled images shot with 2D or 3D imaging, as we do not have the data to support this. However, we are able to say that the images in the Erasmus dataset allow for a significantly more accurate perceived age model than 'in-the-wild' images, even with fewer samples. Furthermore, 3D images contain more information about lighting and shape, meaning that any additional images shot in 3D can be easily preprocessed to match the current dataset. The same cannot be said for lab-grade 2D images, which require more careful consideration of the lighting, background, camera settings and pose. One alternative to this may be images shot with the VISIA system, which includes a chin rest allowing for fixed pose, background and lighting.

In summary, we see 3D images as a promising and future-proof approach to capturing faces for analysis in a healthcare setting. While we do not have the data to suggest that they are more effective training data for deep learning than 2D images, we can say that they provide more flexibility and support consistency when data is gathered across multiple time frames.

6.3.3 Deep Learning as a Replacement for Human Annotators

The primary motivating factor behind this thesis is the cumbersome and costly nature of perceived age annotation. Inspired by the success of models in the Chalearn 2015 PA prediction challenge, we set out to test if deep learning based approaches could replace human annotators and accelerate biological research. From an accuracy perspective deep learning has significantly narrowed the gap between human and computational methods, to the extent that a single deep learning model produces predictions that lay within the standard error of a group of 10 annotators. In this section we review the impact and limitations of our approach with regard to the replacement of human annotators.

Impact

Training costs associated with the deep neural network are non-trivial, requiring a modest Linux server with GPU computing capabilities. Once the system is trained however, prediction of perceived ages can be run on very lightweight devices such as a basic office desktop. The cheapness of this approach unlocks broad swathes of research on perceived age that have previously been seen as cost prohibitive. Morbidity analysis and GWAS studies can be executed in exponentially larger populations, without the requirement for additional ratings of human PA. Furthermore, the potential of our method far exceeds the controlled settings of the 3DMD photography system. Given enough 'in-the-wild' data, such a system could be trained to predict PA from smartphone or selfie images, opening the potential for entirely non-contact population studies.

Limitations

Given the strong performance of deep learning for PA prediction and its associations with health and genetic markers, one may be tempted to assume it is ready for deployment into laboratory and clinical environments. We note several key limitations which must be addressed before our models can be considered robust enough to be applied to face images in datasets containing different demographics and image capture devices.

The first limitation is the bias of our models with regard to age and ethnicity, stemming from the demographic composition of the Erasmus dataset. The Erasmus dataset is sampled from the elderly dutch population, containing faces ranging from 50 to 80 years old, with the majority of individuals being aged between 60 and 70. All faces are of white European descent as ethnic origin was not addressed during the candidate selection phase. These biases in combination result in a model that performs well for white, middle-aged or elderly individuals, but is likely to produce poor predictions on individuals outside of this demographic. The presence of these biases does not detract from our findings, as we reasonably assume that facial ageing links to health across other demographics given the correct dataset. It does however limit the applicability of the models resulting from this data, which is unfortunately an unavoidable constraint. We believe that generating face image datasets with uniform demographics is of the utmost importance, particularly regarding gender, age and ethnic origin.

Secondly, as mentioned in section [ref 3d faces section], the novel nature of 3D face captures is both a strength and limitation of our work. While the benefits of using 3D face capture systems are clear, models trained on this highly normalised and controlled data are not robust in uncontrolled settings. Furthermore, the 3DMD camera system is neither cheap nor portable, requiring an extensive install and dedicated space. In comparison, allowing healthcare professionals to take photographs using standard digital cameras against a neutral backdrop is far less complex and expensive. We note that more images with PA labels exist in-the-wild than in our lab dataset, however none exist with accompanying genetic or morbidity data, limiting our ability to make a direct comparison between an unconstrained CNN and our lab constrained CNN. This indicates with present technology and data, it may already be possible to create such a system, but without additional morbidity or genetic data, it cannot be validated.

The final limitation is neither technological nor is it related to bias, but is instead an ethical limitation. The Erasmus dataset, and in fairness any other medical dataset, comes with highly restrictive privacy terms. These terms are such that the data, and any models resulting from the data cannot be released into the public domain without significant evidence of their anonymity. This leads to a research environment where data is collected in silos which are unable to integrate, trading only with methods and reported findings. Without the ability to integrate face images from different institutions in different localities, it is unlikely that a single model accounting for all demographics will be created. We suggest that an international collaboration is needed, where patients are required to consent for their data to be shared across national boundaries, a condition which we have yet to see included as standard.

6.4 General Discussion

In this section we provide an overview of the discussion points mentioned in this chapter, designed to provide a more high level window into our thinking following the projects completion. We follow the academic discussion structure, first presenting the strengths and weaknesses of out work, then presenting the broader implications of our findings and finishing with a reflection on our research activities and methodologies.

6.4.1 Strengths

The work presented in Chapter 3 is in our opinion our strongest methodological contribution to the field, clearly and systematically evaluating inconsistencies in the face age estimation literature. We review the most popular approaches for face alignment, gathering results which indicate the best and worst performers. In the process of this work, we develop a framework for highly reproducible face alignment and make it available to the wider community on Github. In a similar fashion, the work presented in Chapter 4 addresses inconsistencies in the age estimation literature regarding transfer learning. In this work, we present a concise evaluation of the various stages used as pre-training for perceived age estimation, finding an optimal configuration which we show to outperform all other methods on the APPA-REAL dataset. We also explore a more tangential approach for PA estimation via the use of a fixed deep face recognition feature extractor, showing that useful features can be extracted at higher resolutions than intended. This finding was unexpected and holds significant value in motivating further work. We further evaluate the DFT approach in Chapter 5 using a high resolution highly controlled 3D dataset, finding the same interesting properties are even more significant in this data, reaching a remarkable MAE of 2.84. Using the same dataset we train a novel semi-supervised learning approach which allowed us to make use of unlabelled data from the same 3D image domain as the labelled set. This approach showed a marginal improvement in MAE in both the APPA-REAL and Erasmus datasets. Finally, we validated our algorithmically estimated PA values against objective health and genetic endpoints, showing significant associations with 5 age related diseases and the MC1R gene. Overall we achieved all of our individual research objectives and showed significant contributions toward our high level aims.

6.4.2 Limitations

While our contributions to face alignment for preprocessing are both valuable and impactful, there are some clear limitations to this work that should be addressed. During our literature review we note that an alternative alignment procedure to our proposed eye based alignment is gaining popularity. Namely 'similarity transform based' methods, a class of procedures that minimises the variance of a larger set of facial keypoints, as opposed to our approach which only uses the eyes. This approach may create a bigger reduction in MAE as it is more robust to errors in the positioning of eye keypoints. Another clear limitation of our work on preprocessing is the small size of the CNN backbones we use. Even though these backbones are able to reach and in some cases exceed state-of-the-art accuracies, it is possible that larger CNNs with more parameters may be able to learn face representations which are more invariant to the alignment used. We do not believe that these extensions will invalidate our work, but instead provide more insight true nature of face alignment.

In our opinion Chapter 4 presents the weakest contributions, especially in the context of several recent advancements pertaining to possible pre-training schemes. Our work explores the PA transfer learning formulations as it most commonly exists in the literature. We replicate the methods of DLDL with the addition of a new pre-training dataset and loss for face recognition, the addition of CA pre-training which they do not include in their original work, and the use of a different backbone (iResNet18). While we do show that this formulation exceeds their results, and indeed the current state-of-the-art on the Chalearn 16/APPA-REAL dataset, we do not apply this formulation to other datasets. Furthermore, in the past year two approaches [186, 251] have been developed to train robust face embedding spaces which go further than the face recognition pre-training we apply. These recent works take two very different approaches to train highly general transfer based models for face analysis tasks, which we believe likely constitute much better initialisations for PA estimation. Furthermore, these models and works building on them use transformers as a key building block. The broader face estimation literature has largely avoided using transformer architectures, instead favouring variations of more classical CNNs. We assume that as these newer transformer architectures, such as the ViT, may continue to improve above the contributions we have made. To the best of our knowledge the leading accuracies in deep learning based age estimation are still owned by CNN based models. The incorporation of these backbones into our work on transfer learning would provide a much richer and more current exploration of the technique, challenging the findings in Chapter 4. To address gap, we suggest a more comprehensive set of experiments, including both more variety in general pre-training as well as the inclusion of the Chalearn 15 PA dataset make our results more comparable with legacy approaches.

In the final methodological chapter we proudly present the first work on the application of automated PA estimation in an epidemiological study, showing that deep learning estimated PAs share the same links with health as human PA. While these findings are strong within the dataset they are developed, the demographic composition of this data makes the trained models weaker in a broader context. As the Erasmus data is primarily elderly caucasian individuals, the models we train should not be reused in other demographics, limiting the impact of our work. It is difficult to alleviate this issue as we are not aware of the existence of any other similar datasets, such that further work would also require the acquisition of more 3D face images. We also do not test the models trained on 3D data in standard 'in-the-wild' PA data such as the APPA-REAL dataset, but expect to see a significant performance shortfall in this case. Finally, we see exciting opportunities to revisit face analysis in the Erasmus dataset, making use of their morbidity and genetic labels to train classifiers directly, rather than looking for associations via proxy of perceived age. Training CNNs directly on these endpoints may be a better approach for both accuracy and interpretability, as there is much information in PA that may not be associated with health.

In summary, we find clear limitations across the various chapters presented in this thesis, some which likely have little impact on the implications of our work and others that may be more significant.

6.4.3 Broader Implications

During this project we make several discoveries with wide ranging implications in academia and industry. Here we discuss those implications generally, beginning without the impact of our face alignment framework.

We question the need for such a wide range of face alignment in the age estimation literature, finding that not only is it inconsistent but often not well communicated in recent publications. Our work on face alignment, including the framework we propose has potential implications across all face analysis tasks, making it highly impactful. If work is done to go back and revaluate leading methods using our consistent alignment approach, many may see their results improve or worsen regardless of the novelty of their methodology. Aside from implications in academia, our work to reduce the performance gap when alignment backends are mismatched allows pre-trained models to more easily move from academic works into production applications. This is also true with regard to works moving between disciplines, such as applying models proposed in computer science literature to biological applications.

Our work on transfer learning suggests that recent works on PA estimation which do not include CA as a pre-training stage are leaving a significant amount of performance on the table, making it difficult to compare recent works using FR pre-training with older works which only use CA pre-training. It is possible that with the correct transfer learning formulation older works could be brought forward into the present day and shown to reach the level of current architectures, though without data this is just conjecture. Aside from the formulation we use to reach SOTA results, we also present a novel approach for repurposing FR features at high resolution. This discovery implies that the scale invariance of CNN features should be reconsidered, as in some cases training at low resolution may learn to extract features which are effective at higher resolution. This finding is powerful as training at the lowest resolution we use (112px) is at least an order of magnitude faster than if we were to train the entire CNN at the highest resolution at which we tested DFT (800px). To be specific, training at 800px requires over 44x more calculations simply stemming from the increase number of input features.

Finally, in our last methodological chapter we applied all of our learnings to a novel clinical dataset produced during a previous epidemiological study. It is in this chapter that we make our most impactful findings, laying the foundations for a wealth of future work integrating deep PA estimation with biological research. We review the implications of this work in detail in Section 6.3.3 and as such do not repeat our comments here.

6.4.4 Reflection

During this project we made many interesting discoveries and implemented several strong methodologies for preprocessing, feature extraction, modelling and evaluation. We did however also make some mistakes, including some research which we did not complete to a high enough standard to make it into this thesis. In this section we reflect on the research efforts which did not ultimately produce enough value to become part of this thesis.

As an extension of out work in on preprocessing we invested a significant amount of time developed a face segmentation system to remove background confounders from face images. Ultimately this system was used to remove noise in the 3D face images in Chapter 5, but we had hoped it would also show valuable results in other datasets. Unfortunately we did not reach a point in our experiments fit for publication, nor did we critically evaluate the performance of this approach in Chapter 5. For this reason we see our work on confounder removal with semantic segmentation a somewhat poorly organised research effort, which deserved more attention in the earlier stages of the project.

A large part of the original aim of the project was to collaborate with 3 key institutions carrying out biological research on perceived age, of which we only managed to work with one. In part, these collaborations well through due to the COVID-19 pandemic and its impact on the general appetite for travel. However we do note that our lack of readiness in the earlier years of the project may also have contributed. Instead of pulling the latest and greatest age estimation models from Github and applying them to our own data, we spent the earlier years of the project research more broadly and experimenting with a wide range of deep learning tools. While this did lead to a strong background understanding of the field, a more focussed effort to follow the SOTA would have benefitted our ability to collaborate without physical presence.

During some early experiments we focussed on the formulation of data augmentation for PA estimation, finding that to a degree increasing augmentation levels improved predictions. We did not however explore this in enough detail to produce publishable results, nor did we integrate many of these early findings into our final SSL work, where they may have been more valuable. It remains unclear what augmentation should be used to train the best CA and PA estimation models, a question that limits the consistency of the literature.

In summary, we see clear areas where our research focus could have been improved to better reflect the interdisciplinary nature of this project. We have identified COVID-19 as an additional challenge to our ability to collaborate as intended, but also take some responsibility for this via our technical readiness in the earlier years of the project. In more than one case we began efforts toward promising research questions but did not exhibit enough diligence in the process to reach publishable standards. Overall, this does not detract significantly from the value of our work as a whole, as we met and in some cases exceeded our aims and expectations.

Chapter 7

Future Work

During our work towards the general aims in this thesis, we uncover several areas of potential future work. This future work is in some cases a clear extension of our own work but in many cases it is a collection of research questions we deemed to be out of scope, deserving further attention. In a similar fashion to the discussion chapter we break down future work as it relates to our research objectives, with two additional sections covering work we deemed out of scope. The two additional sections, data collection and deployment, sit either side of our primary objectives in preprocessing, modelling and health; filling in the gaps needed to cover the entire estimation age prediction pipeline.

7.1 Data Collection

Given the scale and ethical complexity of PA data collection we deem it out of scope for this thesis, however it remains an extremely relevant component of the PA prediction pipeline. In addition to the foundational public dataset provided by Chalearn, we are privileged to have access to a number of other private PA datasets which in combination allow us to carry out impactful research without the addition of new data. Given all our findings, what remains unanswered is what impact additional perceived age data would have on the clinical readiness of deep learning PA models for health analysis. In Chapter 2 we state that the lack of annotated PA data creates a challenging research environment compared to other face analysis tasks such as detection, recognition and even CA prediction. These challenges have led researchers to focus on training and architectural methods which allow CNNs to extract robust features even in such a low data environment. One of the key limitations of current data is not only its scale but its heavy class imbalance.

For this reason we suggest future work collecting PA data focuses narrowly on sampling uniformly. This poses a significant challenge due to the natural imbalance present in the human population, as well as the bias associated with images available on the internet. Some multi-step process will be required where images are first collected using the current methodology, but with an additional filtering step. This filtering step may use manual human review or some existing

PA model with reasonable performance to undersample the data, balancing the distribution of classes. We also note that the Chalearn 2016 dataset contains many images of children under the age of 18, which may be useful for other applications but does not add anything to the field in which our study is focusses. Collecting a dataset with an equal number of individuals in each age bracket, ranging from roughly 30 to 100 years old, allows further research to deal with the yet known challenges associated with improving PA prediction, as opposed to methods for learning in the presence of heavy class imbalance. It is not clear what said additional work will uncover, but we can be certain that a uniform dataset will benefit the accuracy of prediction in the elderly.

On the need for more annotated PA data, we see an additional limitation of current datasets that if addressed may significantly accelerate the adoption of subsequent models in a health and wellbeing setting. This limitation is the non-commercial and private nature of current PA datasets, which does not limit the development of further research but instead limits the applicability of said research. Applicability is limited due to the capitalistic nature of our society, with technological developments in healthcare technology seldom reaching patients without some profit-making entity driving their adoption. While it is possible that a grant funded institution could develop a reliable PA model and release it open source, the barrier to adoption is still high. Hospitals, GP surgeries or care homes would need to purchase the correct hardware and subsequently install/-maintain the required software, which is challenging due to the current lack of expertise in these settings. It is far more likely that a private company would take their open source development and productise it, licensing the hardware and software as a bundle including maintenance. The benefits of centralising this responsibility extend further still, with opportunities to collect data with the consent of patients allowing the system to be self improving. All this said, nothing stops a private entity pursuing such an activity using a private self-collected dataset, which is also a likely outcome.

Given the clear need for more data and the cumbersome nature of PA annotation, we identify a research question which when answered may significantly reduce the cost of annotation. Human assessment of perceived age is presently done at a single sample level, the annotator is shown a single face and asked to guess their age in either a bracket or a single year. Having attempted this process themselves, the authors can attest to the challenging nature of this process, and the lack of certainty an annotator has when making a judgement: does this person look 40 or 45? We propose that the task be formulated differently, asking annotators to instead rank faces by who looks younger or older. Qualitatively we see this as a much more straightforward task, removing the need for the human brain to compare the current face to memorable faces, and instead allowing it to act as a simple comparative function. This approach may also reduce the known biases in human perception, as annotators are less likely to make guesses based on the population they are most exposed to. Two works support this theory, firstly [248] uses exemplar faces of known age as comparative signposts, asking annotators to assess if an unlabelled face looks younger or older than a given exemplar face. These older/younger labels are combined using the posterior distributions, representing each assessment as logistic curves and outputting a final value which is the summation of all curves for an individual. What we propose differs from this slightly as we suggest that no exemplar faces are needed, and that given a large enough number of comparisons an entirely unlabelled dataset could be ranked forming a regression value. This regression of faces would not directly represent perceived age, but could be split into n-tiles which capture the same information. We note that another work trains a deep learning model to regress face in a similar way to the previous human study mentioned, they use exemplar images internally within the model, training a CNN to make the older/younger judgement and outputting a final age based on the backet at which the binary value flips. Overall we see this approach to PA annotation as a mostly open question, with some existing work indicating that it may indeed improve the speed and quality of annotations.

Finally, we suggest that more research is done into the samping of both dataset subjects and annotators from a demographic perspective. While a small body of research does exist suggesting that annotator demographic is less important when groups of 10 or more annotators are used [108], we deem it worthy of further investigation nonetheless. Many of the potential biases found in human perception over decades of research have not been interrogated in a perceived age setting, nor have their interactions with the demographics of subjects. We do not know if individuals could receive training allowing them to become more effective PA annotators, nor do we know if they are focussing on facial features which are associated with health. A more in-depth review of the process could give new insights into what drives perceived age, supporting more technical work which attempts to draw the same conclusions from deep learning technology. Perhaps matching the demographics of annotators with that of the subjects may give more accurate estimations, or perhaps the opposite is true. From a more superficial perspective, annotator demographics may hold special significance when considered in a love and attraction setting. Many middle-aged to elderly individuals undergo procedures in an attempt to reduce the appearance of age on their face, projecting youthful looks which are often considered more attractive. What remains unknown is if the perception of youthful attractiveness is different depending on the target demographic, raising the question "in whose eyes do you want to look younger". While arguably less impactful than further work in health and genetics, this line of further work may hold significant value in dermatology and plastic surgery.

7.2 Preprocessing

Our work on preprocessing entailed a broad review of the relevant literature and culminated in a framework for the standardisation of face alignment implementations. Following the completion of our objectives we see several clear extensions of our work that if addressed, may further improve the absolute performance of face analysis systems. In addition to these clear extensions there are several facets of face image preprocessing we did not explore in detail, such that they may provide interesting future work.

We begin with a research objective we identified during the development of the system presented in Chapter 3 but chose not to investigate due to the extensive computational commitment required, as well as reaching far outside the scope of this thesis. Our framework for eye based alignment is parameterised using two continuous values which represent the scale (S) and vertical position (V) of the face within the frame. We select S and V parameters for our experiments naively using manual assessment, where we qualitatively evaluated a number of face images under a range of parameters, choosing the pair which placed all faces entirely within the frame and avoided most background content. Given that this aligns with the face alignment method seen in the literature we are reasonably confident it is near to the optimal value, but we cannot be certain without further investigation. For this reason we propose that a grid search of S and V parameters is carried out, where each pair is evaluated by training and testing a model on the resulting aligned images. We note that this grid search should include several different face analysis tasks, as it is unlikely that a single pair of parameters is optimal across the broad range of possible endpoints. Considerations could be made for tasks such as masked identity prediction, which became particularly relevant during the recent covid-19 pandemic. Masked faces contain little to no unique information in the masked region, such that optimal models would likely benefit from faces positioned with the eyes and nose filling most of the frame. Other tasks such as hair prediction may benefit from wider crops, as opposed to tasks such as emotion recognition which may operate more effectively with more narrow crops. It is also possible that as the resolution and feature recognition capabilities of CNNs improve this line of work may become less valuable, but while face images continue to exist are resolutions far higher than CNNs can process, we see significant value in such a study.

The basis of our work on face alignment stems from the fact that bounding box based face detectors do not have a high level of agreement. This has the effect of cropping faces from the image with different sized and shaped bounding boxes, leading to a misalignment between facial features when different algorithms are used. Under the reasonable assumption that these bounding box biases stem from differences in their respective training data, we propose future work to standardise the bounding box heuristics between different face detection datasets. Following this each face detector could be retrained using identical bounding boxes and our own research could be repeated. It is possible that with variance in bounding boxes removed, the value of eye based alignment becomes far less significant, which would detract from our findings but may also reduce the computational complexity of face analysis pipelines. If bounding boxes are shown to provide consistent face crops which are reliable enough to produce good results at both training and inference time, the need for facial landmarks is removed. Facial landmark extraction is not excessively computationally intensive but does require additional processing over face detection, such that its removal has a positive impact on both the hardware cost and speed of face analysis systems.

Regarding the speed and complexity of face alignment, we see another avenue which may lead to significant optimisation. Current keypoint detectors extract a number of points varying from 5 to 68 keypoints, with some less known approach extending even further. However, in our experiments we find that only two keypoints (the eyes) is enough to reliably and effectively align faces. We suggest that reducing the number of detected key points reduces the complexity of the detection task, making it possible to achieve similar results with simpler models and fewer parameters. A reduction in CNN parameters decreases the time needed at both training and inference time, streamlining face analysis implementations significantly. While the most robust keypoint detectors are successful in over 95% of images, some errors still occur in extremely challenging settings. We suggest that this lack of robustness may also be addressed by reducing the complexity of the task. We cannot be certain that such a task reformulation will benefit the accuracy of face analysis, but this may also be true, motivating further research into both the architecture of face alignment systems and their subsequent impact on performance of downstream models.

In preliminary experiments we show that image augmentation is not a good solution for inducing alignment invariance in age prediction models as it reduces their accuracy when alignment is implemented consistently. We do not however explore how novel network architectures could be combined with more advanced training formulations including augmentation to address this concern. It is our assumption that additional translational augmentation degrades model performance in the same way that highly controlled alignment improves performance, leveraging the fact that CNNs exploit boundary effects and learn to detect features at specific locations within the image frame [198]. This limitation has been shown in current CNN architectures but it remains unclear if new layer configurations could be used to allow CNNs to exploit features in a more translationally invariant way.

7.3 Modelling Perceived Age

Throughout the entirety of this project training deep CNNs for PA estimation has been the goal. Whether it is optimising preprocessing or evaluating their outputs, CNN models are at centre stage. We explore three high level formulations of PA training, including transfer learning, deep feature transfer and semi-supervised learning. Our work on transfer learning for perceived age estimation explores research gaps surrounding its application with modern deep learning technologies. We evaluate semi-supervised learning in a clinical setting, using it to annotate a new dataset of elderly faces to confirm the link between PA and the MC1R gene. Deep feature transfer is a promising approach to efficient face analysis which we show can be applied to perceived age prediction well in highly controlled settings. In this section we review the gaps we see remaining following our work on modelling as well as suggesting open questions that motivate further work.

The activity of modelling perceived age from faces is entirely open-ended with almost unlimited possible research avenues, even when the scope is limited to only deep learning based methods. Throughout our work we identify several areas of study closely aligned to our overarching research goal, but do not explore them thoroughly enough to become part of this thesis. One such obvious extension to our work is the analysis of the contribution of individual parts of the face to perceived age, which has already been studied to some extent by dermatologists [65, 177]. Individual facial ageing features are presently graded as standard using photometric scales, which use expert judgements to compare the presentation of features such as wrinkles to a number of exemplar images [135, 57]. We see a clear opportunity to replace this process with deep learning, but lack the data needed to approach the problem. Many age-related facial features are associated with morphological changes to the face's structure, such as loss of subcutaneous fat deposits and reduction in bone thickness due to osteoporosis [49]. These changes are visible in 2D images but may be more measurable using 3D imaging technology. Given we have access to 3D images with associated health and genetic information, this line of study would be possible with our immediate resources. However, due to the slow development of methods for 3D CNNs and a total lack of pre-training data, we do not see it as technically viable at this point. As 3D imaging becomes more prevalent and large scale datasets are formed, this research should be investigated thoroughly.

In our work we choose to use very small CNN backbones due to their strong performance and efficient training, allowing us to fully explore their potential in an academic setting. It is however evident that in most other deep learning tasks larger models produce better results [211]. In addition to this, it is well known that model size and dataset size are linked, with larger models requiring more data to perform well and vice versa [130, 190]. As it is the case that larger models require more data to learn robust representations, we suggest that we may have already reach model size limits in age estimation, and that without further data gathering it is unlikely that larger models will produce significantly better results. We propose that a study is design to explore this idea in more depth, finding the link between data size and model size in the context of age estimation. A reasonable amount of theory has been proposed regarding model capacity which could be incorporated into this further work, such as the ideas presented in [73]. As we discussed in Chapter 6, we do not explore the lower limit of model size for age estimation. This unknown could also be addressed as part of a study on model size.

Aside from exploring potential avenues to improve performance with CNN architecture, we suggest that algorithmic bias should be considered. We see no work in the existing literature indicating that CNNs for face analysis induce a methodological bias, instead researcher point to the training data as the key source of bias [131]. For example, in a hypothetical situation where preprocessing inadvertently reduces the brightness of an input image, darker faces would be subject to a greater level of information loss due to clipping (negative pixel values not being allowed). As stated previously, we do not see this in current methods and do not expect to find anything confirming its existence. However, it is naive to continue to operate under this assumption, such that further work should be done to measure the biasses associated with different architectures, loss functions and TL formulations.

In this thesis we stick rigidly to the DLDLv2 loss function, which is a combination of distribution learning and l1 regression. During the project however we do experiment with other formulations including SPUDRF and CORAL, both of which are very promising approaches for age estimation. Neither approach claims to outperform DLDL but we see features from each which could be combined to produce an overall better model. CORALs unique class weighting scheme for prediction consistency could be re-design to operate on a distribution age encoding, providing a prior for the expected response from each output neuron. The self-pacing method from SPUDRF may not only provide benefit to DRF based networks, but could indeed be applied to any architecture. We would like to see research done to break down the individual contribution of each of these components, with a fair evaluation on numerous tasks and datasets. More recent models presented since the beginning of this project are also of interest, namely MWR [203] and FP-Age [154]. The comparative search process presented in MWR intuitively seems like the correct approach to age estimation given its links to human cognition, with their results backing this up. FP-Age argues that face parsing attention helps age prediction by teaching the model relevant image regions. As this techniques optimise very different parts of the age estimation process, they are compatible with each other and may produce further improvements.

Recent advances in face analysis may also benefit our DFT approach by providing more robust feature extractors than FR. Two recent approaches train strong general face embedding models using very different methods: FaRL [251] and SwinFace [186]. FaRL uses noisy text image pairs to train a general face model using a large dataset, SwinFace uses a wide range of face analysis tasks and loss functions to train a multi-headed multi-task face analysis model. It is very likely that both these models share the same properties as FR models with regard to the generality of their feature spaces. Our experiments could be re-run using these backbones as feature extractors in place of FR. Furthermore, it is also possible these backbones may be more suitable initialisations for our TL experiments. If these models do work as well as we predict, it is likely that this will become the default approach for all face analysis tasks involving CNNs, making it an exciting research direction.

In Chapter 5 we showed that balanced mini-batch sampling improves both standard TL and SSL. We do not however compare this approach to simply oversampling the entire dataset, which may provide similar benefits with a slightly greater computational cost (oversampling results in more training steps). We also only balance mini-batches by their CA or PA, which may be limiting from a diversity perspective. It may also be beneficial to uniformly sample data based on other facial metadata such as identity, gender, ethnicity and facial expression. Such an approach would require secondary models to annotate those facial features, from which a new sampling algorithm could be design to balance batches in that dataset. This research should also be paired with a fair evaluation using a dataset such as FairFace, where biases toward particular demographics are easier to find.

7.4 Health and Genetics

In Chapter 5 we review our application of DL predicted PA in a cross-sectional study involving morbidity and genetics, confirming the known links between human PA and these endpoints. The next steps in this research are perhaps the most exciting and well motivated future work we propose, as there are so many unknowns in this burgeoning field. We see clear further opportunities to associate deep learning PA with endpoints already known to link with human PA, such as the findings from the Danish morbidity study [110]. Using exactly the same methodology proposed in Chapter 5 it would be possible to extract DL PA from the face images previously shown to human annotators, and make the same comparison between twin pairs. If the same association between PA and morbidity is found, the impact of PA in healthcare would be significantly enlarged, as it would show an ability to predict the onset of mortality. We also note that current studies such as the ERGO cohort study continue to follow individuals who have existing photos with human PA annotation. As these individuals progress through their old age and ultimately reach death, mortality labels could be added to the same dataset in which we have already shown links with morbidity.

This presents a key opportunity to repeat the findings from [110]. One facet of perceived age that has been ignored due to the costly nature of annotation is the tracking of PA longitudinally. Taking multiple PA assessments at intervals through an individual's life allows a second order metric to be calculated measuring the rate of change in PA. This metric may show even stronger links with morbidity and mortality as it is one step closer to the underlying mechanism of ageing. Furthermore, the rate of ageing may not be constant throughout an individual's life, and may even change significantly with the modifiable lifestyle factors. This allows for personalised and actionable suggestions to be made by healthcare professionals to assist in reducing disease and increasing longevity. Deep learning may already be up to the task of predicting PA accurately enough that progression metrics can be calculated when images are taken in a lab setting, however it is unclear if the same is true for in-the-wild images. If its efficacy is however proven in unconstrained images, smartphone applications could be used to allow for highly granular tracking of PA progression, requiring only a selfie photo taken roughly each day. The possibilities for this type of app to influence lifestyle and dietary choices are interesting and should be considered as a possible next step.

While PA prediction from the face has proven to be valuable and provide useful insights about health and genetics, there is a clear next step needed to to begin making deep conclusions about the mechanisms of ageing. Granular or localised PA annotations would allow ageing to be attributed to specific face features, many of which have already been associated with particular lifestyle factors such as sun exposure and smoking. Annotating the perceived age of features such as the nose and mouth is challenging for humans as we naturally have very little experience with the task. Not only are we seldom required to guess the age of a single feature, but when we do assess the age of a face we are presented with a very heterogeneous appearance. Given two individuals of the same gender and perceived age, one may have younger eyes while also having an older looking mouth. It is precisely due to the heterogeneity in face ageing that a photometric grading scale for granular features was introduced. Ordinal scales for localised ageing allow healthcare professionals to rate the severity of features such as wrinkles and skin sagging by minimising the potential bias introduced by a single observer. We would like to see further experiments done to annotate datasets with both PA annotations and granular skin ageing from photometric scales, which can then be used to train deep learning models which attributes ageing to specific features. Previous work [65] has shown that human PA can be predicted by a linear combination of skin parameters, but such a model requires that these parameters are manually annotated for all future experiments. We suggest the involvement of deep learning to extract such features with only the need for a face image, significantly reducing the cost of annotation and allowing for larger studies to be done. If such a model can predict the results of photometric analysis at a similar level of agreeability to dermatologists, smartphone based skin age screening becomes viable.

Novel CNN architectures could be proposed to accommodate learning PA via the proxy of various objective skin ageing parameters. One simple approach is to extend the fully connected portion of the CNN with additional output where a multi-task loss is calculated. This would force the model to learn representations which capture information relevant to both PA and the secondary tasks, which may even improve the accuracy of PA prediction as has been shown with gender [5].

Given the diversity in skin ageing, it is possible some skin parameters require very different convolutional filters due to differences in their spatial nature. For example skin sagging primarily affects the shape of the face region, where sun damage appears mostly as a texture component. To promote learning in a multi-task model, one CNN backbone could be used for each skin parameter, with some degree of weight sharing at deep layers allowing for more effective feature fusion [102] for the final prediction of PA. At present it is entirely unclear how such research would develop, and perhaps PA contains some information which does not exist in any independent skin parameter.

While multi-task learning may be useful for both understanding skin ageing and optimising skin evaluation, another approach could be used which requires little to no additional annotation. Interpretability is a domain of deep learning pertaining to the attribution of predicted value to input features, which in our case is attributing PA predictions to pixels in the image space. In preliminary experiments we show that several interpretability techniques can be used to measure the impact of image features on PA predictions, including the use of semantic segmentation maps to build attribution maps linking to facial features. The results of our preliminaries were inconclusive for two reasons, firstly we see high degrees of variance in attribution maps between samples, making it hard to draw significant conclusions. Secondly, the metadata required to properly interrogate attribution maps is not available in public datasets, and while it is available in the private datasets we work with, we found ourselves limited by time and resources. For this reason we see a clear opportunity to optimise and extract interpretability features from deep PA models using clinical datasets, and look for associations between facial features and underlying health and genetic endpoints. Perhaps when PA predictions are driven by pixels around the jawline, skin sagging is the key feature, which may be associated with osteoporosis. Until time is dedicated to this line of work in a clinical setting we are not able to answer such questions, which should clearly motivate future work.

7.5 Deployment

During the nascent stages of any technological development, more energy is focussed on its discovery and optimisation than on its application in the real world. Regarding DL based PA prediction, we are reaching a stage where application constraints must begin to be considered. For the sake of clarity, we would continue to prioritise all of the other further work defined in this chapter over the work in this section, but see it fit to include regardless. The primary considerations needed for an effective deployment of deep learning in a healthcare setting are trust, impact and cost. Trust represents many of the measurable features of deep learning models such as their accuracy and reliability, specificity and sensitivity being the most commonly used metrics. Trust also encompasses the more subjective nature of the interface between healthcare professionals and technology, which can limit even the most accurate and reliable technologies. For example, if a deep learning based PA model is deployed into a primary care setting as a screen tool for elderly patients, how are the results presented to their physician? Perhaps in this case the results are not useful for individual patients, but instead could be used over longer time periods to allocate resources to localities where the population is on average at higher risk for specific morbidities. Given the interdisciplinary nature of this line of thinking, the inclusion of doctors and health researchers is crucial to finding the best way to interface deep PA prediction with the healthcare system.

A second barrier to the adoption of deep learning that must be addressed is public perception. While great strides have been made in the acceptance of medical technology, there is a far smaller precedent for technologies perceived as artificial intelligence by the general public. Many who do not have an awareness of the widespread adoption of deep learning may be sceptical that their facial appearance reflects their systematic health, and may distrust doctors who promote its use. We are not able to say what the public acceptance of deep PA prediction will be, and suggest that further work in surveying public opinion should be done before making decisions about its use in a clinical setting.

Finally, we suggest that a qualitative review of the ethical implications of such DL PA models be done under the assumption that early stage models will contain unknown biases. Given the black box nature of CNNs and the impressive performance they are beginning to show in many domains, it is possible that they will reach levels of accuracy and usefulness which warrant deployment before they are fully understood. Far too often are machine learning tools deployed without a proper evaluation of their potential biases, only to cause a negative impact in the real world before their ethical implications are fully understood. In this thesis we focus on the early stage developments needed to bring PA prediction technology closer to applicability, but call for other researchers who operate closer to a clinical setting to begin considering how such a technology could shape future healthcare; for better or for worse.

Bibliography

[1] en-US. URL: https://aws.amazon.com/machine-learning/inferentia/.

[2] en. URL: https://www.jetbrains.com/pycharm/.

[3] URL: https://code.visualstudio.com/.

[4] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[5] Amirali Abdolrashidi et al. "Age and Gender Prediction From Face Images Using Attentional Convolutional Network". In: arXiv:2010.03791 (2020). arXiv:2010.03791 [cs, eess]. DOI: 10.48550/arXiv.2010.03791. URL: http://arxiv.org/abs/2010.03791.

[6] Fatma S. Abousaleh et al. "A novel comparative deep learning framework for facial age estimation". en. In: *EURASIP Journal on Image and Video Processing* 2016.1 (2016), p. 47. ISSN: 1687-5281. DOI: 10.1186/s13640-016-0151-4.

[7] O. Agbo-Ajala and S. Viriri. "A Lightweight Convolutional Neural Network for Real and Apparent Age Estimation in Unconstrained Face Images". In: *IEEE Access* 8 (2020), pp. 162800–162808. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3022039.

[8] Olatunbosun Agbo-Ajala and Serestina Viriri. "Deep learning approach for facial age classification: a survey of the state-of-the-art". en. In: *Artificial Intelligence Review* 54.1 (2021), pp. 179–213. ISSN: 1573-7462. DOI: 10.1007/s10462-020-09855-0.

[9] Olatunbosun Agbo-Ajala et al. "Apparent age prediction from faces: A survey of modern approaches". In: *Frontiers in Big Data* 5 (2022). ISSN: 2624-909X. URL: https://www.frontiersin.org/articles/10.3389/fdata.2022.1025806.

[10] Eirikur Agustsson et al. "Apparent and Real Age Estimation in Still Images with Deep Residual Regressors on Appa-Real Database". In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 87–94. ISBN: 978-1-5090-4023-0. DOI: 10.1109/FG.2017.20. URL: http://ieeexplore.ieee.org/document/7961727/.

[11] Shijie Ai, Lili Pan, and Yazhou Ren. "Self-Paced Deep Regression Forests for Facial Age Estimation". In: arXiv:1910.03244 (2020). arXiv:1910.03244 [cs]. DOI: 10.48550/arXiv.1910.03244. URL: http://arxiv.org/abs/1910.03244.

[12] Ali Akbari et al. "A Flatter Loss for Bias Mitigation in Cross-dataset Facial Age Estimation". In: arXiv:2010.10368 (2020). arXiv:2010.10368 [cs]. URL: http://arxiv.org/abs/2010.10368.

[13] Mohammed Jawad Al-Dujaili and Hydr jabar sabat Ahily. "A New Hybrid Model to Predict Human Age Estimation from Face Images Based on Supervised Machine Learning Algorithms". en. In: *Cybernetics and Information Technologies* 23.2 (2023), pp. 20–33. DOI: 10.2478/cait-2023-0011.

[14] Alhanoof Althnian et al. "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain". en. In: *Applied Sciences* 11.2 (2021), p. 796. ISSN: 2076-3417. DOI: 10.3390/app11020796.

[15] Felix Anda et al. "Assessing the Influencing Factors on the Accuracy of Underage Facial Age Estimation". en. In: *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)* (2020). arXiv: 2012.01179, pp. 1–8. DOI: 10.1109/CyberSecurity49315.2020.9138851.

[16] Grigory Antipov et al. "Apparent Age Estimation From Face Images Combining General and Children-Specialized Deep Learning Models". In: 2016, pp. 96–104. URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016_workshops/w18/html/Antipov_Apparent_Age_Estimation_CVPR_2016_paper.html.

[17] Gizelle Anzures et al. "AN OWN-AGE BIAS IN YOUNG ADULTS' FACIAL AGE JUDGMENTS". In: *PSYCHOLOGIA* 54.3 (2011), pp. 166–174. ISSN: 0033-2852, 1347-5916. DOI: 10.2117/psysoc.2011.166.

[18] Gizelle Anzures et al. "Culture Shapes Efficiency of Facial Age Judgments". In: *PLOS ONE* 5.7 (2010), e11679. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0011679.

[19] J. Ayer et al. "A photonumeric scale for the assessment of atrophic facial photodamage". In: *British Journal of Dermatology* 178.5 (2018), pp. 1190–1195. ISSN: 1365-2133. DOI: 10.1111/bjd.16331.

[20] D.L. Baggio. *Mastering OpenCV with Practical Computer Vision Projects*. Community Experience Distilled. Packt Publishing, 2012. ISBN: 978-1-84951-783-6. URL: https://books.google.co.uk/books?id=UjWoIFHcr58C.

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: arXiv:1409.0473 (2016). arXiv:1409.0473 [cs, stat]. DOI: 10.48550/arXiv.1409.0473. URL: http://arxiv.org/abs/1409.0473.

[22] Alexandre Bailly et al. "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models". In: *Computer Methods and Programs in Biomedicine* 213 (2022), p. 106504. ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2021.106504.

[23] Valentin Bazarevsky et al. "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs". In: arXiv:1907.05047 (2019). arXiv:1907.05047 [cs]. DOI: 10.48550/arXiv.1907.05047. URL: http://arxiv.org/abs/1907.05047.

[24] SE. Bekhouche et al. "A comparative study of human facial age estimation: handcrafted features vs. deep features". en. In: *Multimedia Tools and Applications* 79.35 (2020), pp. 26605–26622. ISSN: 1573-7721. DOI: 10.1007/s11042-020-09278-7.

[25] F. Bougourzi, F. Dornaika, and A. Taleb-Ahmed. "Deep learning based face beauty prediction via dynamic robust losses and ensemble regression". en. In: *Knowledge-Based Systems* 242 (2022), p. 108246. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2022.108246.

[26] François Bourlière. "The assessment of biological age in man". In: *The assessment of biological age in man* 37 (1970). URL: http://bases.bireme.br/cgi-bin/wxislind.exe/iah/online/?IsisScript=iah/iah.xis\&src=google\&base=PAHO\&lang=p\&nextAction=lnk\&exprSearch=42285\&indexSearch=ID.

[27] Stevo Bozinovski. "Reminder of the First Paper on Transfer Learning in Neural Networks, 1976". en. In: *Informatica* 44.33 (2020). ISSN: 1854-3871. DOI: 10.31449/inf.v44i3.2828. URL: https://www.informatica.si/index.php/informatica/article/view/2828.

[28] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[29] Marybeth Brown et al. "Physical and Performance Measures for the Identification of Mild to Moderate Frailty". In: *The Journals of Gerontology: Series A* 55.6 (2000), pp. M350–M355. ISSN: 1079-5006. DOI: 10.1093/gerona/55.6.M350.

[30] Adrian Bulat and Georgios Tzimiropoulos. "How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)". In: *International Conference on Computer Vision*. 2017.

[31] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. "Rank-consistent Ordinal Regression for Neural Networks". In: *arXiv:1901.07884 [cs, stat]* (2019). arXiv: 1901.07884. URL: http://arxiv.org/abs/1901.07884.

[32] Vincenzo Carletti et al. "Age from Faces in the Deep Learning Revolution". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (2020), pp. 2113–2132. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2019.2910522.

[33] Alastair Carruthers et al. "A Validated Grading Scale for Crow's Feet". In: *Dermatologic Surgery* 34 (2008), S173–S178. ISSN: 10760512, 15244725. DOI: 10.1111/j.1524-4725.2008.34367.x.

[34] Jean Carruthers et al. “Validated Assessment Scales for the Mid Face:” in: *Dermatologic Surgery* 38.2ptII (2012), pp. 320–332. ISSN: 1076-0512. DOI: 10.1111/j.1524-4725.2011.02251.x.

[35] Ken Chatfield et al. “Return of the Devil in the Details: Delving Deep into Convolutional Nets”. In: arXiv:1405.3531 (2014). arXiv:1405.3531 [cs]. URL: http://arxiv.org/abs/1405.3531.

[36] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. “Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval”. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Cham: Springer International Publishing, 2014, pp. 768–783. ISBN: 978-3-319-10599-4.

[37] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. “Cross-Age Reference Coding for Age-Invariant Face Recognition and Retrieval”. en. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Vol. 8694. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 768–783. ISBN: 978-3-319-10598-7. DOI: 10.1007/978-3-319-10599-4_49. URL: http://link.springer.com/10.1007/978-3-319-10599-4_49.

[38] Fangmei Chen, Xihua Xiao, and David Zhang. “Data-Driven Facial Beauty Analysis: Prediction, Retrieval and Manipulation”. en. In: *IEEE Transactions on Affective Computing* 9.2 (2018), pp. 205–216. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2016.2599534.

[39] Ke Chen et al. “Cumulative Attribute Space for Age and Crowd Density Estimation”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 2467–2474. DOI: 10.1109/CVPR.2013.319.

[40] Ying Chen and John Lyga. “Brain-Skin Connection: Stress, Inflammation and Skin Aging”. In: *Inflammation & Allergy Drug Targets* 13.3 (2014), pp. 177–190. ISSN: 1871-5281. DOI: 10.2174/1871528113666140522104422.

[41] Jingchun Cheng et al. “Exploiting effective facial patches for robust gender recognition”. In: *Tsinghua Science and Technology* 24.3 (2019), pp. 333–345.

[42] Sharan Chetlur et al. “cuDNN: Efficient Primitives for Deep Learning”. In: arXiv:1410.0759 (2014). arXiv:1410.0759 [cs]. DOI: 10.48550/arXiv.1410.0759. URL: http://arxiv.org/abs/1410.0759.

[43] Sung Eun Choi et al. “Age estimation using a hierarchical classifier based on global and local facial features”. In: *Pattern Recognition* 44.6 (2011), pp. 1262–1281. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2010.12.005.

[44] Francois Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[45] K. Christensen et al. “Perceived age as clinically useful biomarker of ageing: cohort study”. en. In: *BMJ* 339.dec11 2 (2009), b5262–b5262. ISSN: 0959-8138, 1468-5833. DOI: 10.1136/bmj.b5262.

[46] Jonathan Hero Chung et al. "Cutaneous photodamage in Koreans: influence of sex, sun exposure, smoking, and skin color." In: *Archives of dermatology* (2001).

[47] Albert Clapés et al. "From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2373–2382.

[48] Victor Gabriel CLATICI et al. "Perceived Age and Life Style. The Specific Contributions of Seven Factors Involved in Health and Beauty". In: *Mædica* 12.3 (2017), pp. 191–201. ISSN: 1841-9038.

[49] Sydney R. Coleman and Rajiv Grover. "The Anatomy of the Aging Face: Volume Loss and Changes in 3-Dimensional Topography". In: *Aesthetic Surgery Journal* 26.1$_{\mathtt{S}}$`upplement` (2006), S4–S9. ISSN: 1090-820X. DOI: `10.1016/j.asj.2005.09.012`.

[50] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam, 2018. URL: `http://www.blender.org`.

[51] T. F. Cootes et al. "Active Shape Models-Their Training and Application". en. In: *Computer Vision and Image Understanding* 61.1 (1995), pp. 38–59. ISSN: 1077-3142. DOI: `10.1006/cviu.1995.1004`.

[52] T.F. Cootes, G.J. Edwards, and C.J. Taylor. "Active appearance models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.6 (2001), pp. 681–685. ISSN: 1939-3539. DOI: `10.1109/34.927467`.

[53] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. "Active Appearance Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Springer, 1998, pp. 484–498.

[54] Ekin D. Cubuk et al. "AutoAugment: Learning Augmentation Strategies From Data". en. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, 2019, pp. 113–123. ISBN: 978-1-72813-293-8. DOI: `10.1109/CVPR.2019.00020`. URL: `https://ieeexplore.ieee.org/document/8953317/`.

[55] Ekin Dogus Cubuk et al. "RandAugment: Practical Automated Data Augmentation with a Reduced Search Space". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 18613–18624. URL: `https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html`.

[56] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* 1 (2005), pp. 886–893. URL: `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467360`.

[57] Doris J Day et al. "The Wrinkle Severity Rating Scale: A Validation Study". In: *American Journal of Clinical Dermatology* 5.1 (2004), pp. 49–52. ISSN: 1175-0561. DOI: `10.2165/00128071-200405010-00007`.

[58] Hedwige Dehon and Serge Brédart. "An 'Other-Race' Effect in Age Estimation from Faces". In: *Perception* 30.9 (2001), pp. 1107–1113. ISSN: 0301-0066. DOI: 10.1068/p3122.

[59] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[60] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2022). arXiv:1801.07698 [cs], pp. 5962–5979. ISSN: 0162-8828, 2160-9292, 1939-3539. DOI: 10.1109/TPAMI.2021.3087709.

[61] Jiankang Deng et al. "RetinaFace: Single-stage Dense Face Localisation in the Wild". In: arXiv:1905.00641 (2019). arXiv:1905.00641 [cs]. DOI: 10.48550/arXiv.1905.00641. URL: http://arxiv.org/abs/1905.00641.

[62] Jiankang Deng et al. "Sub-center ArcFace: Boosting Face Recognition by Large-scale Noisy Web Faces". In: *Proceedings of the IEEE Conference on European Conference on Computer Vision*. 2020.

[63] Zongyong Deng et al. "PML: Progressive Margin Loss for Long-tailed Age Classification". In: arXiv:2103.02140 (2021). arXiv:2103.02140 [cs]. DOI: 10.48550/arXiv.2103.02140. URL: http://arxiv.org/abs/2103.02140.

[64] Terrance DeVries and Graham W. Taylor. "Improved Regularization of Convolutional Neural Networks with Cutout". In: arXiv:1708.04552 (2017). arXiv:1708.04552 [cs]. DOI: 10.48550/arXiv.1708.04552. URL: http://arxiv.org/abs/1708.04552.

[65] Denise Dicanio et al. "Calculation of apparent age by linear combination of facial skin parameters: a predictive tool to evaluate the efficacy of cosmetic treatments and to assess the predisposition to accelerated aging". In: *Biogerontology* 10.6 (2009), p. 757. ISSN: 1573-6768. DOI: 10.1007/s10522-009-9222-6.

[66] Lucas W. M. Diebel and Kenneth Rockwood. "Determination of Biological Age: Geriatric Assessment vs Biological Biomarkers". In: *Current Oncology Reports* 23.9 (2021), p. 104. ISSN: 1523-3790. DOI: 10.1007/s11912-021-01097-9.

[67] F. Dornaika, SE. Bekhouche, and I. Arganda-Carreras. "Robust regression with deep CNNs for facial age estimation: An empirical study". en. In: *Expert Systems with Applications* 141 (2020), p. 112942. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.112942.

[68] F. Dornaika et al. "Efficient deep discriminant embedding: Application to face beauty prediction and classification". en. In: *Engineering Applications of Artificial Intelligence* 95 (2020), p. 103831. ISSN: 0952-1976. DOI: 10.1016/j.engappai.2020.103831.

[69] Fadi Dornaika and Abdelmalik Moujahid. "Multi-View Graph Fusion for Semi-Supervised Learning: Application to Image-Based Face Beauty Prediction". en. In: *Algorithms* 15.66 (2022), p. 207. ISSN: 1999-4893. DOI: 10.3390/a15060207.

[70] Fadi Dornaika et al. “Toward graph-based semi-supervised face beauty prediction”. en. In: *Expert Systems with Applications* 142 (2020), p. 112990. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2019.112990.

[71] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: arXiv:2010.11929 (2021). arXiv:2010.11929 [cs]. DOI: 10.48550/arXiv.2010.11929. URL: http://arxiv.org/abs/2010.11929.

[72] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: arXiv:2010.11929 (2021). arXiv:2010.11929 [cs]. DOI: 10.48550/arXiv.2010.11929. URL: http://arxiv.org/abs/2010.11929.

[73] Simon S Du et al. “How Many Samples are Needed to Estimate a Convolutional Neural Network?” In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/hash/03c6b06952c750899bb03d998e631860-Abstract.html.

[74] Ionut Cosmin Duta et al. “Improved Residual Networks for Image and Video Recognition”. In: arXiv:2004.04989 (2020). arXiv:2004.04989 [cs]. URL: http://arxiv.org/abs/2004.04989.

[75] Matthew Earl. *Switching Eds: Face swapping with Python, dlib, and OpenCV*. 2015. URL: http://matthewearl.github.io/2015/07/28/switching-eds-with-python/ (visited on 04/27/2023).

[76] Natalie C. Ebner, Michaela Riediger, and Ulman Lindenberger. “FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation”. In: *Behavior Research Methods* 42.1 (2010), pp. 351–362. ISSN: 1554-351X, 1554-3528. DOI: 10.3758/BRM.42.1.351.

[77] Eran Eidinger, Roee Enbar, and Tal Hassner. “Age and Gender Estimation of Unfiltered Faces”. In: *IEEE Transactions on Information Forensics and Security* 9.12 (2014), pp. 2170–2179. ISSN: 1556-6013, 1556-6021. DOI: 10.1109/TIFS.2014.2359646.

[78] Maxwell L. Elliott et al. “Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort”. en. In: *Molecular Psychiatry* 26.88 (2021), pp. 3829–3838. ISSN: 1476-5578. DOI: 10.1038/s41380-019-0626-7.

[79] Sergio Escalera et al. “ChaLearn Looking at People 2015: Apparent Age and Cultural Event Recognition Datasets and Results”. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015, pp. 243–251. DOI: 10.1109/ICCVW.2015.40.

[80] William Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. Mar. 2019. DOI: 10.5281/zenodo.3828935. URL: https://github.com/Lightning-AI/lightning.

[81] Yang-Yu Fan et al. “Label Distribution-Based Facial Attractiveness Computation by Deep Residual Learning”. In: *IEEE Transactions on Multimedia* 20.8 (2018), pp. 2196–2208. ISSN: 1941-0077. DOI: 10.1109/TMM.2017.2780762.

[82] Songhe Feng et al. "Human Facial Age Estimation by Cost-Sensitive Label Ranking and Trace Norm Regularization". In: *IEEE Transactions on Multimedia* 19.1 (2017), pp. 136–148. ISSN: 1941-0077. DOI: 10.1109/TMM.2016.2608786.

[83] Bernhard Fink, Karl Grammer, and Paul J. Matts. "Visible skin color distribution plays a role in the perception of age, attractiveness, and health in female faces". In: *Evolution and Human Behavior* 27.6 (2006), pp. 433–442. ISSN: 1090-5138. DOI: 10.1016/j.evolhumbehav.2006.08.007.

[84] Frederic Flament et al. "Objective and automatic grading system of facial signs from selfie pictures of South African women: Characterization of changes with age and sun-exposures". en. In: *Skin Research and Technology* 28.4 (2022), pp. 596–603. ISSN: 1600-0846. DOI: 10.1111/srt.13153.

[85] Yun Fu, Guodong Guo, and Thomas S. Huang. "Age Synthesis and Estimation via Faces: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.11 (2010), pp. 1955–1976. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2010.36.

[86] Kunihiko Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". en. In: *Biological Cybernetics* 36.4 (1980), pp. 193–202. ISSN: 0340-1200, 1432-0770. DOI: 10.1007/BF00344251.

[87] Toshiyuki Furukawa et al. "Assessment of Biological Age by Multiple Regression Analysis". In: *Journal of Gerontology* 30.4 (1975), pp. 422–434. ISSN: 0022-1422. DOI: 10.1093/geronj/30.4.422.

[88] Junying Gan et al. "Deep self-taught learning for facial beauty prediction". en. In: *Neurocomputing* 144 (2014), pp. 295–303. ISSN: 09252312. DOI: 10.1016/j.neucom.2014.05.028.

[89] Bin-Bin Gao et al. "Age Estimation Using Expectation of Label Distribution Learning". en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 712–718. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/99. URL: https://www.ijcai.org/proceedings/2018/99.

[90] Bin-Bin Gao et al. "Deep Label Distribution Learning with Label Ambiguity". In: *IEEE Transactions on Image Processing* 26.6 (2017). arXiv: 1611.01731, pp. 2825–2838. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2017.2689998.

[91] Bin-Bin Gao et al. "Learning Expectation of Label Distribution for Facial Age and Attractiveness Estimation". en. In: *arXiv:2007.01771 [cs]* (2020). arXiv: 2007.01771. URL: http://arxiv.org/abs/2007.01771.

[92] Bin-Bin Gao et al. "Learning Expectation of Label Distribution for Facial Age and Attractiveness Estimation". In: arXiv:2007.01771 (2021). arXiv:2007.01771 [cs]. URL: http://arxiv.org/abs/2007.01771.

[93] Xin Geng, Chao Yin, and Zhi-Hua Zhou. "Facial Age Estimation by Learning from Label Distributions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.10 (2013), pp. 2401–2412. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2013.51.

[94] Xin Geng et al. "Learning from facial aging patterns for automatic age estimation". en. In: *Proceedings of the 14th ACM international conference on Multimedia*. Santa Barbara CA USA: ACM, 2006, pp. 307–316. ISBN: 978-1-59593-447-5. DOI: 10.1145/1180639.1180711. URL: https://dl.acm.org/doi/10.1145/1180639.1180711.

[95] Patricia A George and Graham J Hole. "Factors Influencing the Accuracy of Age Estimates of Unfamiliar Faces". In: *Perception* 24.9 (1995), pp. 1059–1073. ISSN: 0301-0066, 1468-4233. DOI: 10.1068/p241059.

[96] Ross Girshick et al. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.

[97] K. Godfrey. "Identification of parametric models from experimental data [Book Review]". In: *IEEE Transactions on Automatic Control* 44.12 (1999), pp. 2321–2322. ISSN: 1558-2523. DOI: 10.1109/TAC.1999.811220.

[98] William B. Goggins et al. "Frailty Index as a Measure of Biological Age in a Chinese Population". In: *The Journals of Gerontology: Series A* 60.8 (2005), pp. 1046–1051. ISSN: 1079-5006. DOI: 10.1093/gerona/60.8.1046.

[99] Yves Grandvalet and Yoshua Bengio. "Semi-supervised Learning by Entropy Minimization". In: *Advances in Neural Information Processing Systems*. Vol. 17. MIT Press, 2004. URL: https://proceedings.neurips.cc/paper_files/paper/2004/hash/96f2b50b5d3613adf9c27049b2a Abstract.html.

[100] C. E. Griffiths et al. "A photonumeric scale for the assessment of cutaneous photodamage". In: *Archives of Dermatology* 128.3 (1992), pp. 347–351. ISSN: 0003-987X.

[101] D. E. Grobbee et al. "[Coronary heart disease in the elderly; the ERGO study (Erasmus Rotterdam Health and the Elderly)]". dut. In: *Nederlands Tijdschrift Voor Geneeskunde* 139.39 (Sept. 1995), pp. 1978–1982. ISSN: 0028-2162.

[102] O. Guehairia et al. "Feature fusion via Deep Random Forest for facial age estimation". en. In: *Neural Networks* 130 (2020), pp. 238–252. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2020.07.006.

[103] Christiane Guinot et al. "Relative Contribution of Intrinsic vs Extrinsic Factors to Skin Aging as Determined by a Validated Skin Age Score". en. In: *ARCH DERMATOL* 138 (2002), p. 7.

[104] Christiane Guinot et al. "Relative Contribution of Intrinsic vs Extrinsic Factors to Skin Aging as Determined by a Validated Skin Age Score". In: *Archives of Dermatology* 138.11 (2002). ISSN: 0003-987X. DOI: 10.1001/archderm.138.11.1454. URL: http://archderm.jamanetwork.com/article.aspx?doi=10.1001/archderm.138.11.1454.

[105] Asuman Gunay and Vasif V. Nabiyev. "Automatic age classification with LBP". In: *2008 23rd International Symposium on Computer and Information Sciences*. 2008, pp. 1–4. DOI: 10.1109/ISCIS.2008.4717926.

[106] D.A. Gunn et al. "Lifestyle and youthful looks". In: *British Journal of Dermatology* 172.5 (2015), pp. 1338–1345. ISSN: 0007-0963. DOI: 10.1111/bjd.13646.

[107] David A. Gunn et al. "Facial Appearance Reflects Human Familial Longevity and Cardiovascular Disease Risk in Healthy Individuals". en. In: *The Journals of Gerontology: Series A* 68.2 (2013), pp. 145–152. ISSN: 1758-535X, 1079-5006. DOI: 10.1093/gerona/gls154.

[108] David A. Gunn et al. "Perceived age as a biomarker of ageing: a clinical methodology". In: *Biogerontology; Dordrecht* 9.5 (2008), pp. 357–64. ISSN: 1389-5729. DOI: http://dx.doi.org.libproxy.ncl.ac.uk/10.1007/s10522-008-9141-y.

[109] David A. Gunn et al. "Why Some Women Look Young for Their Age". en. In: *PLoS ONE* 4.12 (2009). Ed. by Tom Tregenza, e8021. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0008021.

[110] David Andrew Gunn et al. "Mortality is Written on the Face". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 71.1 (2016), pp. 72–77. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glv090.

[111] Guodong Guo et al. "Human age estimation using bio-inspired features". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 112–119. DOI: 10.1109/CVPR.2009.5206681.

[112] Guodong Guo et al. "Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression". In: *IEEE Transactions on Image Processing* 17.7 (2008), pp. 1178–1188. ISSN: 1941-0042. DOI: 10.1109/TIP.2008.924280.

[113] Yandong Guo et al. "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition". In: arXiv:1607.08221 (2016). arXiv:1607.08221 [cs]. DOI: 10.48550/arXiv.1607.08221. URL: http://arxiv.org/abs/1607.08221.

[114] Danielle Gutman et al. "Exceptionally Long-Lived Individuals (ELLI) Demonstrate Slower Aging Rate Calculated by DNA Methylation Clocks as Possible Modulators for Healthy Longevity". In: *International Journal of Molecular Sciences* 21.22 (2020), p. 615. DOI: 10.3390/ijms21020615.

[115] Muhammad Sabirin Hadis et al. "The Impact of Preprocessing on Face Recognition using Pseudorandom Pixel Placement". In: *2022 29th International Conference on Systems, Signals and Image Processing (IWSSIP)*. Vol. CFP2255E-ART. 2022, pp. 1–5. DOI: 10.1109/IWSSIP55020.2022.9854474.

[116] M. A. Hamer et al. "Validation of image analysis techniques to measure skin aging features from facial photographs". In: *Skin Research and Technology* 21.4 (2015), pp. 392–402. ISSN: 1600-0846. DOI: 10.1111/srt.12205.

[117] Hu Han, Charles Otto, and Anil K. Jain. "Age estimation from face images: Human vs. machine performance". In: *2013 International Conference on Biometrics (ICB)*. 2013, pp. 1–8. DOI: 10.1109/ICB.2013.6613022.

[118] Mark Harris. "Many-core GPU computing with NVIDIA CUDA". In: *Proceedings of the 22nd annual international conference on Supercomputing*. ICS '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 1. ISBN: 978-1-60558-158-3. DOI: 10.1145/1375527.1375528. URL: https://dl.acm.org/doi/10.1145/1375527.1375528.

[119] Y. He et al. "Deep embedding network for robust age estimation". In: *2017 IEEE International Conference on Image Processing (ICIP)*. 2017, pp. 1092–1096. DOI: 10.1109/ICIP.2017.8296450.

[120] Ronald Henss. "Perceiving Age and Attractiveness in Facial Photographs". In: *Journal of Applied Social Psychology* 21.11 (1991), pp. 933–946. ISSN: 1559-1816. DOI: 10.1111/j.1559-1816.1991.tb00451.x.

[121] Shakediel Hiba and Yosi Keller. "Hierarchical Attention-based Age Estimation and Bias Estimation". In: arXiv:2103.09882 (2021). arXiv:2103.09882 [cs]. URL: http://arxiv.org/abs/2103.09882.

[122] G. E. Hinton and R. R. Salakhutdinov. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647.

[123] Sepp Hochreiter. "Untersuchungen zu dynamischen neuronalen Netzen". Available at https://people.idsia.ch/~juergen/SeppHochreiter1991ThesisAdvisorSchmidhuber.pdf. Diploma thesis. ETU Munich: Technische Universitat Munchen, 1991.

[124] R. Iga et al. "A gender and age estimation system from face images". In: *SICE 2003 Annual Conference (IEEE Cat. No.03TH8734)*. Vol. 1. 2003, 756–761 Vol.1.

[125] Woobin Im et al. "Scale-Varying Triplet Ranking with Classification Loss for Facial Age Estimation". en. In: *Computer Vision – ACCV 2018*. Ed. by C.V. Jawahar et al. Vol. 11365. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 247–259. ISBN: 978-3-030-20872-1. DOI: 10.1007/978-3-030-20873-8_16. URL: http://link.springer.com/10.1007/978-3-030-20873-8_16.

[126] Marwa Jabberi et al. "68 landmarks are efficient for 3D face alignment: what about more?" en. In: *Multimedia Tools and Applications* (2023). ISSN: 1573-7721. DOI: 10.1007/s11042-023-14770-x. URL: https://doi.org/10.1007/s11042-023-14770-x.

[127] Randa Jdid et al. "Validation of digital photographic reference scales for evaluating facial aging signs". In: *Skin Research and Technology* 24.2 (2018), pp. 196–202. ISSN: 0909752X. DOI: 10.1111/srt.12413.

[128] Zhe Jia et al. "Dissecting the Graphcore IPU Architecture via Microbenchmarking". In: arXiv:1912.03413 (2019). arXiv:1912.03413 [cs]. DOI: 10.48550/arXiv.1912.03413. URL: http://arxiv.org/abs/1912.03413.

[129] Norman P. Jouppi et al. "In-Datacenter Performance Analysis of a Tensor Processing Unit". In: arXiv:1704.04760 (2017). arXiv:1704.04760 [cs]. DOI: 10.48550/arXiv.1704.04760. URL: http://arxiv.org/abs/1704.04760.

[130] Jared Kaplan et al. "Scaling Laws for Neural Language Models". In: arXiv:2001.08361 (2020). arXiv:2001.08361 [cs, stat]. DOI: 10.48550/arXiv.2001.08361. URL: http://arxiv.org/abs/2001.08361.

[131] Kimmo Karkkainen and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation". en. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, 2021, pp. 1547–1557. ISBN: 978-1-66540-477-8. DOI: 10.1109/WACV48630.2021.00159. URL: https://ieeexplore.ieee.org/document/9423296/.

[132] Vahid Kazemi and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees". en. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus, OH: IEEE, 2014, pp. 1867–1874. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.241. URL: https://ieeexplore.ieee.org/document/6909637.

[133] Henry J. Kelley. "Gradient Theory of Optimal Flight Paths". en. In: *ARS Journal* 30.10 (1960), pp. 947–954. ISSN: 1936-9972. DOI: 10.2514/8.5282.

[134] Jehandad Khan et al. *MIOpen: An Open Source Library For Deep Learning Primitives*. 2019. arXiv: 1910.00078 [cs.LG].

[135] Eugene J. Kim, Jay B. Reeck, and Corey S. Maas. "A Validated Rating Scale for Hyperkinetic Facial Lines". In: *Archives of Facial Plastic Surgery* 6.4 (2004), p. 253. ISSN: 1521-2491. DOI: 10.1001/archfaci.6.4.253.

[136] Sangkyu Kim et al. "The frailty index outperforms DNA methylation age and its derivatives as an indicator of biological age". In: *GeroScience* 39.1 (2017), pp. 83–92. ISSN: 2509-2723. DOI: 10.1007/s11357-017-9960-3.

[137] Taewoon Kim. "Generalizing MLPs With Dropouts, Batch Normalization, and Skip Connections". In: arXiv:2108.08186 (2021). arXiv:2108.08186 [cs]. DOI: 10.48550/arXiv.2108.08186. URL: http://arxiv.org/abs/2108.08186.

[138] Davis E. King. "Dlib-ml: A Machine Learning Toolkit". In: *Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[139] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. "Susceptibility to Image Resolution in Face Recognition and Trainings Strategies". In: *Leibniz Transactions on Embedded Systems* (2022). arXiv:2107.03769 [cs], 01:1–01:20 Pages. DOI: 10.4230/LITES.8.1.1.

[140] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[141] Pei-Lun Kuo et al. "Epigenetic Age Acceleration and Hearing: Observations From the Baltimore Longitudinal Study of Aging". In: *Frontiers in Aging Neuroscience* 13 (2021). ISSN: 1663-4365. URL: https://www.frontiersin.org/articles/10.3389/fnagi.2021.790926.

[142] Young H Kwon and Niels da Vitoria Lobo. "Age Classification from Facial Images". In: *Computer Vision and Image Understanding* 74.1 (1999), pp. 1–21. ISSN: 1077-3142. DOI: 10.1006/cviu.1997.0549.

[143] Samuli Laine and Timo Aila. "Temporal Ensembling for Semi-Supervised Learning". In: arXiv:1610.02242 (2017). arXiv:1610.02242 [cs]. DOI: 10.48550/arXiv.1610.02242. URL: http://arxiv.org/abs/1610.02242.

[144] A. Lanitis, C. Draganova, and C. Christodoulou. "Comparing Different Classifiers for Automatic Age Estimation". In: *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)* 34.1 (2004), pp. 621–628. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2003.817091.

[145] Irina Lebedeva, Fangli Ying, and Yi Guo. "Personalized facial beauty assessment: a meta-learning approach". en. In: *The Visual Computer* 39.3 (2023), pp. 1095–1107. ISSN: 1432-2315. DOI: 10.1007/s00371-021-02387-w.

[146] Yann LeCun et al. "Handwritten Digit Recognition with a Back-Propagation Network". In: *Advances in Neural Information Processing Systems*. Vol. 2. Morgan-Kaufmann, 1989. URL: https://proceedings.neurips.cc/paper/1989/hash/53c3bce66e43be4f209556518c2fcb54-Abstract.html.

[147] Kai Li et al. "D2C: Deep cumulatively and comparatively learning for human age estimation". en. In: *Pattern Recognition* 66 (2017), pp. 95–105. ISSN: 00313203. DOI: 10.1016/j.patcog.2017.01.007.

[148] Peipei Li et al. "Deep label refinement for age estimation". en. In: *Pattern Recognition* 100 (2020), p. 107178. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2019.107178.

[149] Shichao Li and Kwang-Ting Cheng. "Visualizing the decision-making process in deep neural decision forest". In: *CoRR* abs/1904.09201 (2019). arXiv: 1904.09201. URL: http://arxiv.org/abs/1904.09201.

[150] Wanhua Li et al. "BridgeNet: A Continuity-Aware Probabilistic Network for Age Estimation". In: *arXiv:1904.03358 [cs]* (2019). arXiv: 1904.03358. URL: http://arxiv.org/abs/1904.03358.

[151] Wanhua Li et al. "MetaAge: Meta-Learning Personalized Age Estimators". en. In: *IEEE Transactions on Image Processing* 31 (2022). arXiv:2207.05288 [cs], pp. 4761–4775. ISSN: 1057-7149, 1941-0042. DOI: 10.1109/TIP.2022.3188061.

[152] Xia Li et al. "Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up". In: *eLife* 9 (2020). ISSN: 2050-084X. DOI: 10.7554/eLife.51507. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7012595/.

[153] Luojun Lin, Lingyu Liang, and Lianwen Jin. "Regression Guided by Relative Ranking Using Convolutional Neural Network (R3CNN) for Facial Beauty Prediction". In: *IEEE Transactions on Affective Computing* (2019), pp. 1–1. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2019.2933523.

[154] Yiming Lin et al. "FP-Age: Leveraging Face Parsing Attention for Facial Age Estimation in the Wild". In: arXiv:2106.11145 (2022). arXiv:2106.11145 [cs]. DOI: 10.48550/arXiv.2106.11145. URL: http://arxiv.org/abs/2106.11145.

[155] Fan Liu et al. "The MC1R Gene and Youthful Looks". en. In: *Current Biology* 26.9 (2016), pp. 1213–1220. ISSN: 09609822. DOI: 10.1016/j.cub.2016.03.008.

[156] Hao Liu et al. "Ordinal Deep Learning for Facial Age Estimation". en. In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.2 (2019), pp. 486–501. ISSN: 1051-8215, 1558-2205. DOI: 10.1109/TCSVT.2017.2782709.

[157] Hao Liu et al. "Similarity-Aware and Variational Deep Adversarial Learning for Robust Facial Age Estimation". In: *IEEE Transactions on Multimedia* 22.7 (2020), pp. 1808–1822. ISSN: 1941-0077. DOI: 10.1109/TMM.2020.2969793.

[158] Wei Liu et al. "SSD: Single Shot MultiBox Detector". In: vol. 9905. arXiv:1512.02325 [cs]. 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0_2. URL: http://arxiv.org/abs/1512.02325.

[159] Xin Liu et al. "AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation". In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015, pp. 258–266. DOI: 10.1109/ICCVW.2015.42.

[160] Xinhua Liu et al. "Face Image Age Estimation Based on Data Augmentation and Lightweight Convolutional Neural Network". en. In: *Symmetry* 12.11 (2020), p. 146. DOI: 10.3390/sym12010146.

[161] Ake T. Lu et al. "DNA methylation GrimAge strongly predicts lifespan and healthspan". In: *Aging (Albany NY)* 11.2 (2019), pp. 303–327. ISSN: 1945-4589. DOI: 10.18632/aging.101684.

[162] Heinz-Theo Lübbers et al. "Precision and accuracy of the 3dMD photogrammetric system in craniomaxillofacial application". In: *Journal of Craniofacial Surgery* 21.3 (2010), pp. 763–767.

[163] Camillo Lugaresi et al. "MediaPipe: A Framework for Building Perception Pipelines". In: arXiv:1906.08172 (2019). arXiv:1906.08172 [cs]. DOI: 10.48550/arXiv.1906.08172. URL: http://arxiv.org/abs/1906.08172.

[164] Debbie S. Ma, Joshua Correll, and Bernd Wittenbrink. "The Chicago face database: A free stimulus set of faces and norming data". en. In: *Behavior Research Methods* 47.4 (2015), pp. 1122–1135. ISSN: 1554-3528. DOI: 10.3758/s13428-014-0532-5.

[165] Raju Machupalli, Masum Hossain, and Mrinal Mandal. "Review of ASIC accelerators for deep neural network". In: *Microprocessors and Microsystems* 89 (2022), p. 104441. ISSN: 0141-9331. DOI: 10.1016/j.micpro.2022.104441.

[166] Refik Can Malli, Mehmet Aygun, and Hazim Kemal Ekenel. "Apparent Age Estimation Using Ensemble of Deep Learning Models". In: *arXiv:1606.02909 [cs]* (2016). arXiv: 1606.02909. URL: http://arxiv.org/abs/1606.02909.

[167] Stefano Markidis et al. "NVIDIA Tensor Core Programmability, Performance & Precision". In: *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. arXiv:1803.04014 [cs]. 2018, pp. 522–531. DOI: 10.1109/IPDPSW.2018.00091. URL: http://arxiv.org/abs/1803.04014.

[168] Markus Mathias et al. "Face Detection without Bells and Whistles". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Vol. 8692. Lecture Notes in Computer Science. Springer International Publishing, 2014, pp. 720–735. ISBN: 978-3-319-10592-5. DOI: 10.1007/978-3-319-10593-2_47. URL: http://link.springer.com/10.1007/978-3-319-10593-2_47.

[169] Masakazu Matsugu et al. "Subject independent facial expression recognition with robust face detection using a convolutional neural network". en. In: *Neural Networks*. Advances in Neural Networks Research: IJCNN '03 16.5 (2003), pp. 555–559. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00115-1.

[170] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". en. In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133. ISSN: 1522-9602. DOI: 10.1007/BF02478259.

[171] Selma Mekić et al. "Younger facial looks are associate with a lower likelihood of several age-related morbidities in the middle-aged to elderly". en. In: *British Journal of Dermatology* (2023), ljac100. ISSN: 0007-0963, 1365-2133. DOI: 10.1093/bjd/ljac100.

[172] Meredith Minear and Denise C. Park. "A lifespan database of adult facial stimuli". en. In: *Behavior Research Methods, Instruments, & Computers* 36.4 (2004), pp. 630–633. ISSN: 1532-5970. DOI: 10.3758/BF03206543.

[173] Stylianos Moschoglou et al. "AgeDB: The First Manually Collected, In-the-Wild Age Database". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2017, pp. 1997–2005. DOI: 10.1109/CVPRW.2017.250.

[174] Rhoda S. Narins et al. "Validated Assessment Scales for the Lower Face:" in: *Dermatologic Surgery* 38.2ptII (2012), pp. 333–342. ISSN: 1076-0512. DOI: 10.1111/j.1524-4725.2011.02247.x.

[175] Zhenxing Niu et al. "Ordinal Regression with Multiple Output CNN for Age Estimation". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4920–4928. DOI: 10.1109/CVPR.2016.532.

[176] A. Nkengne et al. "Influence of facial skin attributes on the perceived age of Caucasian women". In: *Journal of the European Academy of Dermatology and Venereology* 22.8 (2008), pp. 982–991. ISSN: 1468-3083. DOI: 10.1111/j.1468-3083.2008.02698.x.

[177] Alex Nkengne, Georgios N Stamatas, and Christiane Bertin. “Facial Skin Attributes and Age Perception”. In: (2015), p. 12.

[178] Alex Nkengne et al. “The skin aging index: a new approach for documenting anti-aging products or procedures”. In: *Skin Research and Technology* 19.3 (2013), pp. 291–298. ISSN: 1600-0846. DOI: 10.1111/srt.12040.

[179] Haruko C. Okada et al. “Facial Changes Caused by Smoking: A Comparison between Smoking and Nonsmoking Identical Twins”. en-US. In: *Plastic and Reconstructive Surgery* 132.5 (2013), p. 1085. ISSN: 0032-1052. DOI: 10.1097/PRS.0b013e3182a4c20a.

[180] Lili Pan et al. “Self-Paced Deep Regression Forests with Consideration on Underrepresented Examples”. In: *arXiv:2004.01459 [cs]* (2020). arXiv: 2004.01459. URL: http://arxiv.org/abs/2004.01459.

[181] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[182] Gavin Rowe Paul Willner. “Alcohol Servers’ Estimates of Young People’s Ages”. In: *Drugs: Education, Prevention and Policy* 8.4 (2001), pp. 375–383. ISSN: 0968-7637. DOI: 10.1080/09687630010019299.

[183] P. Jonathon Phillips et al. “The FERET database and evaluation procedure for face-recognition algorithms”. In: *Image and Vision Computing* 16.5 (1998), pp. 295–306. ISSN: 0262-8856. DOI: 10.1016/S0262-8856(97)00070-X.

[184] A. Porcheron et al. “Influence of skin ageing features on Chinese women’s perception of facial age and attractiveness”. In: *International Journal of Cosmetic Science* 36.4 (2014), pp. 312–320. ISSN: 1468-2494. DOI: 10.1111/ics.12128.

[185] *Post-NIPS*95 Workshop on Transfer in Inductive Systems*. URL: https://plato.acadiau.ca/courses/comp/dsilver/NIPS95_LTL/transfer.workshop.1995.html.

[186] Lixiong Qin et al. “SwinFace: A Multi-task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023), pp. 1–1. ISSN: 1558-2205. DOI: 10.1109/TCSVT.2023.3304724.

[187] Rajat Raina, Anand Madhavan, and Andrew Y. Ng. “Large-scale deep unsupervised learning using graphics processors”. en. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal Quebec Canada: ACM, 2009, pp. 873–880. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553486. URL: https://dl.acm.org/doi/10.1145/1553374.1553486.

[188] U L Raj et al. “Impact of Dietary Supplements on Skin Aging”. en. In: (), p. 13.

[189] K. Ricanek and T. Tesafaye. “MORPH: a longitudinal image database of normal adult age-progression”. In: *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*. 2006, pp. 341–345. DOI: 10.1109/FGR.2006.78.

[190] Jonathan S. Rosenfeld et al. “A Constructive Prediction of the Generalization Error Across Scales”. In: arXiv:1909.12673 (2019). arXiv:1909.12673 [cs, stat]. DOI: 10.48550/arXiv.1909.12673. URL: http://arxiv.org/abs/1909.12673.

[191] Rasmus Rothe, Radu Timofte, and Luc Van Gool. “Deep expectation of real and apparent age from a single image without facial landmarks”. In: *International Journal of Computer Vision* 126.2-4 (2018), pp. 144–157.

[192] Rasmus Rothe, Radu Timofte, and Luc Van Gool. “DEX: Deep EXpectation of Apparent Age from a Single Image”. In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2015, pp. 252–257. ISBN: 978-1-4673-9711-7. DOI: 10.1109/ICCVW.2015.41. URL: http://ieeexplore.ieee.org/document/7406390/.

[193] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. “Learning representations by back-propagating errors”. en. In: *Nature* 323.60886088 (1986), pp. 533–536. ISSN: 1476-4687. DOI: 10.1038/323533a0.

[194] Berthold Rzany et al. “Validated Composite Assessment Scales for the Global Face:” in: *Dermatologic Surgery* 38.2ptII (2012), pp. 294–308. ISSN: 1076-0512. DOI: 10.1111/j.1524-4725.2011.02252.x.

[195] Carl F. Sabottke and Bradley M. Spieler. “The Effect of Image Resolution on Deep Learning in Radiography”. In: *Radiology: Artificial Intelligence* 2.1 (2020), e190015. DOI: 10.1148/ryai.2019190015.

[196] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. “Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/30ef30b64204a3088a26bc2e6ecf7602-Abstract.html.

[197] Nadine Samson et al. “Visible changes of female facial skin surface topography in relation to age and attractiveness perception”. en. In: *Journal of Cosmetic Dermatology* 9.2 (2010), pp. 79–88. ISSN: 1473-2165. DOI: 10.1111/j.1473-2165.2010.00489.x.

[198] Osman Semih Kayhan and Jan C. van Gemert. “On Translation Invariance in CNNs: Convolutional Layers Can Exploit Absolute Spatial Location”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 14262–14273. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.01428. URL: https://ieeexplore.ieee.org/document/9156444/.

[199] Sefik Ilkin Serengil and Alper Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE. 2021, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: https://doi.org/10.1109/ICEET53442.2021.9659697.

[200] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE. 2020, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802. URL: https://doi.org/10.1109/ASYU50717.2020.9259802.

[201] Wei Shen et al. "Deep Regression Forests for Age Estimation". en. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT: IEEE, 2018, pp. 2304–2313. ISBN: 978-1-5386-6420-9. DOI: 10.1109/CVPR.2018.00245. URL: https://ieeexplore.ieee.org/document/8578343/.

[202] Ryota Shimizu et al. "Balanced Mini-Batch Training for Imbalanced Image Data Classification with Neural Network". In: *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*. 2018, pp. 27–30. DOI: 10.1109/AI4I.2018.8665709.

[203] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. "Moving Window Regression: A Novel Approach to Ordinal Regression". In: arXiv:2203.13122 (2022). arXiv:2203.13122 [cs]. URL: http://arxiv.org/abs/2203.13122.

[204] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[205] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: arXiv:1409.1556 (2015). arXiv:1409.1556 [cs]. DOI: 10.48550/arXiv.1409.1556. URL: http://arxiv.org/abs/1409.1556.

[206] Timo E. Strandberg et al. "Association of Telomere Length in Older Men With Mortality and Midlife Body Mass Index and Smoking". In: *The Journals of Gerontology: Series A* 66A.7 (2011), pp. 815–820. ISSN: 1079-5006. DOI: 10.1093/gerona/glr064.

[207] Li Sun et al. "Facial age estimation through self-paced learning". In: *2017 IEEE Visual Communications and Image Processing (VCIP)*. 2017, pp. 1–4. DOI: 10.1109/VCIP.2017.8305113.

[208] Avinash Swaminathan et al. "Gender Classification using Facial Embeddings: A Novel Approach". en. In: *Procedia Computer Science*. International Conference on Computational Intelligence and Data Science 167 (2020), pp. 2634–2642. ISSN: 1877-0509. DOI: 10.1016/j.procs.2020.03.342.

[209] Christian Szegedy et al. "Going Deeper with Convolutions". In: arXiv:1409.4842 (2014). arXiv:1409.4842 [cs]. DOI: 10.48550/arXiv.1409.4842. URL: http://arxiv.org/abs/1409.4842.

[210] Shahram Taheri and Önsen Toygar. "On the use of DAG-CNN architecture for age estimation with multi-stage features fusion". en. In: *Neurocomputing* 329 (2019), pp. 300–310. ISSN: 0925-2312. DOI: 10.1016/j.neucom.2018.10.071.

[211] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: arXiv:1905.11946 (2020). arXiv:1905.11946 [cs, stat]. DOI: 10.48550/arXiv.1905.11946. URL: http://arxiv.org/abs/1905.11946.

[212] Xu Tang et al. "PyramidBox: A Context-assisted Single Shot Face Detector". In: arXiv:1803.07737 (2018). arXiv:1803.07737 [cs]. DOI: 10.48550/arXiv.1803.07737. URL: http://arxiv.org/abs/1803.07737.

[213] Imad Eddine Toubal et al. "Single View Facial Age Estimation Using Deep Learning with Cascaded Random Forests". In: *Computer Analysis of Images and Patterns*. Ed. by Nicolas Tsapatsoulis et al. Cham: Springer International Publishing, 2021, pp. 285–296.

[214] Gerhard Tutz. "Ordinal regression: A review and a taxonomy of models". en. In: *WIREs Computational Statistics* 14.2 (2022), e1545. ISSN: 1939-0068. DOI: 10.1002/wics.1545.

[215] F Valet, K Ezzedine, and D Malvy. "Assessing the reliability of four severity scales depicting skin ageing features". In: *British Journal of Dermatology* (2009), p. 6.

[216] Fabien Valet et al. "Assessing Quality of Ordinal Scales Depicting Skin Aging Severity". In: *Textbook of Aging Skin*. Ed. by Miranda A. Farage, Kenneth W. Miller, and Howard I. Maibach. Springer, 2017, pp. 1569–1577. ISBN: 978-3-662-47398-6. DOI: 10.1007/978-3-662-47398-6_87. URL: https://doi.org/10.1007/978-3-662-47398-6_87.

[217] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[218] Jenny Vestlund et al. "Experts on age estimation". en. In: *Scandinavian Journal of Psychology* 50.4 (2009), pp. 301–307. ISSN: 00365564, 14679450. DOI: 10.1111/j.1467-9450.2009.00726.x.

[219] Jenny Vestlund et al. "Experts on age estimation". In: *Scandinavian Journal of Psychology* 50.4 (2009), pp. 301–307. ISSN: 00365564, 14679450. DOI: 10.1111/j.1467-9450.2009.00726.x.

[220] P. Viola and M. Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Vol. 1. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

[221] Hee Lin Wang et al. "Effects of facial alignment for age estimation". In: *2010 11th International Conference on Control Automation Robotics & Vision*. 2010, pp. 644–647. DOI: 10.1109/ICARCV.2010.5707877.

[222] Shengzheng Wang, Dacheng Tao, and Jie Yang. "Relative Attribute SVM+ Learning for Age Estimation". In: *IEEE Transactions on Cybernetics* 46.3 (Mar. 2016), pp. 827–839. ISSN: 2168-2275. DOI: 10.1109/TCYB.2015.2416321.

[223] Duorui Xie et al. "SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception". In: *arXiv:1511.02459 [cs]* (2015). arXiv: 1511.02459. URL: http://arxiv.org/abs/1511.02459.

[224] Qizhe Xie et al. "Unsupervised Data Augmentation for Consistency Training". In: arXiv:1904.12848 (2020). arXiv:1904.12848 [cs, stat]. DOI: 10.48550/arXiv.1904.12848. URL: http://arxiv.org/abs/1904.12848.

[225] Jin Xu et al. "Novel Gene Expression Profile of Women with Intrinsic Skin Youthfulness by Whole Transcriptome Sequencing". en. In: *PLOS ONE* 11.11 (2016), e0165913. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0165913.

[226] Lu Xu, Jinhai Xiang, and Xiaohui Yuan. "Transferring Rich Deep Features for Facial Beauty Prediction". en. In: *arXiv:1803.07253 [cs]* (2018). arXiv: 1803.07253. URL: http://arxiv.org/abs/1803.07253.

[227] Lu Xu, Jinhai Xiang, and Xiaohui Yuan. "Transferring Rich Deep Features for Facial Beauty Prediction". en. In: *arXiv:1803.07253 [cs]* (2018). arXiv: 1803.07253. URL: http://arxiv.org/abs/1803.07253.

[228] Yichong Xu et al. "Scale-Invariant Convolutional Neural Networks". In: arXiv:1411.6369 (2014). arXiv:1411.6369 [cs]. DOI: 10.48550/arXiv.1411.6369. URL: http://arxiv.org/abs/1411.6369.

[229] Haibin Yan. "Cost-sensitive ordinal regression for fully automatic facial beauty assessment". en. In: *Neurocomputing* 129 (2014), pp. 334–342. ISSN: 09252312. DOI: 10.1016/j.neucom.2013.09.025.

[230] Huei-Fang Yang et al. "Automatic Age Estimation from Face Images via Deep Ranking". en. In: *Procedings of the British Machine Vision Conference 2015*. Swansea: British Machine Vision Association, 2015, pp. 55.1–55.11. ISBN: 978-1-901725-53-7. DOI: 10.5244/C.29.55. URL: http://www.bmva.org/bmvc/2015/papers/paper055/index.html.

[231] Xiangli Yang et al. "A Survey on Deep Semi-supervised Learning". In: arXiv:2103.00550 (2021). arXiv:2103.00550 [cs]. URL: http://arxiv.org/abs/2103.00550.

[232] Xiao Yang. "An Overview of the Attention Mechanisms in Computer Vision". en. In: *Journal of Physics: Conference Series* 1693.1 (2020), p. 012173. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1693/1/012173.

[233] Xu Yang et al. "Deep Label Distribution Learning for Apparent Age Estimation". In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015, pp. 344–350. DOI: 10.1109/ICCVW.2015.53.

[234] Dong Yi, Zhen Lei, and Stan Z Li. "Age Estimation by Multi-scale Convolutional Network". en. In: (2014).

[235] Dong Yi et al. "Learning Face Representation from Scratch". In: arXiv:1411.7923 (2014). arXiv:1411.7923 [cs]. DOI: 10.48550/arXiv.1411.7923. URL: http://arxiv.org/abs/1411.7923.

[236] T. Ylonen and C. Lonvick. *The Secure Shell (SSH) Connection Protocol*. RFC 4254 (Proposed Standard). Internet Engineering Task Force, 2006. URL: http://www.ietf.org/rfc/rfc4254.txt.

[237] Changqian Yu et al. "BiSeNet V2: Bilateral Network with Guided Aggregation for Real-time Semantic Segmentation". In: arXiv:2004.02147 (Apr. 2020). arXiv:2004.02147 [cs]. DOI: 10.48550/arXiv.2004.02147. URL: http://arxiv.org/abs/2004.02147.

[238] Xusheng Zeng et al. "Soft-ranking Label Encoding for Robust Facial Age Estimation". In: arXiv:1906.03625 (2019). arXiv:1906.03625 [cs]. URL: http://arxiv.org/abs/1906.03625.

[239] Yikui Zhai et al. "BeautyNet: Joint Multiscale CNN and Transfer Learning Method for Unconstrained Facial Beauty Prediction". en. In: *Computational Intelligence and Neuroscience* 2019 (2019), e1910624. ISSN: 1687-5265. DOI: 10.1155/2019/1910624.

[240] Bob Zhang, Xihua Xiao, and Guangming Lu. "Facial beauty analysis based on features prediction and beautification models". en. In: *Pattern Analysis and Applications* 21.2 (2018), pp. 529–542. ISSN: 1433-7541, 1433-755X. DOI: 10.1007/s10044-017-0647-2.

[241] Chao Zhang et al. "C3AE: Exploring the Limits of Compact Model for Age Estimation". In: arXiv:1904.05059 (2019). arXiv:1904.05059 [cs]. DOI: 10.48550/arXiv.1904.05059. URL: http://arxiv.org/abs/1904.05059.

[242] Kaipeng Zhang et al. "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (2016). arXiv:1604.02878 [cs], pp. 1499–1503. ISSN: 1070-9908, 1558-2361. DOI: 10.1109/LSP.2016.2603342.

[243] Kaipeng Zhang et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503. DOI: 10.1109/lsp.2016.2603342. URL: https://doi.org/10.1109%2Flsp.2016.2603342.

[244] Meng M. Zhang et al. "Exploring artificial intelligence from a clinical perspective: A comparison and application analysis of two facial age predictors trained on a large-scale Chinese cosmetic patient database". eng. In: *Skin research and technology: official journal of International Society for Bioengineering and the Skin (ISBS) [and] International Society for Digital Imaging of Skin (ISDIS) [and] International Society for Skin Imaging (ISSI)* 29.7 (2023), e13402. ISSN: 1600-0846. DOI: 10.1111/srt.13402.

[245] Shifeng Zhang et al. "S^{3}FD: Single Shot Scale-invariant Face Detector". In: arXiv:1708.05237 (2017). arXiv:1708.05237 [cs]. DOI: 10.48550/arXiv.1708.05237. URL: http://arxiv.org/abs/1708.05237.

[246] Song Yang Zhang Zhifei and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017.

[247] Yu Zhang and Dit-Yan Yeung. "Multi-task warped Gaussian process for personalized age estimation". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 2622–2629. DOI: 10.1109/CVPR.2010.5539975.

[248] Yunxuan Zhang et al. "Quantifying Facial Age by Posterior of Age Comparisons". en. In: *Procedings of the British Machine Vision Conference 2017*. London, UK: British Machine Vision Association, 2017, p. 108. ISBN: 978-1-901725-60-5. DOI: 10.5244/C.31.108. URL: http://www.bmva.org/bmvc/2017/papers/paper108/index.html.

[249] Qilu Zhao et al. "Distilling Ordinal Relation and Dark Knowledge for Facial Age Estimation". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.7 (2021), pp. 3108–3121. ISSN: 2162-2388. DOI: 10.1109/TNNLS.2020.3009523.

[250] SONG ZHENG. "Visual Image Recognition System with Object-Level Image Representation". en. Accepted: 2013-04-30T18:00:17Z. Thesis. 2012. URL: https://scholarbank.nus.edu.sg/handle/10635/37543.

[251] Yinglin Zheng et al. "General Facial Representation Learning in a Visual-Linguistic Manner". In: arXiv:2112.03109 (2022). arXiv:2112.03109 [cs]. DOI: 10.48550/arXiv.2112.03109. URL: http://arxiv.org/abs/2112.03109.

[252] Yu Zhu et al. "A Study on Apparent Age Estimation". In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. 2015, pp. 267–273. DOI: 10.1109/ICCVW.2015.43.

[253] Zhuoting Zhu et al. "Retinal age gap as a predictive biomarker for mortality risk". en. In: *British Journal of Ophthalmology* 107.4 (2023), pp. 547–554. ISSN: 0007-1161, 1468-2079. DOI: 10.1136/bjophthalmol-2021-319807.