
# Age-Related Facial Analysis with Deep Learning

by

**Haoyi Wang**

**Thesis**

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

**Doctor of Philosophy**

**Department of Computer Science**

September 2020

# Contents

# List of Tables

# List of Figures

# Acknowledgments

I would like to express my gratitude to my supervisors, Professor Chang-Tsun Li and Dr. Victor Sanchez, who guided me throughout PhD time. I really appreciate their patience support and helps by providing me with many insightful comments and suggestions on my academic research. I would also like to thank Professor Dong Xu and Dr. Wanli Ouyang at the University of Sydney and Professor Liang Wang at the Institute of Automation at Chinese Academic of Sciences for providing me with excellent research environments during my 12 months secondment under the EU IDENTITY project. I also want to thank all my friends, my colleagues, my team members at the University of Warwick for their support. Finally, my deep and sincere gratitude to my parents and my wife for their continuous and unparalleled love, help and support.

# Declarations

Parts of this thesis have been previously published by the author in the following papers:

- Haoyi Wang, Xingjie Wei, Victor Sanchez, and Chang-Tsun Li. Fusion Network for Face-based Age Estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 2675–2679, 2018.
- Haoyi Wang, Victor Sanchez, Wanli Ouyang, and Chang-Tsun Li. Using Age Information as a Soft Biometric Trait for Face Image Analysis. In *Deep Biometrics*, pages 1–20, 2020.

Parts of this thesis are based on the following manuscripts that are currently being reviewed:

- Haoyi Wang, Victor Sanchez, and Chang-Tsun Li. Age-Oriented Face Synthesis with Conditional Discriminator Pool and Adversarial Triplet Loss. Under reviewed at *IEEE Transactions on Image Processing*.
- Haoyi Wang, Victor Sanchez, and Chang-Tsun Li. Improving Face-Based Age Estimation with Attention-Based Dynamic Patch Fusion. Under reviewed at *IEEE Transactions on Image Processing*.

# Abstract

Age, as an important soft biometric trait, can be inferred based on the appearance of human faces. However, compared to other facial attributes like race and gender, age is rather subtle due to the underlying conditions of individuals (i.e., their upbringing environment and genes). These uncertainties make age-related facial analysis (including age estimation, age-oriented face synthesis and age-invariant face recognition) still unsolved. In this thesis, we study these age-related problems and propose several deep learning-based methods, each tackle a problem from a specific aspect.

We first propose a customised Convolutional Neural Network architecture called the FusionNet and also its extension to study the age estimation problem. Although faces are composed of numerous facial attributes, most deep learning-based methods still consider a face as a typical object and do not pay enough attention to facial regions that carry age-specific features for this particular task. Therefore, the proposed methods take several age-specific facial patches as part of the input to emphasise the learning of age-specific patches. Through extensive evaluation, we show that these methods outperform existing methods on age estimation benchmark datasets under various evaluation matrices.

Then, we propose a Generative Adversarial Network (GAN) model for age-oriented face synthesis. Specifically, to ensure that the synthesised images are within target age groups, this method tackles the mode collapse issue in vanilla GANs with a novel Conditional Discriminator Pool (CDP), which consists of multiple discriminators, each targeting one particular age category. To ensure the identity information

is unaltered in the synthesised images, our method uses a novel Adversarial Triplet loss. This loss, which is based on the Triplet loss, adds a ranking operation to further pull the positive embedding towards the anchor embedding resulting in significantly reduced intra-class variances in the feature space. Through extensive experiments, we show that our method can precisely transform input faces into the target age category while preserving the identity information on the synthesised faces.

Last but not least, we propose the disentangled contrastive learning (DCL) for unsupervised age-invariant face recognition. Different from existing AIFR methods, DCL, which aims to learn disentangled identity features, can be trained on any facial datasets and further tested on age-oriented datasets. Moreover, by utilising a set of three augmented samples derived from the same input image, Disentangled Contrastive Learning can be directly trained on small-sized datasets with promising performance. We further modify the conventional contrastive loss function to fit this training strategy with three augmented samples. We show that our method dramatically outperforms previous unsupervised methods and other contrastive learning methods.

# Sponsorships and Grants

All research works of this thesis are supported by the EU Horizon 2020 - Marie Sklodowska-Curie Actions through the project Computer Vision Enabled Multimedia Forensics and People Identification (Project No. 690907, Acronym: IDENTITY).

# Acronyms

**ADPF** Attention-based Dynamic Patch Fusion.

**AGES** AGing pattErn Subspace.

**AIFR** Age-Invariant Face Recognition.

**AIFV** Age-Invariant Face Verification.

**AOFS** Age-Oriented Face Synthesis.

**ASM** Active Shape Model.

**BIF** Bio-Inspired Features.

**CACD** Cross-Age Celebrity Dataset.

**CMC** Cumulative Match Curve.

**CNN** Convolutional Neural Network.

**CS** Cumulative Score.

**D2GAN** Dual Discriminator Generative Adversarial Nets.

**EM** Earth-Mover.

**FC** Fully-Connected.

**FID** Fréchet Inception Distance.

**FRD** Fréchet ResNet Distance.

**IPCGAN** Identity-Preserving Conditional Generative Adversarial Networks.

**IS** Inception Score.

**KL-divergence** Kullback-Leibler divergence.

**LBP** Local Binary Pattern.

**LDA** Linear Discriminant Analysis.

**LFW** Labelled Faces in the Wild.

**LOPO** Leave-one-person-out.

**LSGAN** Least Square Generative Adversarial Network.

**MAE** Mean Absolute Error.

**mAP** Mean Average Precision.

**MHSA** Multi-Head Self-Attention.

**MTFE** Multi-Task Feature Extractor.

**NLP** Natural Language Processing.

**PCA** Principal Component Analysis.

**PIE** Pose, Illumination, Expression.

**RMHHA** Ranking-based Multi-Head Hybrid Attention.

**RS** ResNet Score.

**SAGAN** Self-Attention Generative Adversarial Network.

**SELU** Scaled Exponential Linear Unit.

**SGD** Stochastic Gradient Descent.

**SVM** Support Vector Machine.

**SVR** Support Vector Regression.

**WGAN** Wasserstein Generative Adversarial Network.

# Symbols

| | |
|---|---|
| $x\|y$ | Conditional Probability |
| $D_{KL}(\cdot)$ | KL divergence |
| $\mathbb{E}$ | Expectation |
| $\cdot$ | Inter-Product |
| $\|x - y\|$ | L1 Distance |
| $\|\|x\|\|_2$ | L2 Normalisation |
| $(\cdot)^T$ | Matrix or Vector Transpose |

# Chapter 1

# Introduction

## 1.1 Age as a Soft Biometric Trait

Biometrics aim to determine the identity of an individual by leveraging the subjects' physiological or behavioural attributes [75]. Physiological attributes refer to the physical characteristics of the human body, like the face, iris, fingerprint, etc. On the other hand, behavioural attributes indicate the particular patterns of the behaviour of a person, which include gait, voice, keystroke dynamics, etc. Among all these biometrics attributes, the face is the most commonly used one due to its accessibility and the fact that face-based biometric systems require little cooperation from the subject.

Besides the identity information, other ancillary information like age, race and gender (often referred to as soft biometrics) can also be retrieved from the face. Soft biometrics is the set of traits that provide some information to describe individuals, but do not have the capability to discriminate identities due to their lack of distinctiveness and permanence [74]. Although soft biometric traits alone cannot distinguish among individuals, they can be used in conjunction with the identity information to boost the recognition or verification performance or be leveraged in other scenarios. For example, locating persons-of-interest based on a combination of soft biometric traits by using surveillance footage.

Compared to traditional biometrics, soft biometrics have the following merits. First, when the identity information is not available, soft biometrics can generate human-understandable descriptions to track the person-of-interest, such as in the 2013 Boston bombings [82]. Second, as the data abuse issue becomes more and more severe in the information era, using soft biometric traits to capture subjects' ancillary information can preserve their identity while achieving the expected goals. For example, companies can efficiently recommend merchandises by merely knowing the age or the gender of their potential customers. Third, collecting soft biometric traits do not require the participation of the subject, which makes them easy to compute.

Among all the soft biometric traits (age, gender, race, etc.) that can be obtained from facial images, age has the widest range of real-life applications. To begin with, the age information is widely utilised in security control and surveillance monitoring systems. By determining the user's age, vending machines or websites that contains adult-exclusive content can prevent teenagers from access. Moreover, faces within different age groups can be synthesised to predict the outcome of cosmetic surgeries and generate special visual effects on characters of video games and films [40]. Furthermore, age information can aid face recognition and verification systems to track person-of-interest such as missing children, people with dementia, or suspects over several years span [148].

## 1.2 Age-related Problems

Based on the form of expected output, the age-related problem can be categorised into three sub-problems: age estimation, AOFS, and AIFR. Specifically, age estimation is concerned with inferring the specific age from facial images; AOFS is concerned with the rendering of facial images with natural ageing or rejuvenating effects; AIFR involves the recognition of the identity of subjects correctly regardless of their age.

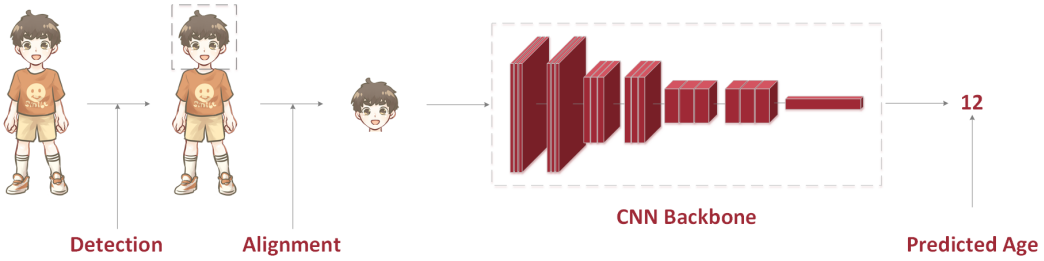


Figure 1.1: A simplified diagram of a deep learning-based age estimation model.

### 1.2.1 Age Estimation

The purpose of age estimation is to estimate the real age (cumulated years after birth) of the subject. The predicted age is mainly deduced based on the age-specific features extracted by the feature extractor. Modern face-based age estimation methods typically consist of two components, a feature extractor and an estimator. The feature extractor is used to extract age-specific features from raw facial images, and the estimator is used to predict the age based on the extracted features. Before deep learning-based methods dominated the computer vision field, researchers used to estimate ages with hand-crafted features [36, 43, 44]. With the growing size of age-oriented datasets, CNNs are now the foundation of feature extractors. A block diagram of a deep learning-based age estimation model can be found in Figure 1.1. Since we are only interested in the face region, the face is located and aligned from the original image before fed into the CNN model.

### 1.2.2 Age-Oriented Face Synthesis

Compared to age estimation, AOFS has not gained much attention from the research community yet. AOFS methods aim to generate elder or younger faces by rendering facial images with natural ageing or rejuvenating effects. The synthesis is usually conducted between age categories (e.g. the 20s, 30s, 40s) rather than specific ages (e.g. 22, 25, 29) since there is no noticeable visual change of a face over a several-year span. A block diagram of an AOFS model with two parallel processes, an ageing

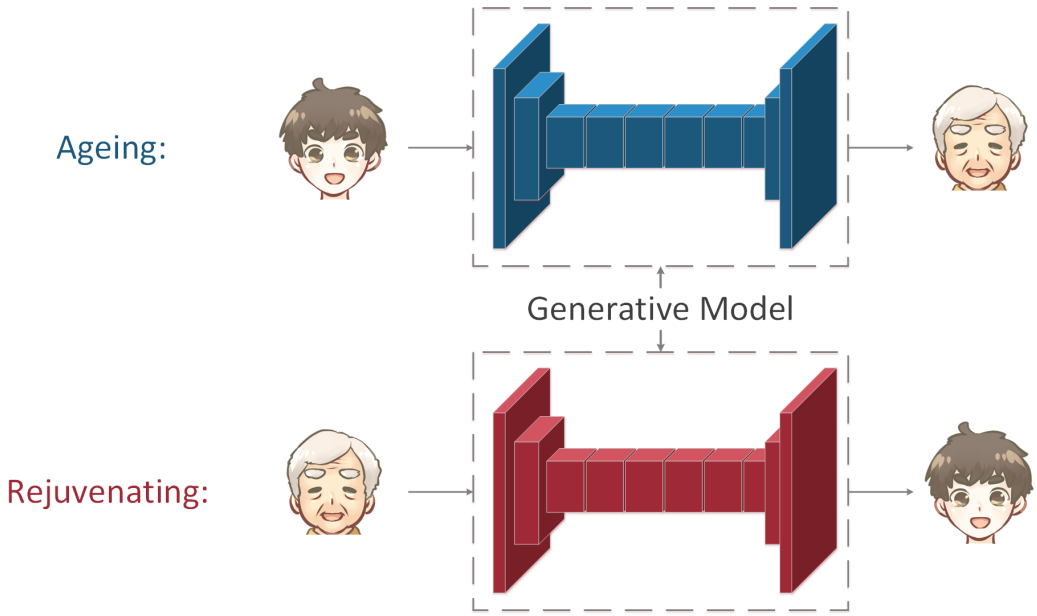


Figure 1.2: A simplified block diagram of an AOFS model.

process and a rejuvenating process, can be found in Figure 1.2.

### 1.2.3 Age-Invariant Face Recognition

AIFR aims to recognise the identity of subjects regardless of their age and is an important yet less studied topic compared to other sub-problems of face recognition. Different from the conventional face recognition problem, AIFR needs to consider the intra-class variance caused by the age information.

Existing AIFR methods can be categorised as either a discriminative model or a generative model [94, 152]. Discriminative models [49, 132, 152, 159] aim to learn and extract age-invariant features directly from input images while generative models [88, 119] synthesise samples that match the target age before the feature extraction.

## 1.3 Motivations

All three age-related facial analysis problems have gained more and more attention from the research community, and the performance has been boosted significantly thanks to the recent development of machine learning methods, especially deep learning methods. However, there is still a large margin to improve by paying attention to details for each problem. Here, we discuss the limitations of existing works tackling each problem and how the performance can be further improved.

As aforementioned, modern age estimation methods typically consist of two components, a feature extractor and an estimator. Most state-of-the-art works [14, 21, 39, 67, 93, 97, 98, 114, 117, 134] focus on designing customised estimators while treating the facial image as an ordinary input, hence paying no attention to the relative importance of the extracted features. However, related studies [53, 56] show that age-specific patches are useful when predicting the age of the subject from an image. In other words, customised feature extractors can be designed to exploit age-specific patches during training to boost the performance of face-based age estimation methods. Therefore, we focus on design customised feature extractors to further boost the performance of age estimation. Figure 1.3 exemplifies discovered age-specific patches represented as heatmaps. Each row in the figure depicts an age-specific patch cross different subjects.

Regarding the AOFS problem, in order to synthesise realistic images, the vanilla GAN [50] is commonly used as the backbone of state-of-the-art AOFS methods [5, 47, 92, 118, 171]. One of the biggest advantages of the vanilla GAN over other generative methods, like the Variational Autoencoder [79], is that it can generate sharp and realistic images by playing a minimax game between the generator and the discriminator.

However, the vanilla GAN suffers from the mode collapse issue caused by the vanishing gradient due to the involvement of the negative log-likelihood loss [6]. Specifically, once the discriminator converges, the loss does not penalise the

Figure 1.3: Five most informative age-specific patches.


Figure 1.4: A demonstration of mode collapse in AOFS.

generator any further [17]. This allows the generator to find a specific mode (i.e., a distribution) that can easily fool the discriminator [10]. The mode collapse issue may also occur in the AOFS problem, where a mode is represented by an age group. Within this context, the vanilla GAN may generate faces with limited variations as exemplified in Figure 1.4. The figure uses the ageing process as an example where the top row depicts images generated by a vanilla GAN suffering from the mode collapse issue, and the bottom row depicts images with rich and natural ageing effects.

For the AIFR problem, it is commonly known that cross-age facial images are usually expensive to collect, which makes the size of noise-free age-oriented datasets relatively small compared to that of widely-used large-scale facial datasets. A statistical comparison between a widely used noise-free age-oriented face dataset and a general large-scale face dataset is tabulated in Table 1.1. In the table, *#images* indicates the number of images in the dataset, *#images per subject* indicates the number of images per subject, and *SOTA performance* indicates the state-of-the-art performance achieved on the corresponding dataset. It is worth noting that the

Table 1.1: Comparison between a noise-free age-oriented face dataset and a large-scale face dataset.

| Dataset | #images | #images per subject | SOTA performance |
|---|---|---|---|
| Age-oriented | 1,002 | 12 | 60.01% [145] |
| General | 3,310,901 | 363 | 96.10% [13] |

performance achieved on the age-oriented dataset requires pre-training on large-scale datasets before fine-tuning and evaluating on the target dataset. Additionally, in real scenarios, images of the same subject at different ages are usually hard or even impossible to obtain, which further limits the versatility of supervised AIFR methods.

## 1.4 Contributions

Motivated by the ideas mentioned in Section 1.3, this thesis focuses on developing novel deep learning-based methods to tackle age-related facial analysis tasks. This thesis proposed four methods in total, two for age estimation, one for AOFS, and one for AIFR. The main contributions of this thesis are summarised as follows:

- We propose a customised CNN named FusionNet to solve the age estimation problem. To the best of our knowledge, our network is the first CNN-based model in which the learning of age-specific features is enhanced by using selected input patches. The facial patch selection process is based on the BIF and the AdaBoost algorithm. Moreover, these input patches form short-cut connections that complement the learning process, which is useful to boost the performance.

- To further improve the training efficiency and the performance of the FusionNet, we propose a framework called ADPF for the age estimation problem. Instead of using the BIF and the AdaBoost algorithm to locate age-specific patches, ADPF uses an AttentionNet, which includes a novel attention mechanism. The proposed attention mechanism dynamically produces ranked single-channel

attention maps, where each attention map highlights a particular patch. Additionally, to reduce the overlap among patches, we propose a diversity loss to force the attention mechanism to reveal diverse age-specific regions.

- Given the mode collapse issue in the GANs, we study this specific issue in the AOFS task. To the best of our knowledge, our work is the first to tackle the AOFS task from the aspect of mode learning. Specifically, to address the mode collapse issue in the vanilla GAN and attain a high synthesis accuracy, we propose the CDP, which allows our AOFS method to learn multiple modes explicitly and independently. To preserve the identity information in the synthesised images, we propose the Adversarial Triplet loss. Smaller intra-class variance can be achieved by forcing triplets to play zero-sum games during training.

- Instead of studying supervised AIFR problem, given the small-sized age-oriented datasets, we tackle the unsupervised AIFR problem by proposing the DCL that utilises three augmented samples from each input image. To learn disentangled identity features, the DCL maximise the similarity between features that represent the facial images of the same subject within different age groups. We also modify the conventional contrastive loss to fit the training strategy with three augmented samples.

## 1.5 Outline

This thesis is organised as follows:

- **Chapter 2: Literature Review**

  This chapter provides a comprehensive overview of machine learning and deep learning-based works from the research community that tackle the three age-related facial analysis problem. We also review several widely used age-oriented face dataset and various evaluation metrics for each problem.

- **Chapter 3: FusionNet for Age Estimation**

  This chapter discusses the proposed FusionNet that tackles the age estimation problem. CNNs have been applied to age-related research as the core framework. Although faces are composed of numerous facial attributes, most works with CNNs still consider a face as a typical object and do not pay enough attention to facial regions that carry age-specific feature for this particular task. To this end, we propose the FusionNet. Apart from the whole facial image, the FusionNet successively takes several age-specific facial patches as part of the input to emphasise the age-specific features. Through experiments, we show that the FusionNet significantly outperforms other state-of-the-art models on the MORPH II benchmark.

- **Chapter 4: Improving Age Estimation with Attention-Based Dynamic Patch Fusion**

  Chapter 4 presents the ADPF that is built based on the FusionNet. In ADPF, two separate CNNs are implemented, namely the AttentionNet and the FusionNet. The AttentionNet dynamically locates and ranks age-specific patches by employing a novel RMHHA mechanism. The FusionNet uses the discovered patches along with the facial image to predict the age of the subject. Since the proposed RMHHA mechanism ranks the discovered patches based on their importance, the length of the learning path of each patch in the FusionNet is proportional to the amount of information it carries (the longer, the more important). ADPF also introduces a novel diversity loss to guide the training of the AttentionNet and reduce the overlap among patches so that the diverse and important patches are discovered. Through extensive experiments, we show that our proposed framework outperforms state-of-the-art methods on several age estimation benchmark datasets.

- **Chapter 5: Age-Oriented Face Synthesis with Conditional Discriminator Pool and Adversarial Triplet Loss**

Chapter 5 focus on tackling the AOFS problem. The vanilla GANs are commonly used to generate realistic images depicting aged and rejuvenated faces. However, the performance of vanilla GANs in the AOFS problem is often compromised by the mode collapse issue, which may result in the generation of faces with minimal variations and a poor synthesis accuracy. In addition, recent AOFS methods use the L1 or L2 constraint to preserve the identity information on synthesised faces, which implicitly limits the identity permanence capabilities when these constraints are associated with a trivial weighting factor. To this end, we propose a method for the AOFS that achieves a high synthesis accuracy with strong identity permanence capabilities. Specifically, to achieve a high synthesis accuracy, our method tackles the mode collapse issue with a novel CDP, which consists of multiple discriminators, each targeting one particular age group. To achieve strong identity permanence capabilities, our method uses a novel Adversarial Triplet loss. This loss, which is based on the Triplet loss [131], adds a ranking operation to further pull the positive embedding towards the anchor embedding resulting in significantly reduced intra-class variances in the feature space. Through extensive experiments, we show that our proposed method outperforms state-of-the-art methods in terms of synthesis accuracy and identity permanence capabilities, qualitatively and quantitatively.

- **Chapter 6: Unsupervised Age-Invariant Face Recognition with Disentangled Contrastive Learning**

  Cross-age facial images are usually expensive to collect, which makes the size of noise-free age-oriented datasets relatively small compared to that of widely-used large-scale facial datasets. Additionally, in real scenarios, images of the same subject at different ages are usually hard or even impossible to obtain, which limits the versatility of supervised methods. To this end, we tackle the problem of unsupervised AIFR by proposing the DCL. DCL aims to learn disentangled identity features and can be trained on any facial datasets and further tested on

age-oriented datasets. Moreover, by utilising a set of three augmented samples derived from the same input image, DCL can be directly trained on small-sized datasets with promising performance. We further modify the conventional contrastive loss function to fit this training strategy with three augmented samples. To demonstrate the effectiveness of the proposed method, we conduct both homogeneous-dataset and cross-dataset experiments using several AIFR benchmark datasets and general facial datasets. Experimental results show that DCL outperforms state-of-the-art unsupervised method based on several evaluation metrics.

- **Chapter 7: Conclusion and Future Trends**

  This chapter concludes this thesis and discusses the future research trend by discussing the unaddressed issues in three age-related facial analysis problems.

# Chapter 2

# Literature Review

This chapter presents the survey of related datasets, evaluation metrics, and works in age estimation, AOFS, and AIFR. Section 2.1 begins by presenting the commonly used benchmark datasets for the age estimation problem, which followed by evaluation metrics for this problem. Then, traditional machine learning-based and deep learning-based age estimation methods and related techniques we use to tackle the problem are discussed. Section 2.2 follows the same presentation style and focuses on the AOFS problem. Section 2.3 focuses on the AIFR problem. After these three sections, Section 2.4 reviews related machine learning concepts that are related to our works. We conclude this chapter in Section 2.5.

## 2.1 Age Estimation

### 2.1.1 Datasets for Age Estimation

Among all the age-oriented datasets, the MORPH II dataset [126] is the most broadly used to evaluate age estimation models. This dataset contains more than 55,000 facial images from about 13,000 subjects with ages ranging from 16 to 77 with an average age of 33. Each image in the MORPH II dataset is associated with identity, age, race and gender labels. The second most commonly used dataset to evaluate age estimation models is the FG-NET dataset [28] which contains 1002 images from

Table 2.1: Most commonly used datasets to evaluate age estimation models.

| Dataset | #images | #subjects | age range | noise-free label | Mugshot |
|---|---|---|---|---|---|
| MORPH II | 55,134 | 13,618 | 16-77 | Yes | Yes |
| FG-NET | 1,002 | 82 | 0-69 | Yes | No |
| CACD | 163,446 | 2000 | 16-62 | No | No |
| IMDB-WIKI | 523,051 | 20,284 | 0-100 | No | No |

82 subjects. However, due to the limited number of images, the FG-NET dataset is usually only used during the evaluation phase. Since the training of CNN-based models requires a large number of training samples, to meet this requirement, two large-scale age-oriented datasets have been built, the Cross-Age Celebrity Dataset (CACD) [18] and the IMDB-WIKI dataset [130]. The CACD contains more than 160,000 facial images from 2000 individuals with ages ranging from 16 to 62. The IMDB-WIKI dataset contains 523,051 facial images (460,723 images from IMDB and 62,328 images from Wikipedia) from 20,284 celebrities. However, both datasets contain noisy (incorrect) labels. The details of these four datasets are tabulated in Table 2.1.

### 2.1.2 Evaluation Metrics for Age Estimation Models

There are two evaluation metrics commonly used for age estimation models. The first one is the MAE, which measures the average absolute difference between the predicted age and the ground truth:

$$MAE = \frac{\sum_{i=1}^{M} e_i}{M}, \tag{2.1}$$

where $e_i$ is the absolute error between the predicted age $\hat{l}_i$ and the input age label $l_i$ for the $i$-th sample. The denominator $M$ is the total number of testing samples.

The other evaluation metric is the CS, which measures the percentage of

images that are correctly classified in a certain range:

$$CS(n) = -\frac{M_n}{M} \times 100\%, \tag{2.2}$$

where $M_n$ is the number of images whose predicted age $\hat{l_i}$ is in the range of $[l_i - n, l_i + n]$, and and $n$ indicates the number of years.

### 2.1.3 Traditional Machine Leaning-based Age Estimation

In the past few decades, many works have been conducted on face-based age estimation. One of the earliest works can be traced back to [85], in which the researchers classify faces into three age groups based on the cranio-facial development theory and wrinkle analysis. Later, [153] reveals that wrinkles play an important role in modelling ageing faces and determining ages.

Before deep learning-based methods dominated the computer vision field, researchers used to develop face-based age estimation methods with hand-crafted features. For example, the Statistical Face Model [36] used in [88] is adopted to extract features and reveal the relationship between features and the corresponding age labels. Geng *et al.* [44, 45] propose the AGES to learn ageing pattern vectors in a representative subspace from training images. Unseen faces are then projected to this newly constructed subspace to predict their ages. Later, [43] reveals the ambiguity of mapping ages to age groups and proposes the Fuzzy LDA to build the classifier as an estimator. The authors define an Age Membership Function to encode the relevance between ages and age groups and integrate this function as a weighting factor into the conventional LDA. Guo *et al.* [51] propose a kernel-based regression method to tackle the face-based age estimation problem. A worth-noting algorithm designed to extract hand-crafted features for face-based age estimation is BIF [53]. The BIF algorithm is based on the HMAX feature extraction method [127], which models the visual processing in the cortex. Specifically, it adopts the first two layers of HMAX, where the first layer convolves facial images with a set of Gabor

filters [41] and the second layer performs maximum (max) pooling over the features extracted by the first layer. The authors improve this bio-inspired method by adding a normalisation operation after max pooling. They find that using only the first two layers of HMAX achieves better results in the age estimation scenario than using the entire HMAX method. Recently, Han *et al.* [56] attach binary decision trees after the feature extraction process performed by the BIF algorithm to predict the age, gender and race simultaneously.

### 2.1.4 Deep Leaning-based Age Estimation

Due to the appearance differences among different images of the same individual, extracting age-specific features and predicting the precise age can be onerous. Due to the extraordinary capability of CNN for feature extraction, [149] first employ a CNN to tackle the age estimation problem. In [149], the authors design a two-layer CNN to extract the age-specific features and use manifold learning algorithms (SVR and SVMs) to compute the final output. Their results show a dramatic improvement on the MORPH II dataset compared to the methods that use traditional machine learning [15, 45, 166].

As aforementioned, recent deep learning-based attempts for age estimation can be classified into two categories. The first category is about improving the accuracy by leveraging customised loss functions rather than using conventional classification loss functions, such as the cross-entropy loss. The second category boosts the estimation performance by modifying the network architecture of a plain CNN model. We first review the recent age estimation works based on these two categories. Then, we discuss some works that involve multi-task learning frameworks to learn age information along with other tasks.

Generally, the age estimation problem can be treated as a multi-class classification problem [116] or a regression problem [114]. Rothe *et al.* [130] propose a formulation that combines regression and classification for this particular task. Since age estimation usually involves a large number of classes (approximately 50 to 100)

and based on the fact that the discretisation error becomes smaller for the regressed signal when the number of classes becomes larger, they compute the final output value by using the following equation:

$$\mathbb{E}(O) = \sum_{i=1}^{n} p_i y_i, \tag{2.3}$$

where $O$ is the output from the final layer of the network after a softmax function, $y_i$ is the discrete year representing the $i$-th class and $n$ indicates the number of classes. Evaluation results demonstrate that this method outperforms both conventional regression and classification in the ChaLearn LAP 2015 apparent age estimation challenge [37] and other benchmarks.

Recent solutions for age estimation have shown that there is an ordinal relationship among ages and leveraged this relationship to design customised loss functions. The ordinal relation indicates that the age of an individual increase as time elapses since ageing is a non-stationary process. Specifically, in [99], the authors construct a label ordinal graph based on a set of quadruplets from training batches and use a hinge loss to force the topology of this graph to remain constant in the feature space. On the other hand, [114] treats the age estimation problem as an ordinal regression problem [91]. The ordinal regression is a type of classification method which transforms the conventional classification into a series of simpler binary classification subproblems. In [114], each binary classification subproblem is used to determine whether the estimated age is younger or elder than a specific age. To this end, the authors replace the final output layer with $n$ binary classifiers, where $n$ equals the number of classes. Let us assume that there are $N$ samples $\{x_i, y_i\}_{i=1}^{N}$, where $x_i$ is the $i$-th input image and $y_i$ is the corresponding age label, and $T$ binary classifiers (tasks). The loss function to optimise the multi-output CNN can then be formulated as [114]:

$$\mathbb{E}_m = -\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \lambda^t 1\{o_i^t = y_i^t\} w_i^t log(p(o_i^t \mid x_i, W^t)), \tag{2.4}$$

where $o_i^t$ indicates the output of the $t$-th binary linear layer, $y_i^t$ indicates the label for the $t$-th task of the $i$-th input, and $w_i^t$ indicates the weight of the $i$-th image for the $t$-th task. Moreover, $W^t$ is the weight parameter for the $t$-th task, and $\lambda^t$ is the importance coefficient of the $t$-th task. Chen *et al.* [22] take a step further by training separate networks for each age group so that each network can learn specific features for the target age group rather than sharing the common features as in [114]. Experiments show that this separate training strategy leads to a significant performance gain on the MORPH II dataset under both evaluation metrics. Li *et al.* [90] also consider the ordinal relation among ages in their work. However, instead of applying the age estimation model on the entire dataset, they take the different ageing pattern of different races and genders into consideration and leverage the domain adaptation methodology to tackle the problem. As stated in their paper, it is difficult to collect and label sufficient images of every population (one particular race or gender) to train the network. Therefore, an age estimation model that is trained on the population with an insufficient number of images would have lower accuracy than models trained on other populations. In their work, they first train an age estimation model under the ranking based formulation on the source population (the population with sufficient images). Then, they fine-tune the pre-trained model on the target population (the population with a limited number of images) by adopting a pairwise loss function to align the age-specific features of the two populations. The loss function used for feature alignment is [90]:

$$\sum_{i=1}^{N^s}\sum_{j=1}^{N^t}\{1 - l_{ij}(\eta - d(\hat{x_i^s}, \hat{x_j^t})) \cdot \omega(y_i^s, y_j^t)\}, \tag{2.5}$$

where $\hat{x_i^s}$ and $\hat{x_j^t}$ are the high-level features extracted from the network, $y_i^s$ and $y_j^t$ are the labels of the images from the source and target populations, respectively. $d(\cdot)$ is the Euclidean distance. $\eta$ and $\omega(\cdot)$ are a predefined threshold value and a weighting function, respectively. $l_{ij}$ is set to 1 if $y_i^s = y_j^t$ or -1 otherwise. The basic idea behind this function is that when the two images have the same age label, the

model tries to minimise [90]:

$$d(\hat{x_i^s}, \hat{x_j^t})) - 1, \tag{2.6}$$

which reduces the Euclidean distance between two features. When the two images have different labels, i.e. $y_i^s \neq y_j^t$, the model tries to minimise [90]:

$$\frac{3}{\omega(y_i^s, y_j^t)} - d(\hat{x_i^s}, \hat{x_j^t})), \tag{2.7}$$

where $\omega(y_i^s, y_j^t)$ is a number smaller than one. This pushes the two features away from each other with a large distance value. In addition, the distance value is proportional to the age difference between the two images.

Another research trend based on customised loss functions is to involve joint loss functions to optimise the age estimation model. Current works that involve joint loss functions include [68] and [117]. [68] studies the problem where the labelled data are not sufficient. In that work, the authors use the Gaussian distributions as the labels rather than specific numbers, which allows the model to learn the similarity between adjacent ages. Since the labels are distributions, they use the KL-divergence to minimise the dissimilarity between the output probability and the label. The KL-divergence can be formulated as:

$$D_{KL}(P \parallel Q) = \mathbb{E}_{x \sim P}[log(P) - log(Q)], \tag{2.8}$$

where $P$ and $Q$ are two distributions. Besides the KL divergence, their model also involves an entropy loss and a cross-entropy loss. The entropy loss is used to make sure the output probability only has one peak since an image can only be associated with one specific age. The cross-entropy loss is used to consider the age difference between images for the non-labelled datasets. Moreover, for the non-labelled datasets, their model accepts two images as input simultaneously. For example, for two images $a$ and $b$, where $a$ is $K$ years younger than $b$, then the age of $a$ should not be larger than $K$. For the image $a$, the authors split the output layer into two parts, the

first part is the neurons with the indices 0 to $K$, and the second part is the neurons with the indices $K$ to $M$, where $M$ is the total number of classes. Based on the aforementioned assumption, the sum of the values in the second part should be 0 while the sum of the values in the first part should be a positive number. The authors treat this problem as a binary classification problem and use the cross-entropy loss to minimise the probability error.

[117] also uses the Gaussian distribution to represent the age label. In addition, it proposes a mean-variance loss to penalise the mean and variance value of the predicted age distribution. The mean-variance loss is used alongside the classification loss to optimise the model, which currently achieves the best performance on the MORPH II dataset and the FG-NET dataset under the MAE metric.

Other worth noting works that also use customised loss function are [100] and [60]. [100] considers both the ordinal relation among ages and the age distribution and involve the metric learning method to cluster the age-specific features in the feature domain. On the other hand, [60] adopts the triplet loss [131] from the conventional face recognition task and uses it for age estimation.

Instead of using plain CNN models (a stack of convolutional layers), some works modify the network architecture to design efficient age estimation models, which is another trending research topic to boost the estimation performance.

Yi *et al.* [162] design a multi-column CNN for age estimation. They take the facial attributes (the eyes, nose, mouth, etc.) into consideration and train several sub-networks for each attribute. All the features extracted from different attributes are then fused before the final layer. [162] is also one of the earliest works that use a CNN for age estimation.

Taheri and Toygar [140] also fuse the information during the learning process. They design a fusion framework to fuse the low-level features, the middle-level features, and the high-level features from a CNN to estimate the age.

Another challenging research area is multi-task learning, which combines age estimation with other facial attribute classification problems or with face recognition.

Multi-task learning is a learning scheme that can learn several tasks simultaneously, which allows the network to learn the correlation among all the tasks and saves training time and computational resources.

Levi and Hassner [89] first design a three-layer CNN to classify both the age and the race. Recently, Hsieh *et al.* [65] design a CNN with ten layers for age estimation, gender classification and face recognition. Results show that this joint learning scheme can boost the performance of all three tasks. Similarly, Ranjan *et al.* [125] propose an all-in-one face analyser which can detect and align faces, detect smiles, and classify age, gender and identity simultaneously. They use a pre-trained network for face recognition and fine-tune it using the target datasets. Authors argue that the network pre-trained for the face recognition task can capture the fine-grained details of the face better than a randomly-initialised one. Each sub-network used for each task is then branched out from the main path based on the level of features on which they depend. Experimental results demonstrate a robust performance on all the tasks.

Lately, Han *et al.* [57] also involve age estimation in a multi-task learning scheme for the face attribute classification problem. Different from the aforementioned works, they group attributes based on their characteristics. For example, since the age is an ordinal attribute, it is grouped with other ordinal attributes like the hair length. Rather than sharing the high-level features among all the attributes, each group of attributes has independent high-level features.

Different from above methods, our FusionNet and ADPF focus on the customised feature extractor by involving dynamically detected age-specific patches. Since the facial images and cropped patches are processed by a different number of convolutional layers, i.e., the length of the learning path varies for different learning sources, the FusionNet in both works involves fusing different levels of features. One work that also fuses different levels of features is [156]. However, the fused features in our work are from various inputs while the fused features in [156] are all from the input facial image.

## 2.2 Age-Oriented Face Synthesis

### 2.2.1 Datasets for Age-Oriented Face Synthesis

Since the age synthesis models also require age information for the training phase, they can also rely on the datasets mentioned in Section 2.1 for training and evaluation. The most broadly used datasets to evaluate age synthesis models are the MORPH II dataset, the CACD and the FG-NET dataset. Typically, the MORPH II dataset and the CACD are used for both training and evaluation, and the FG-NET dataset is only involved in the evaluation phase due to its limited number of samples.

### 2.2.2 Evaluation Metrics for Age-Oriented Face Synthesis Models

Although age synthesis methods have attracted important attention from the research community, several challenges make the synthesis process hard to achieve. First, age synthesis benchmark datasets like the CACD involve other variations like the PIE and occlusion. With these unexpected factors, extracting age-specific features is onerous. Second, existing datasets do not have enough images covering a wide age range for each subject. For example, the MORPH II dataset only captures a time span of 164 days, on average, which may make the learning of long-term personalised ageing and rejuvenating features an unsupervised task. Third, the underlying conditions of the individuals, such as their upbringing environment and genes, make the whole synthesis process a difficult prediction task.

Based on these aforementioned challenges, researchers have established two criteria to measure the quality of synthesised faces. One is the synthesis accuracy, under which synthesised faces are fed into an age classification model to test whether the faces have been transformed into the target age category. Another criterion is the identity permanence, which relies on face verification algorithms to test whether the synthesised face and the original face belong to the same person [160].

### 2.2.3 Traditional Machine Leaning-based Age-Oriented Face Synthesis

The first AOFS methods can be traced back to [106, 107, 143], in which craniofacial growth in young faces is studied. In the early stage, geometry-based methods were a popular choice among researchers, and one of the most representative works is the ASM [29]. The authors model the shape of faces by adjusting the positions of a number of points. Each point marks one part of the face, such as the position of the eyes and the boundary of the face. Synthetic facial images of different shapes and ages can then be obtained by adjusting the position of these points. Another approach to rendering ageing or rejuvenating effects is to directly synthesise or remove wrinkles on a given facial image [8, 102, 112, 153, 154]. Later, Ramanathan and Chellappa [124] propose an ageing-focused method called the craniofacial growth model for synthesising elderly faces by leveraging facial landmark movements. Another worth-noting early AOFS method is [136], where the authors use dictionary learning to learn a personalised ageing process and associate an ageing dictionary to each subject to represent their ageing characteristics.

### 2.2.4 Deep Leaning-based Age-Oriented Face Synthesis

With the increasing popularity of deep learning, several attempts have been made to tackle the AOFS problem using various network architectures. Both Wang *et al.* [151] and Zhang *et al.* [170] use conditional adversarial learning [109] to synthesise aged faces. Wang *et al.* further employ an age category classifier to boost the synthesis accuracy and an L2 constraint on the identity-specific features to preserve the identity information. Yang *et al.* [160] propose a GAN framework by implementing a customised discriminator with a pyramid architecture, which leads to more realistic results than a conventional discriminator as images can be discriminated based on features at multiple scales. They further adopt a pre-trained identity classifier to preserve the identity in the synthesised images. AOFS methods based on the

Wavelet transform are proposed recently in [92, 101], where this transform is used to enhance the texture information in the frequency domain so that richer ageing and rejuvenating effects can be synthesised. He *et al.* [61] implement a GAN model with a customised generator, where a number of decoders are implemented, each one learning an age category. All the decoders are associated with a weight factor to control their relative importance in each transformation. Since all the decoders in the above methods are trained in parallel, the computational complexity of the method is proportional to the number of age categories to be learned. On the contrary, by selecting a particular discriminator from a discriminator pool, our CDP only uses one discriminator for each transformation, which does not increase the computational complexity.

Our work is different from the aforementioned deep learning-based methods as it tackles the AOFS problem from a different angle (i.e., mode learning). Our method can achieve high synthesis accuracy by learning multiple modes explicitly and independently. Additionally, compared to the L1 loss, the L2 loss, and the simple classifiers used in those methods, our AOFS method uses the proposed Adversarial Triplet loss to keep the identity information unaltered in the synthesised facial images.

## 2.3 Age-Invariant Face Recognition

### 2.3.1 Datasets for Age-Invariant Face Recognition

The datasets commonly used for evaluation of AIFR models are the MORPH II dataset and the FG-NET dataset. Moreover, the CACD-VS, which is a noise-free dataset derived from the CACD for cross-age face verification, is also used for AIFR. The CACD-VS contains 2,000 positive cross-age image pairs and 2,000 negative pairs. In addition, researchers also test their AIFR models on the conventional face datasets such as the LFW dataset to demonstrate the generalisation ability of their models.

### 2.3.2 Evaluation Metrics for Age-Invariant Face Recognition Models

Rank-1 accuracy and the mAP are the two widely used evaluation metrics for AIFR models. Given a set of query images $Q$, a set of retrieval results $R_i$ and the number of correct retrieval results $m_i$ for a query image $q_i$, we first define the average precision of $q_i$ as:

$$AP(q_i) = \frac{1}{m_i} \sum_{i=1}^{m_i} Precision(R_i). \tag{2.9}$$

Then, the mAP of $Q$ can be formulated as:

$$mAP(Q) = \frac{1}{Q} \sum_{i=1}^{Q} AP(q_i). \tag{2.10}$$

### 2.3.3 Traditional Machine Leaning-based Age-Invariant Face Recognition

The problem of AIFR has not gained much attention from the research community yet as there are relatively limited works on it compared to works that study other facial variations like pose, illumination and expression. One of the early works is [119], in which the authors used 3D modelling to simulate facial ageing and compensate for the age variations to improve the face recognition performance. In detail, 3D models are built from 2D images, and separate modelling methods are used to generate aged faces. Although they considered both the shape and texture in ageing simulation, the generated faces are not well constructed due to the lack of efficient age estimation algorithm. Later, Li *et al.* [94] defined two general approaches for AIFR. The aforementioned method [119] is categorised as a generative approach, which first synthesises the face that matches the target age and then performs recognition. The other approach is called discriminative approach in which age-invariant features are learned and extracted, and the recognition is based on these features.

Due to the low-quality samples synthesised by early generative models, most existing AIFR methods are discriminative approaches [49, 81, 132, 152, 159]. Studies

on human age [87, 112] show that age information on faces is associated with skin textures, i.e., the texture becomes rough as the age progresses. To this end, early discriminative approaches [94, 138, 158, 159] use the LBP to extract features from facial images and then use techniques like PCA or LDA to perform dimensionality reduction on extracted features. Gong *et al.* [49] model the extracted low-dimensional features as a combination of multiple components, among which one component represents the age information that can be decomposed from the global features before performing the recognition.

### 2.3.4 Deep Leaning-based Age-Invariant Face Recognition

With the increased popularity of the CNNs, researchers have started to use them as features extractors in discriminative AIFR methods. Wen *et al.* [152] are the first using a CNN to tackle the AIFR problem. Instead of directly applying a CNN, the authors designed a customised network with the latent identity analysis that learns disentangled features. Zheng *et al.* [173] proposed a multi-task framework for AIFR with one learning path for the face recognition task and another for the age estimation task. To obtain age-invariant identity features, the authors subtract age features from global features. However, they did not consider the correlation between age features and identity features. Later, Wang *et al.* [150] followed the same multi-task strategy and proposed a novel decomposition method to disentangle the age features from the identity features by using a spherical coordinate system. They also used a regression loss to learn finer age features in order to boost the effectiveness of the decomposition process. The authors further proposed a discriminative method based on adversarial learning and canonical mapping module to reduce the correlation between the age features and the identity features [145]. This adversarial learning-based method demonstrated a superior performance than their previous method on several benchmark datasets.

With the dramatically improved quality of synthesised images, researchers have moved their attention back to the generative approach. Zhao *et al.* [171]

proposed an end-to-end method that can simultaneously synthesise faces at different age groups and performed feature disentanglement. Specifically, the disentanglement is achieved by leveraging a gradient reverse layer [42] that can reverse the gradient of the age information during the back-propagation. Recently, Zhao *et al.* [172] proposed a GAN model for AIFV. The model can synthesise realistic facial images within different age groups by manipulating the latent features between the encoder and the decoder. The verification is then conducted between the input image and the synthesised one.

While all the aforementioned methods are designed for supervised AIFR, unsupervised AIFR is rarely studied. The only worth noting work is [157] in which a pair of auto-encoder is implemented to learn the ageing and de-ageing process simultaneously. However, this method requires image pairs of the same subject as input which may not be applicable in some extreme cases.

## 2.4 Review of Machine Learning Concepts

This section reviews related machine learning concepts that are related to our works. Specifically, we first review various works on the attention mechanism since we use it to discover age-specific patches in age estimation problem. Then, we discuss works that alleviate the mode collapse issue in GANs as we tackle the AOFS from the aspect of mode learning. Additionally, we review works related to the triplet loss. Last, we present a review of contrastive learning.

### 2.4.1 Attention Mechanisms

We used both MHSA and Channel-wise Attention in our work. MHSA is first proposed in [144] and has been widely deployed as the backbone model for various NLP tasks [32]. MHSA can attend to multiple informative segments of the input with an attention head attending to one specific segment. Therefore, the number of segments MHSA can attend to is determined by the number of attention heads.

MHSA has been recently used for imaging data. For example, Zhang *et al.* [165] uses MHSA for the image synthesis task. Specifically, the authors propose the self-attention GAN (SAGAN) by adding MHSA layers to both the generator and the discriminator of a GAN [50]. With the help of MHSA layers, SAGAN can synthesise images with finer details than other state-of-the-art GAN models like [12]. Several recent works [9, 121] also use MHSA for image classification and object detection tasks.

Ever since Zeiler *et al.* [164] visualised the feature maps learned by each channel in each layer of the AlexNet [83] trained on the ImageNet dataset [31], researchers have been exploiting channel-wise attention mechanism to guide the network to pay attention to those channels that learn representative feature maps. Hu *et al.* [66] integrate channel-wise attention into various CNN architectures [58, 64, 137, 139] to boost their performance on image classification and object detection tasks. Similarly, Zhang *et al.* [167] and Chen *et al.* [20] employ channel-wise attention to generate high-resolution images and image captions, respectively. Different from the aforementioned works where channel-wise attention is used to highlight informative channels in the input, in the proposed RMHHA mechanism, we use the computed channel-wise attention weights to merge the multi-channel self-attention maps into a single-channel attention map that reveals a particular age-specific patch.

### 2.4.2 Mode collapse in GANs

The vanilla GAN, which is introduced by Goodfellow *et al.* [50], is capable of generating sharp and realistic images by playing a minimax game between its generator and its discriminator. When training the vanilla GAN, the generator and the discriminator try to reach a Nash equilibrium [108] by minimising the negative log-likelihood loss and minimising the JS-divergence [96]. However, the involvement of the negative log-likelihood loss may cause the discriminator to converge faster than the generator [63]. Once the discriminator finds its global minima, the loss

function stops penalising the generator [17]. This is also known as the vanishing gradient problem [6, 38, 76] and is the main cause of the mode collapse issue. Since the parameters in the discriminator are not further updated, the generator may then find a specific mode that can easily fool the discriminator. When such an issue occurs, the vanilla GAN can only generate limited varieties of samples. Solving this mode collapse issue has become one of the most trending research topics on GANs.

Since the mode collapse issue is caused by the vanishing gradient problem due to the involvement of the negative log-likelihood loss, one strategy to alleviate it is to use an alternative loss function that minimises a different divergence. Nowozin *et al.* [115] first show that the optimisation of GANs is a general process that can be done by minimising any $f$-divergence [30, 95], which is a family of divergences aiming to minimise the distance between two distributions. Some commonly used members of the $f$-divergence family are the JS-divergence, the KL-divergence [84], the squared Hellinger divergence, and the Pearson $\chi^2$ divergence [123]. The authors show that GANs trained with other divergences, like the KL-divergence or the squared Hellinger divergence, can generate images with more variations compared to those generated by the vanilla GAN. Although the work in [115] does not tackle the mode collapse issue directly, it shows the possibility of using other loss functions to optimise GANs.

Arjovsky *et al.* [7] propose the WGAN and use the Wasserstein or EM distance to calculate the distance between distributions of the real and synthesised data. Intuitively, the EM distance computes the cost of transforming one distribution to another, which is more sensitive to the difference between two distributions [7]. Therefore, even if the discriminator is well-trained, it can still keep rejecting the data synthesised by the generator. The LSGAN [104], on the other hand, replaces the negative log-likelihood loss by the L1 loss. Minimising the L1 loss is equivalent to minimising the Pearson $\chi^2$ divergence, which can produce overdispersed approximations and thus makes the LSGAN less mode-seeking [33, 105].

Although the methods discussed before may alleviate the mode collapse issue, their discriminators still have to learn from all the modes. Therefore, recently

proposed methods now focus on modifying the GAN structure. For example, Nguyen *et al.* [113] propose the D2GAN where each discriminator favours data from a different distribution. By using this strategy, their method can compute the KL and reverse KL divergence simultaneously, which in turn increases the variety of samples. Based on this idea, Zhang *et al.* [168] propose a D2GAN variation with two customised discriminators. Specifically, one discriminator consists of residual blocks to form a deep network aiming to increase the variety of generated samples. The other discriminator uses the SELU function [80] as the non-linear activation function. Adopting the SELU function guarantees that this discriminator produces a non-zero value even if the distributions of the synthesised and real data are similar. The authors further propose the D2PGGAN [169] to stabilise the training by leveraging the idea of progressively increasing the complexity of the generator [77]. Durugkar *et al.* [35] propose a GAN with multiple discriminators. Their method may alleviate the mode collapse issue to some extent since the generator has to fool a set of discriminators, which in turn makes the generated samples diverse. It is important to note that by introducing additional discriminators in parallel, the aforementioned methods are also more computationally complex than their plain counterparts (e.g., the vanilla GAN). On the contrary, by selecting a particular discriminator from a discriminator pool, our CDP only uses one discriminator for each transformation, which does not increase the computational complexity.

### 2.4.3 Triplet Loss

The Triplet loss is proposed in [131] aiming to learn feature embeddings for images by optimising the geometric relationship, in the feature space, within a triplet consisting of an *anchor*, a *positive* and a *negative*. Within this context, the *anchor* and *positive* represent feature embeddings of the same class and the *negative* represents a feature embedding of a different class. The goal is to minimise the distance between the *anchor* and the *positive* and simultaneously push the *negative* away from the *anchor*. Since then, a number of variations to this loss have been proposed. For instance,

Chen *et al.* [24] uses an additional *negative* embedding alongside the original triplet to form a quadruplet. Huang *et al.* [69] implement three ranking operations in total by using an *anchor*, a *negative* and three *positives*. Ye *et al.* [161], on the other hand, adopt additional images from other modalities. It is worth noting that all these variants leverage additional samples either within the same or from another modality. Therefore, these losses can no longer help to optimise the geometric relationship within a triplet.

### 2.4.4 Contrastive Learning

The first related work on contrastive learning can be traced back to [55] which learns robust feature representations by contrasting positive pairs against negative pairs. Dosovitskiy *et al.* [34] then used a similar strategy to train a CNN for an object recognition task by discriminating samples generated by different augmentation processes. Later, Wu *et al.* [155] replaced the linear classifier in [34] with a memory bank to store representations for each class and used the noise contrastive estimation to compare samples. The memory bank has been widely used in recent works [110, 142]. He *et al.* [59] explored the contrastive learning from a different perspective where feature representations are produced by a momentum encoder rather than a pre-trained CNN. Most recently, Chen *et al.* [23] demonstrated that the aforementioned contrastive learning methods can be simplified as long as the batch size is large enough.

## 2.5 Concluding Remarks

In this chapter, we briefly reviewed the datasets, evaluation metrics, and related works on the problem basis. Related works cover both traditional machine learning-based methods and deep learning-based methods. At the end of this chapter, we reviewed several machine learning concepts used in our methods and state-of-the-art works related to these concepts.

# Chapter 3

# FusionNet for Age Estimation

## 3.1 Introduction

Age estimation is an active research topic, which is intended to predict the age of a subject based on the appearance of his or her face. Recently, CNNs have been proved to be capable of dramatically boosting the performance of many mainstream computer vision problems [58, 71, 120].

Neuroscience shows that when the primate brain is processing the facial information, different neurons respond to different facial features [16]. Inspired by this fact, we intuitively assume that the accuracy of age estimation may be largely improved if the CNN could learn from age-specific patches. Consequently, in this chapter, we propose the FusionNet, a novel CNN architecture for face-based age estimation. Specifically, FusionNets take the face and several age-specific facial patches as successive inputs. This data feeding sequence is shown in Figure 3.1. As illustrated in the figure, there are a total of $n + 1$ inputs (one face and $n$ facial patches) being fed into the network. The aligned face, which provides most of the information, is the primary input that is fed to the lowest layer to have the longest learning path. After all the inputs are fed into the network, the final prediction is calculated based on this fused information that is learned through the convolutional layers. We show later that the input at the middle-level layers can be viewed as

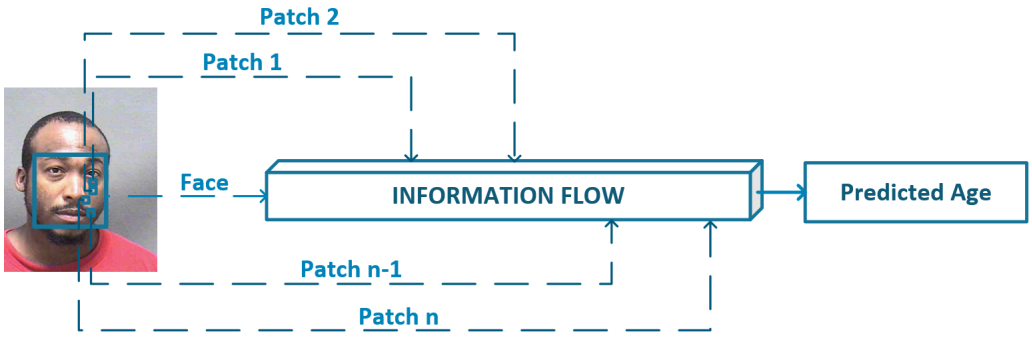


Figure 3.1: Data feeding sequence in the FusionNet. The model takes the original face and a total of $n$ facial patches as inputs.

shortcut connections that boost the flow of the age-specific features.

Unlike previous multi-patch-based work [162] which use major facial attributes (e.g. the eyes and the mouth) as input patches, our network takes adaptively-selected features as the secondary learning source. Different from those dominating attributes which may introduce certain and sophisticated patterns that cannot be learned together with the original face, the selected patches in our case are mainly those regions representing smooth facial skin with aged textures. Our results demonstrate that these textures can be used to complement the features learned from the whole face to emphasise the age-specific patterns.

## 3.2 FusionNet

The proposed method consists of three components, the facial patch selection, the convolutional network and the age regression. The facial patch selector is based on the BIF [53] and the AdaBoost algorithm. Selected patches are subsequently fed into the convolutional network, in a sequential manner, together with the face. The final prediction is calculated based on the output of the network by using a regression method.

### 3.2.1 Facial Patch Selection

We use the BIF [53] to extract age-specific feature from aligned faces. Faces are convolved with a bank of Gabor filters [41], which can be formulated as:

$$G(x, y) = \exp(-\frac{(x'^2 + \gamma^2 y'^2)}{2\sigma^2}) \times \cos(2\pi \frac{x'}{\lambda}) \tag{3.1}$$

where $(x, y)$ are the spatial coordinates, and $x' = x\cos\theta + y\sin\theta$ and $y' = -x\sin\theta + y\cos\theta$ denote the orientation of the filters with the angle $\theta \in [0, \pi]$. $\gamma$, $\sigma$, and $\lambda$ are the parameters of the filters. We convolve each face with a total of 8 bands and 8 orientations of Gabor filters to generate a $k$-dimensional feature vector to detect textures in different sizes and orientations with minimum redundancy. In our experiments, $k$ is greater than 10,000 with each element encoding one potential input for the subsequent CNN. Since we cannot use this high-dimensional feature vector in the feeding sequence directly, we need to select $k'$ features from the BIF feature vector to form a subset where $k' << k$. We experimentally set $k'$ to 1000 and use the top 5 most informative features as the input to the subsequent network to keep a balance between the training time and the performance. We observe that 5 features have a good coverage of age-specific regions, and including more features can lead to redundancy. The top 5 selected features are represented as the 5 patches marked in the face in Figure 3.2.

The multi-class AdaBoost is used to select the subset $k'$ from the high-dimensional feature vector. A Decision Tree is built as the weak classifier in AdaBoost, which is similar to the implementation in [56]. Briefly, for a dataset with $m$ samples, we pick the $k'$ most informative features from a $k$-dimensional vector by using the weak classifier $h$,

$$\mathcal{F}_j = argmin_k(\sum_{i=1}^{m} w_i^{k'} e(h_k(x_i), y_i)) \tag{3.2}$$

where $\mathcal{F}_j$ is the $j$-th selected feature and $j \in [1, k']$. $x_i$ is the high dimensional feature vector after the $i$-th sample is filtered by Gabor filters and $y_i$ is the associated age

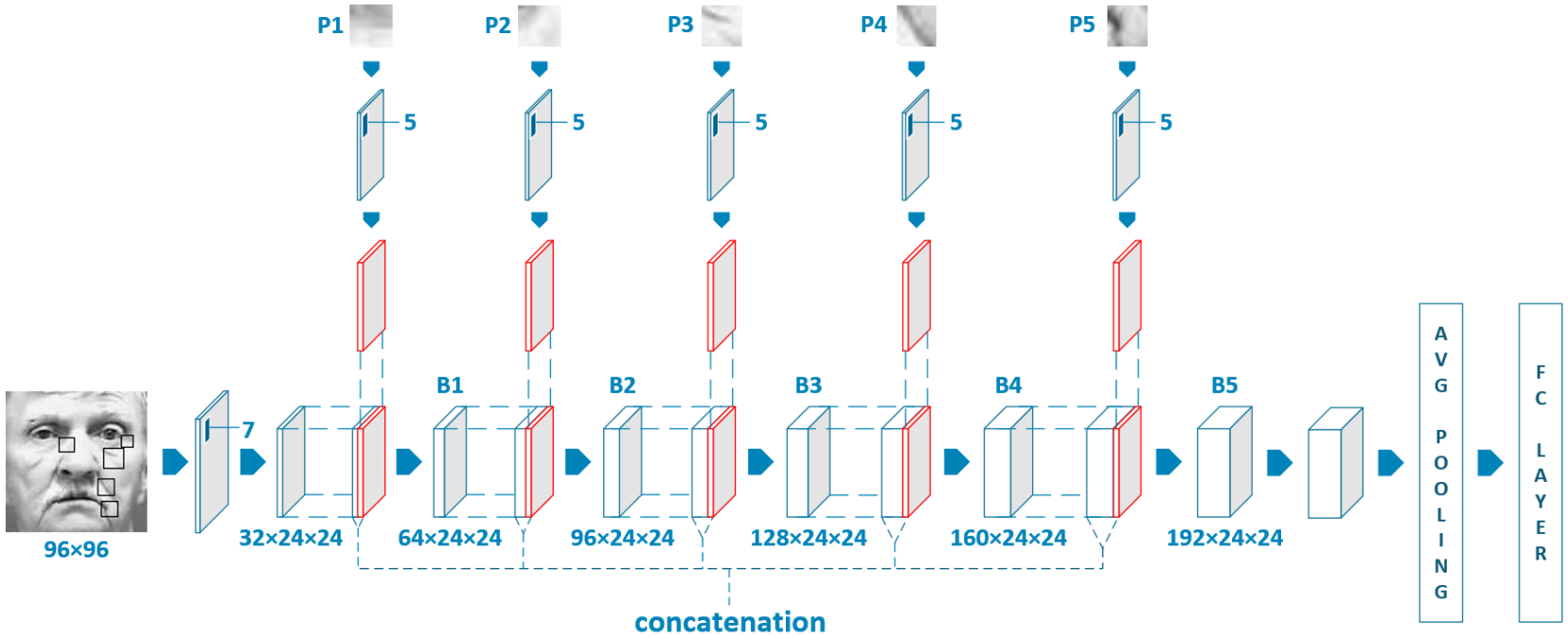


Figure 3.2: The architecture of the Fusion Network for face-based age estimation. The selected patches are fed to the network sequentially as the secondary learning source. The input of patches can be viewed as shortcut connections to enhance the learning of age-specific feature. We use five patches (P1 to P5) to keep the balance between the training efficiency and the performance. The final output is produced by a single FC layer.

label. In addition, $w_i^{k'}$ is the weight in AdaBoost, which is updated and normalised after each $\mathcal{F}_j$ is found. The error function $e(h_k(x_i), y_i)$ in Eq. (3.2) is defined as follows:

$$e(h_k(x_i), y_i) = \begin{cases} 0 & h_k(x_i) = y_i \\ 1 & otherwise \end{cases} \tag{3.3}$$

We find that a 28-level Decision Tree can be implemented as the weak classifier in our case to give us a good classification performance while keeping the training time manageable.

### 3.2.2 Network Architecture

The architecture of the FusionNet is illustrated in Figure 3.2. In the figure, the block arrows indicate the feature extraction process and the dashed lines between blocks denote copying. All of the blocks shown in Figure 3.2 are residual blocks [58], and each block after concatenation (B1 to B5) contains bottleneck layers. Note that we do not apply feature reduction to B5 in Figure 3.2, since we have found that lowering the number of feature maps right before the global pooling largely reduces the performance. Moreover, we apply a batch normalization layer [72] before each convolutional layer to improve the training speed and overall accuracy. After the convolutional stage, a global average pooling layer and a FC layer are attached to generate the final output of the network.

Instead of training separate shallow CNNs for each input and concatenating the information before the final fully-connected layer, we merge the features in the convolution stage. In the FusionNet, all the features from different inputs have a longer and more efficient learning path compared to the multi-path CNN in [162]. Moreover, the common age-specific features among the inputs can be extracted and emphasised. For example, the skin feature, which has ordinal relationship to the age, can be enhanced since all the simultaneous inputs share almost the same skin texture.

The use of concatenation is inspired by the DenseNet [71]. In a DenseNet, the network is divided into several dense blocks, and layers within the same block typically share the identical spatial dimension. More importantly, inside each dense block, the output of each layer flows directly into all of the subsequent layers. As a result, the $l$-th layer receives feature maps from all the previous layers within the same block as the input [71]:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \tag{3.4}$$

where $x$ represents the output of each layer and $H_l$ denotes the learning hypothesis of the $l$-th layer. $[\cdot]$ is used to represent the concatenation operation.

In the FusionNet, the formulation is based on blocks, and the output of each residual block after concatenations can be represented as:

$$x_i = B_i([x_{i-1}, s_i]) \tag{3.5}$$

where $B_i[\cdot]$ denotes the synthesised learning function of the $i$-th block and $i \in [1, 5]$ since we decide to use 5 input patches in our network. Therefore, the shortcut connections in FusionNet are block-wise operations rather than layer-wise operations as in [71]. In addition, $x_{i-1}$ is the output from the previous residual block and $s_i$ is the feature map learned from the $i$-th input patch. Since the patches share common features with the original face, and based on Eq. (3.4) and (3.5), the incoming patches can be viewed as shortcut connections that refresh and amplify the flow of age-specific information.

### 3.2.3 Age Regression

Based on the fact that the discretization error becomes smaller for the regressed signal when the number of classes becomes larger [128], we calculate the final prediction through a regression approach.

After the features are processed by the fully-connected layer, we first eliminate all the negative values in the output vector and feed it to a Softmax function to form a probability distribution. Then, we normalise the distribution to make it sum up to 1.

The final prediction is the summation of products of the probabilities by the corresponding age labels.

$$\mathbf{E}(O) = \sum_{i=1}^{j} p_i y_i \tag{3.6}$$

where $p_i$ denotes the normalized probability for the $i$-th class, $y_i$ is the associated age label, and $j$ is the number of classes.

## 3.3 Experiments

### 3.3.1 Experimental Settings

We use the most frequently used MORPH II benchmark [126] for age estimation to test the performance of our network. Following the previous works [19, 114, 128], in this work, the dataset is randomly divided into two parts, about 80% for training and the other 20% for testing. There is no overlap between the training and testing sets. To perform statistical analysis and follow previous works [22, 114, 128], we use 20 different partitions (with same ratio but different distribution) in the experiment and report the mean values.

We use the open-source computer vision library dlib [78] for the image preprocessing in our work. All the faces are cropped to $96 \times 96$ pixels and converted to gray-scale images since the MORPH II dataset suffers from the colour cast issue. After the facial patches are selected, the cropped patches are then resized to $24 \times 24$ pixels.

The proposed network is implemented based on the open-source deep learning framework Pytorch and trained with the SGD algorithm with momentum. The batch size is set to 64. We train our network for 200 epochs with an initial learning rate of

0.1. The learning rate drops by a factor of 0.1 after every 50 epochs.

### 3.3.2 Results

To demonstrate the efficiency of our proposed network, we use the CS criteria to evaluate the performance of the FusionNet compared with a baseline model, which is a plain network with all selected patches removed. In Table 3.1, the model in the second row represents a FusionNet taking major facial attributes like the eyes, the nose and the mouth as secondary inputs and using classification method to calculate the predicted age. The model in third row uses age-specific patches and the model in the last row uses regression to produce the final age. The reason why the second row (FusionNet + FAttrs + Cls) performs worse compared to the baseline may due to that major facial attributes carry identity-specific details rather than age-specific features, which could be treated as noise during training and degrade the performance.

We compare our approach with other recent state-of-the-art CNN-based models: DEX [128], OR-CNN [114], and Ranking-CNN [21]. To have a fair comparison, only works with the same data partition ratio are evaluated. In [128], authors use a pre-trained VGG-16 [137] as the core model and further fine-tune it on the IMDB-WIKI dataset [128]. In the comparison, we use the result without fine-tuning on the additional dataset. As shown in Table 3.2, the FusionNet achieves the lowest MAE of 2.82, which outperforms other state-of-the-art models. This result shows that our network has a much more efficient feature extraction architecture. Moreover, the modern network design philosophy used (i.e., the residual blocks and bottleneck layers) helps to improve the performance even further.

## 3.4 Conclusion

In this chapter, we presented the FusionNet to tackle the face-based age estimation problem. Our model takes not only the face but also other age-specific facial

Table 3.1: Comparison between FusionNet and a baseline model. The best result is highlighted in **bold**.

| Method | CS(n=1) | CS(n=2) | CS(n=3) | CS(n=4) | CS(n=5) | CS(n=6) | CS(n=7) | CS(n=8) |
|---|---|---|---|---|---|---|---|---|
| baseline | 30.06% | 51.07% | 63.51% | 74.00% | 82.70% | 88.07% | 92.45% | 95.04% |
| FusionNet + FAttrs + Cls | 29.94% | 50.51% | 63.02% | 73.26% | 82.02% | 87.50% | 91.94% | 94.96% |
| FusionNet + AdaP + Cls | **31.22%** | 51.72% | 67.24% | 78.40% | 85.26% | 90.55% | 93.57% | 96.01% |
| FusionNet + AdaP + Reg | 30.96% | **53.07%** | **68.35%** | **79.59%** | **86.16%** | **91.00%** | **93.97%** | **96.37%** |

Table 3.2: MAE values of three state-of-the-art CNN-based models and our method on MORPH II dataset. The best result is highlighted in **bold**.

| Method | MAE |
|---|---|
| OR-CNN [114] | 3.27 |
| DEX [128] | 3.25 |
| Ranking-CNN [22] | 2.96 |
| baseline | 3.05 |
| FusionNet + FAttrs + Cls | 3.18 |
| FusionNet + AdaP + Cls | 2.95 |
| FusionNet + AdaP + Reg | **2.82** |

patches as inputs. The input facial patches can be considered as being shortcut connections in the network, which amplify the learning efficiency for age-specific features. Experiments show that our network significantly outperforms other CNN-based state-of-the-art methods on the MORPH II benchmark.

However, we find that the proposed method takes too much time to train (normally a few days). This is mainly caused by the involvement of the BIF and Adaboost algorithm. To reduce the training complexity, we further propose a modified method based on attention mechanisms. The details of this modified method is presented in the next chapter.

# Chapter 4

# Improving Age Estimation with Attention-Based Dynamic Patch Fusion

## 4.1 Introduction

Modern face-based age estimation methods typically consist of two components, a feature extractor and an estimator. The feature extractor is used to extract age-specific features from raw facial images and the estimator is used to predict the age based on the extracted features. Many recent works [14, 21, 39, 67, 93, 97, 98, 114, 117, 134] focus on designing customised estimators while treating the facial image as an ordinary input, hence paying no attention to the relative importance of the extracted features. However, related studies [53, 56, 146] show that age-specific patches are useful when predicting the age of the subject from an image. In other words, customised feature extractors can be designed to exploit age-specific patches during training to boost the performance of face-based age estimation methods. As a consequence, many works now tackle the face-based age estimation problem by leveraging cropped age-specific patches as complementary inputs to their estimator

[4, 26, 56, 146, 162]. The patches used in most of these works are those depicting dominant facial attributes like the eyes, nose, and mouth. However, early studies on face-based age estimation [2, 11, 48, 86, 87, 112, 153] show that the most informative patches for this problem are where wrinkles typically appear, like eye bags and laugh lines. To locate these age-specific patches, Han *et al.* [56] leverage the BIF proposed in [53]. Later, Wang *et al.* [146] design a customised CNN to fuse the features learned from the facial image and the BIF-based patches. Unfortunately, the computed BIF-based patches in these methods are fixed in every image, which prevents extracting features that are robust to the location and shape variations of age-specific regions.

In this chapter, we propose a novel framework named ADPF based on our preliminary work [146] to tackle the face-based age estimation problem. ADPF comprises a customised feature extractor that consists of an AttentionNet and a FusionNet. The AttentionNet dynamically discovers age-specific patches by employing a novel attention mechanism, while the FusionNet predicts the age of the subject by fusing features learned from the facial image and the discovered age-specific patches. To improve performance, the discovered patches are fed into the FusionNet sequentially in a descending order based on the amount of age-specific information they carry. To this end, we introduce the RMHHA mechanism into the AttentionNet. RMHHA is inspired by the MHSA mechanism [144]. However, instead of using the multi-channel feature maps produced by MHSA, each attention head in RMHHA yields a compact single-channel attention map, which is used to crop the corresponding age-specific patch from the facial image. RMHHA assigns a learnable weight to the produced attention maps to rank their importance. Hence, RMHHA not only helps to dynamically learn age-specific patches, but it also ensures the discovered patches are fed into the FusionNet in the desired order.

## 4.2 Attention-based Dynamic Patch Fusion

In this section, we explain in detail ADPF by first discussing the core of the AttentionNet, i.e., the proposed RMHHA mechanism. Then, we formulate the diversity loss followed by explaining the FusionNet used to fuse features from various learning sources. The architecture of ADPF is illustrated in Fig. 4.1.

### 4.2.1 Ranking-guided Multi-Head Hybrid Attention

Since RMHHA is based on MHSA and the key component in MHSA is the self-attention mechanism, we first discuss the self-attention mechanism followed by the proposed hybrid attention mechanism. Then, we detail the complete RMHHA mechanism.

Let us consider an input tensor $X$, as shown in Fig. 4.1, that has a dimension of $H \times W \times C$, where $H$ denotes the height, $W$ denotes the width and the $C$ denotes the number of channels. $X$ is convolved into three separate tensors: $Q$ with a shape of $H \times W \times C_Q$, $K$ with a shape of $H \times W \times C_K$, and $V$ with a shape of $H \times W \times C_V$, where $C_Q$, $C_K$, and $C_V$ indicate the number of channels in the corresponding tensor. The intuition behind self-attention is to compute a weighted summation of the values, $V$, where the weights are computed as the similarities between the query, $Q$, and the corresponding key, $K$. Therefore, in order to compute the similarity, $Q$ and $K$ normally have the same shape, i.e., $C_Q = C_K$. The output of a single self-attention mechanism is computed as [144]:

$$SA = Softmax(\frac{Q' \cdot K'^T}{\sqrt{C_K}}) \cdot V, \tag{4.1}$$

where $Q'$ and $K'$ are flattened tensors in order to perform the dot product.

After the scaling operation, i.e., dividing the similarity matrix $Q' \cdot K'^T$ by a factor of $\sqrt{C_K}$ and applying the softmax function, we perform a dot product between the normalized similarity matrix and $V$ to generate the self-attention maps $SA$ with a dimension of $H \times W \times C_V$.

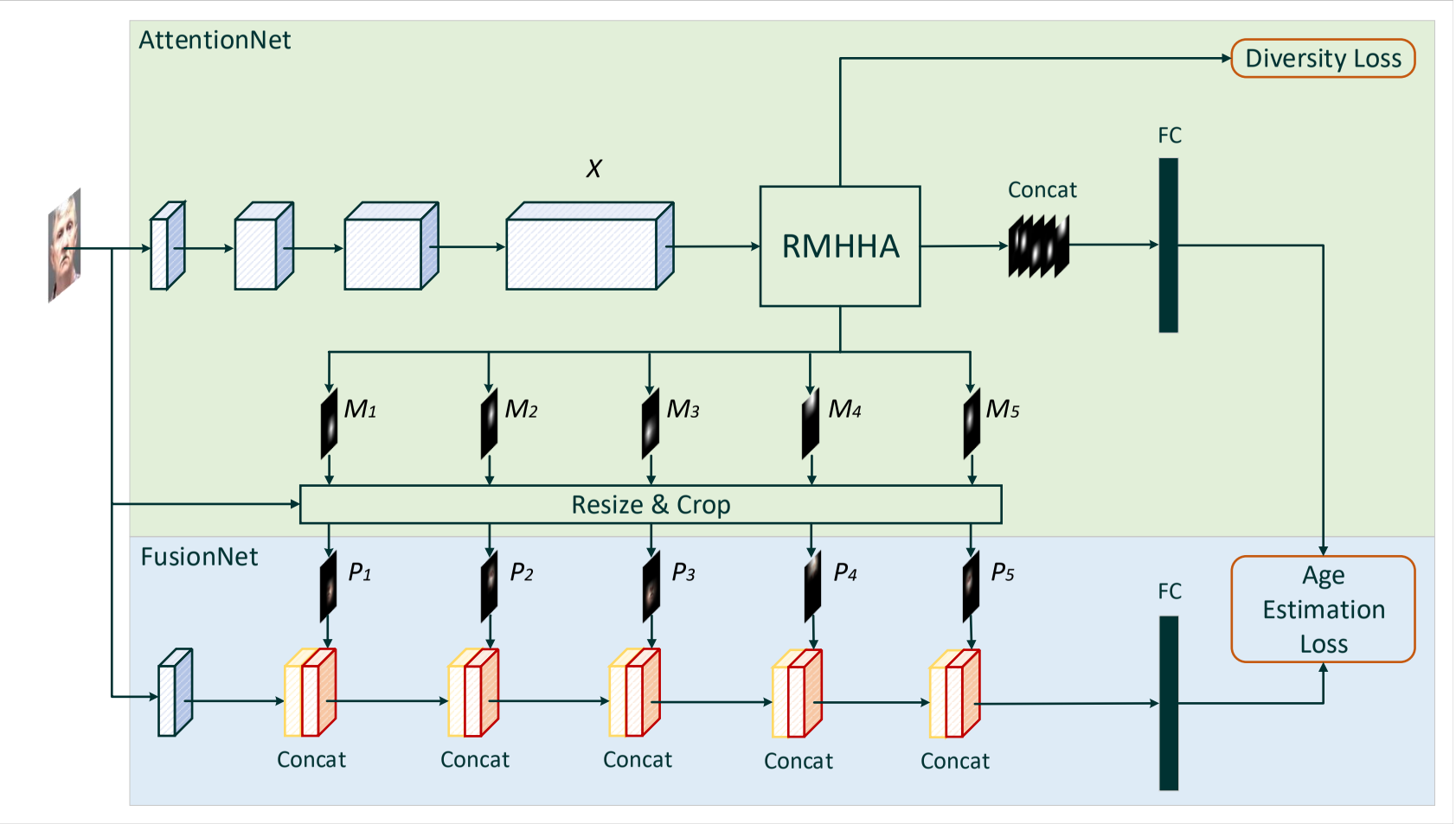


Figure 4.1: Architecture of ADPF. It consists of two networks, the AttentionNet and the FusionNet. The AttentionNet is used to train the proposed RMHHA to learn and rank age-specific features. Once the features are learned and ranked, denoted as $M1$ to $M5$ in the figure, we resize them to crop the corresponding patches from the input facial image. The cropped patches are listed as $P1$ to $P5$ in a descending order based on the amount of age-specific information they carry. Blocks represents CNN layers and *Concat* indicates concatenation operations. In particular, yellows blocks are from the previous layer in the main stream and red ones are from one particular age-specific patch. In addition, $X$ is the input tensor to the RMHHA mechanism.

Since we flatten two-dimensional feature maps into an one-dimensional vector in Eq. (4.1), the original structure of the feature maps is therefore distorted. To make it efficient when dealing with structured data like images and multi-dimensional features, we adopt the relative positional encoding in [133] and [9]. Specifically, the relative positional encoding is represented by the attention logit, which encodes how much an entry in $Q'$ attends to an entry in $K'$. The attention logit is computed as [9]:

$$l_{i,j} = \frac{q_i^T}{\sqrt{C_K}}(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y}), \tag{4.2}$$

where $q_i$ is the $i$-th row in $Q'$ indicating the feature vector for pixel $i := (i_x, i_y)$ and $k_j$ is the $j$-th row in $K'$ indicating the feature vector for pixel $j := (j_x, j_y)$. $r^W_{j_x - i_x}$ and $r^H_{j_y - i_y}$ are learnable parameters encoding the positional information within the relative width $j_x - i_x$ and relative height $j_y - i_y$. With the relative positional encoding, the output of a single self-attention mechanism can be reformulated as [9]:

$$SA = Softmax(\frac{Q' \cdot K'^T + M_H + M_W}{\sqrt{C_K}}) \cdot V, \tag{4.3}$$

where $M_H[i,j] = q_i^T r^H_{j_y - i_y}$ and $M_W[i,j] = q_i^T r^W_{j_x - i_x}$ are matrices of relative positional logits.

The output of the self-attention mechanism in Eq. (4.3) has a dimension of $H \times W \times C_V$. However, we want each attention head to produce a single-channel attention map to depict one particular age-specific patch. To this end, we use channel-wise attention alongside self-attention to form a hybrid attention mechanism. Channel-wise attention is used to compute weights for each channel and a weighted summation is performed along the channel axis of the self-attention maps to generate the final single-channel attention map, indicated as the hybrid attention map in Fig. 4.2.

As depicted in Fig. 4.2, in the proposed hybrid attention mechanism, we first use a 1x1 convolutional layer on the input tensor, $Z$, to ensure the number of channels before computing the channel-wise attention weights matches the number of

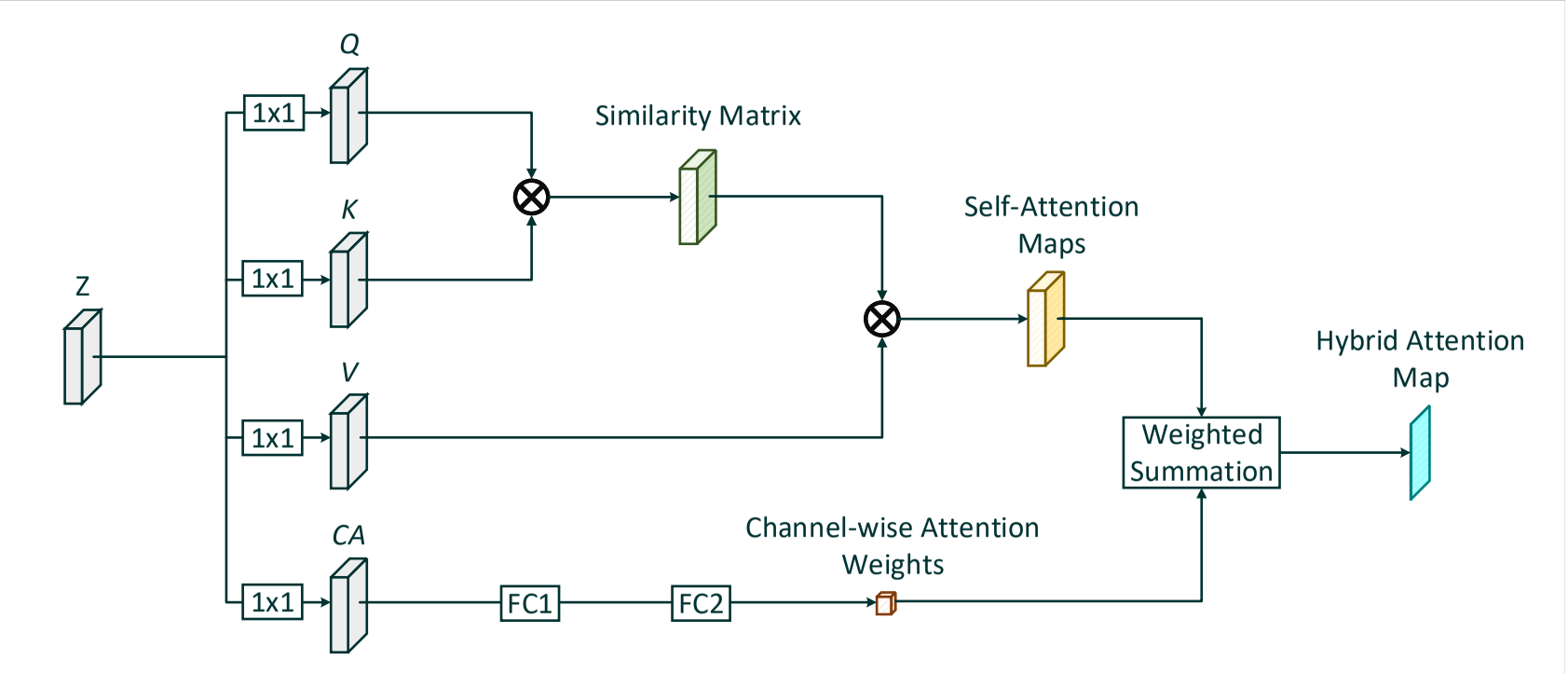


Figure 4.2: Structure of the proposed hybrid attention mechanism. $Q$, $K$, and $V$ are the *query*, *keys*, and *value*, respectively, for the self-attention mechanism, and $CA$ is the input tensor to the channel-wise attention mechanism. The final hybrid attention map is computed as weighted summation, where the input tensor comprises the attention maps from the self-attention mechanism and the weights are computed from the channel-wise attention mechanism. 1x1 represents convolutional layers with kernel size of 1 and FC1 and FC2 indicate two fully-connected layers.

channels in the self-attention maps, i.e., $C_V$. The tensor after this 1x1 convolution is denoted as $CA$. We then aggregate each feature map in $CA$ with a pooling operation to produce a feature vector, in which each entry represents the features for the corresponding channel. Different from [21, 66], in which average pooling is used, we use max pooling as we want to emphasise the most important features with high activation values. Following the procedure in [66], we use a gating mechanism with two sequential FC layers to form a bottleneck. The first FC layer reduces the dimentionality, i.e., the number of channels, and the second FC layer increases the dimentionality of the previous layer to match the original shape. The output from the second FC layer is the set of channel-wise attention weights that we need, which are computed as:

$$W^{CA} = \sigma(W^{FC2}\delta(W^{FC1}\delta(CA))), \tag{4.4}$$

where $\delta$ indicates the non-linear ReLU function, $\sigma$ refers to the Sigmoid function used to normalise the attention weights, and $W^{FC1}$ and $W^{FC2}$ are learnable parameters in the two FC layers.

After the self-attention maps and channel-wise attention weights are computed, we perform a weighted summation over these two tensors along the channel dimension to get the single-channel hybrid attention map. The hybrid attention map is then computed as:

$$HA = \sum_{c}^{C_V} SA_c W_c^{CA}, \tag{4.5}$$

where $c$ is the channel index and $SA$ is computed using Eq. (4.3).

To perform hybrid attention in a multi-head manner, each hybrid attention head takes a certain number of feature maps from the previous convolutional layer as the input. Specifically, assume there are $C_P$ feature maps in the tensor produced by the previous layer. Then, we have $C_P = C_{HEAD} \times N$, where $N$ denotes the number of heads.

Different from MHSA [144], in which the attention maps from each head are concatenated right after the attention operation, we assign a learnable scale to each

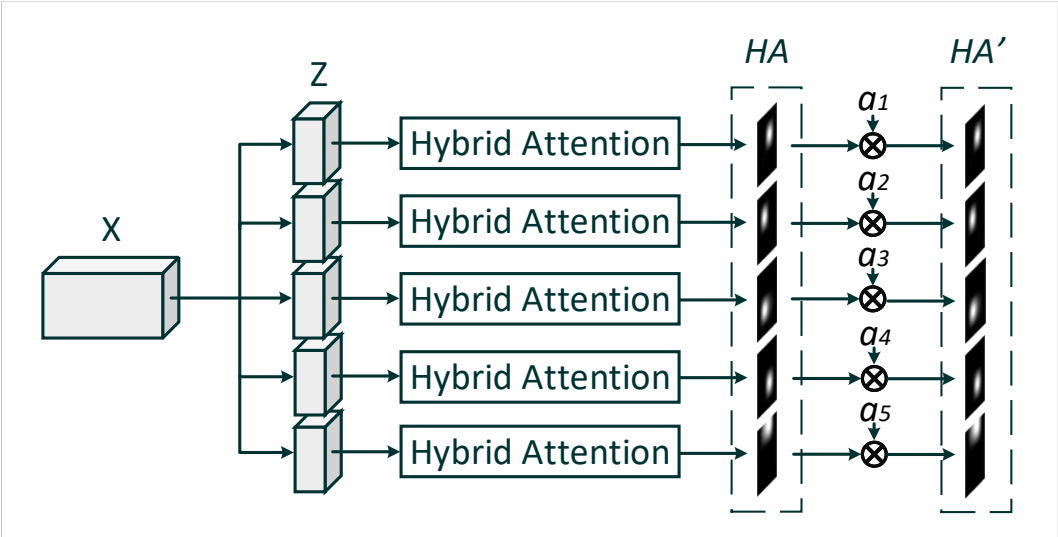


Figure 4.3: Architecture of the proposed RMHHA, where five attention heads are implemented.

hybrid attention map to rank their importance when predicting ages, as shown in Fig. 4.3. RMHHA can then be formulated as:

$$RMHHA = [HA_1 a_1, HA_2 a_2, ..., HA_n a_n], \tag{4.6}$$

where $a_n$ indicates the learnable scale, which is updated by using the age estimation loss function presented in section 4.2.4 and $n \in [1, N]$. $HA_n a_n$ is equivalent to $HA'_n$ in Fig. 4.3. All weighted hybrid attention maps used in ADPF are then concatenated before the final FC layer in the AttentionNet.

It is worth noting that multi-head attention methods always involve heavy matrix multiplications, which may be computationally expensive especially when the input matrices have a high dimentionality, which is common in CNNs. Therefore, differently from [9, 121], which stack dozens of MHSA models to compute the output, our work only uses one multi-head attention model to discover age-specific patches.

### 4.2.2 Diversity Loss

The number of patches that can be discovered is determined by the number of attention heads implemented in RMHHA. However, during implementation, we find that when using more than four heads, patches tend to overlap with each other especially in informative regions. As demonstrated in Section IV, without further supervision, two attention maps may overlap in the nose region. This overlap of attended patches may led to redundant learning sources and leave other age-specific patches undiscovered. To alleviate this overlap issue, we propose a diversity loss to learn diverse and non-overlapping patches by minimizing the summation of product of corresponding entries in two hybrid attention maps, $HA_m$ and $HA_n$. The diversity loss is formulated as:

$$\mathcal{L}_{diversity} = \sum_{\substack{m,n \\ m \neq n}}^{N} \sum_{h}^{H} \sum_{w}^{W} HA_m(h,w) HA_n(h,w), \tag{4.7}$$

where $(h, w)$ denotes the location of the corresponding entry in a hybrid attention map.

### 4.2.3 FusionNet

The architecture of the FusionNet is illustrated in Fig. 4.2. To get the input patches, i.e., $P_1$ to $P_5$, we first rank the learned hybrid attention maps based on their associated weights, i.e., $a_1$ to $a_5$. $M_1$ has the highest weight indicating that the corresponding age-specific patch represents the most age-specific information. After the hybrid attention maps are ranked, they are resized into the same spatial size as the original facial image and used to crop the corresponding highlighted area by keeping all the pixels where the activation values in the resized feature maps are non-zero.

Instead of training separate shallow CNNs for each input and concatenating the information before the final FC layer, we merge the features in the convolution

stage. In the FusionNet, the length of the path to learn from an input is directly proportional to the amount of information it carries. This approach also allows extracting and emphasising common age-specific features among all inputs. For example, the skin feature, which has an ordinal relationship with the age, can be emphasised since all inputs are expected to share the same skin texture.

In the FusionNet, we preform concatenation operations on pairs of feature maps, one from the previous layer in the main stream (yellow blocks in Fig. 4.1), $I$ , and the other representing the features learned from one particular age-specific patch (red blocks in Fig. 4.1), $P$. Therefore, the concatenation in the FusionNet is formulated as:

$$R = Concate[I, P]. \tag{4.8}$$

This formulation is also commonly used in modern CNN architectures like the ResNet [58] and the DenseNet [71]. Therefore, a sub-stream in the FusionNet can be treated as a shortcut connection, which emphasises the learning of the age-specific information shared by all inputs.

### 4.2.4 Age Estimation Loss

To estimate the age, we use a commonly used method that combines a regression loss to learn the exact age and a divergence loss to learn the age distribution. Specifically, after the features are processed by a Softmax function, we eliminate all the negative values in the output vector and normalise the remaining values so that they can form a probability distribution that sums up to 1:

$$o_p := \begin{cases} 0 & o_t \leq 0 \\ \frac{\sum_{p=1}^{q} \max(0, o_p)}{o_p} & o_t > 0, \end{cases} \tag{4.9}$$

where $o_p$ is the $p$-th element in the output vector $O \in \mathbb{R}^q$ and $q$ is the total number of classes.

The final prediction is the summation of products of the probabilities by the corresponding age labels:

$$E = \sum_{p=1}^{q} o_p g_p, \tag{4.10}$$

where $o_p$ denotes the normalized probability from Eq. (4.9) and $g_p$ is the associated age label for class $p$.

We use the MAE to compute the error between the prediction and the corresponding ground truth label:

$$\mathcal{L}_{MAE} = \frac{1}{B} \sum_{b}^{B} |E_b - GT_b|, \tag{4.11}$$

where $B$ is the batch size and $GT$ refers to the ground truth label.

To learn the age distribution, we use the KL-divergence to measure the difference between a Gaussian distribution derived from the label [117] and the learned distribution. The KL-divergence is formulated as:

$$\mathcal{L}_{KL} = \sum_{p=1}^{q} P(p) log \left( \frac{P(p)}{P'(p)} \right), \tag{4.12}$$

where $P$ is the ground truth distribution and $P'$ is the learned distribution. The complete age estimation loss is then defined as a summation of these two losses:

$$\mathcal{L}_{AE} = \mathcal{L}_{MAE} + \mathcal{L}_{KL}. \tag{4.13}$$

### 4.2.5 Training Strategy

Since the training of the FusionNet requires well-learned and stabilised patches, we first train the AttentionNet with RMHHA until convergence. The overall loss to train this network is the summation of two loss functions:

$$\mathcal{L}_{AttentionNet} = \mathcal{L}_{AE} + \lambda \mathcal{L}_{diversity}, \tag{4.14}$$

where $\lambda$ controls the relative importance between two learning objectives.

When the AttentionNet converges, we freeze its parameters and start training the FusionNet. The loss function used to train the FusionNet is the loss formulated in Eq. (4.13).

## 4.3 Experiments

### 4.3.1 Experimental Settings

**Data Pre-processing**. We use the open-source computer vision library dlib [78] for image pre-processing. Firstly, 68 facial points are detected in each facial image to crop them based on the location of the eyes to a size of $128 \times 128$ pixels.

Further, data augmentation is used to increase the dataset size. Specifically, images are zero-padded first and then cropped to the original size. Finally, the cropped images are randomly flipped horizontally.

**Dataset Partition**. We conduct experiments on three commonly used face-based age estimation benchmark datasets, the MORPH II dataset [126], the FG-NET dataset [28], and the CACD [18]. For the MORPH II dataset, three commonly used settings are adopted. In the first setting, i.e., *Setting I*, following prior works [21, 98, 114, 117, 141, 146, 156], we randomly split the whole dataset into two subsets, one with 80% of the data for training and the other with 20% for testing. In this setting, there is no identity overlap between the two subsets. To perform statistical analysis, we use 20 different partitions (with the same ratio but different distribution) and report mean values. In the second setting, i.e., the *Setting II*, to compensate for the imbalance of race distribution, we randomly split the dataset into three subsets, denoted as *S1*, *S2*, and *S3*, and ensure the ratio between Black and White labels is 1:1 and that between Male and Female labels is 1:3. In order to follow the same protocol as other works [25, 26, 51, 93, 162], the results under this setting are reported in three different ways: 1) training on *S1* and testing on *S2+S3*; 2) training on *S2* and testing on *S1+S3* and 3) the average value from the previous two scenarios. Finally,

Table 4.1: MAE values for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting I.

| Method | MAE |
|---|---|
| OR-CNN [114] | 3.27 |
| DEX [130] | 3.25 |
| SMMR [70] | 3.24 |
| ARN [1] | 3.00 |
| Ranking-CNN [21] | 2.96 |
| MSFCL [156] | 2.90 |
| DAG-GoogleNet [141] | 2.87 |
| DAG-VGG16 [141] | 2.81 |
| Mean-Variance Loss [117] | 2.80 |
| MSFCL-LR [156] | 2.79 |
| Hu *et al.* [67] | 2.78 |
| BIF + FusionNet [146] | 2.76 |
| MSFCL-KL [156] | 2.73 |
| ADPF (ours) | **2.54** |

Table 4.2: MAE values for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting II.

| Method | MAE | | |
|---|---|---|---|
| | S1/S2+S3 | S2/S1+S3 | Average |
| KPLS [51] | 4.21 | 4.15 | 4.18 |
| MS-CNN [162] | 3.63 | 3.63 | 3.63 |
| MRNPE (AlexNet) [25] | 2.98 | 2.73 | 2.86 |
| MRNPE (VGG16) [25] | 2.85 | 2.60 | 2.73 |
| ARAN [26] | 2.77 | **2.48** | 2.63 |
| BridgeNet [93] | 2.74 | 2.51 | 2.63 |
| ADPF (ours) | **2.63** | 2.50 | **2.55** |

in the third setting, i.e., the *Setting III*, we select 5,492 facial images of White people to reduce the variance caused by imbalanced race distribution [1, 52, 130, 149]. Then, these 5,492 facial images are randomly split into two subsets, 80% of the them are used for training and the remaining 20% for testing. To further reduce the data distribution variance, in this setting, we use 5-fold cross validation to produce the final results.

For the FG-NET dataset, we use the LOPO strategy [44, 46, 98, 103, 135, 156]. In each fold, we use facial images of one subject for testing and the remaining images for training. Since there are 82 subjects, this process consists of 82 folds and the reported results are the average values.

For the CACD, following the setup in [25, 130, 135], the whole dataset is divided into three subsets, denoted as the training set, validation set, and testing set. The training set has facial images from 1,800 subjects, the validation set has facial images from 120 subjects, and the testing set has facial images from 80 subjects. The reported results are computed by training either on the training set or the validation set and evaluating on the testing set.

**Implementation Details**. ADPF is implemented based on the open-source

Table 4.3: MAE values for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting III.

| Method | MAE |
|---|---|
| AGES [45] | 8.83 |
| MTWGP [166] | 6.28 |
| CA-SVR [19] | 5.88 |
| DLA [149] | 4.77 |
| Rothe *et al.* [129] | 3.45 |
| DLDLF [135] | 2.94 |
| DRF [135] | 2.80 |
| ADPF (ours) | **2.71** |

deep learning framework Pytorch [122] and trained with the SGD algorithm with a batch size of 32. We first train the AttentionNet for 200 epochs and then the FusionNet for another 200 epochs with the parameters of the AttentionNet fixed. The initial learning rate for both networks is set to 0.1 and drops by a factor of 0.1 after every 50 epochs. When training the AttentionNet, we empirically set $\lambda$ in Eq. 4.14 to 0.01. Following our prior work, we use 5 patches when comparing with other state-of-the-art methods. All experiments are run on a single NVIDIA GTX 2080Ti GPU. To have a fair comparison against our prior work, we replace the age regression model used by our prior work with the age estimation loss in Eq. 4.13.

### 4.3.2 Evaluations on the MORPH II Dataset

The MAE values for the three aforementioned settings of the MORPH II dataset are tabulated in Table 4.1-4.3, respectively. In Table 4.2, the headings indicate the subsets used to compute the results. For example, *S1/S2+S3* indicates the model is trained on the *S1* subset and evaluated on the *S2* and *S3* subsets, and the *Average* column tabulates the mean value of the two columns on the left. The CS curves for the three settings are presented in Fig. 4.4-4.6, respectively. Note that not all

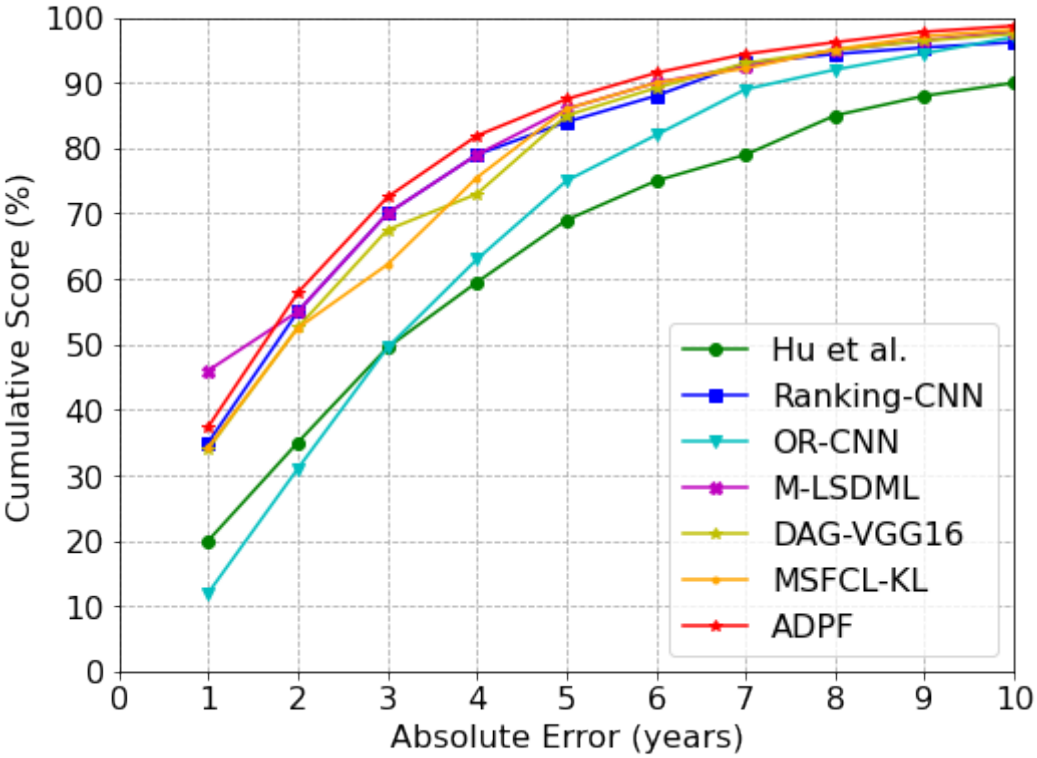


Figure 4.4: CS curves for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting I.

methods report the results under this metric. As can be seen from these tables and figures, ADPF outperforms all state-of-the-art methods that focus on improving the feature extractor like the DAG family (DAG-GoolgeNet and DAG-VGG16) [141], MSFCL family (MSFCL, MSFCL-LR, and MSFCL-KL) [156], and our prior work [146]. Also note that ADPF achieves comparable results to other methods that use customized estimators. For all three settings, the superior performance demonstrate that ADPF can predict ages accurately regardless of the imbalanced data distribution caused by other information like race.

### 4.3.3 Evaluations on the FG-NET Dataset

The MAE values and the CS curve are tabulated in Table 4.4 and depicted in Fig. 4.7, respectively, for the FG-NET dataset. Again, not all methods report the results

Table 4.4: MAE values for several state-of-the-art Face-based Age Estimation Methods on the FG-NET Dataset.

| Method | MAE |
|---|---|
| AGES [45] | 6.77 |
| IIS-LLD [46] | 5.77 |
| LARR [52] | 4.87 |
| Feng *et al.* [39] | 5.05 |
| BIF [53] | 4.77 |
| CPNN [46] | 4.76 |
| DEX [130] | 4.63 |
| CS-LBFL [103] | 4.43 |
| CS-LBMFL [103] | 4.36 |
| Mean-Variance Loss [117] | 4.10 |
| GA-DFL [97] | 3.93 |
| LSDML [98] | 3.92 |
| ARAN [26] | 3.79 |
| M-LSDML [98] | 3.74 |
| DLDLF [135] | 3.71 |
| DRF [135] | 3.47 |
| DAG-VGG16 [141] | 3.08 |
| DAG-GoogleNet [141] | 3.05 |
| ADPF (ours) | **2.86** |

Table 4.5: MAE values for several state-of-the-art Face-based Age Estimation Methods on the CACD.

| Method | MAE | |
|---|---|---|
| | train | val |
| DEX [130] | 4.79 | 6.52 |
| DLDLF [135] | 4.68 | 6.16 |
| DRF [135] | **4.61** | 5.63 |
| ADPF (ours) | 4.72 | **5.39** |

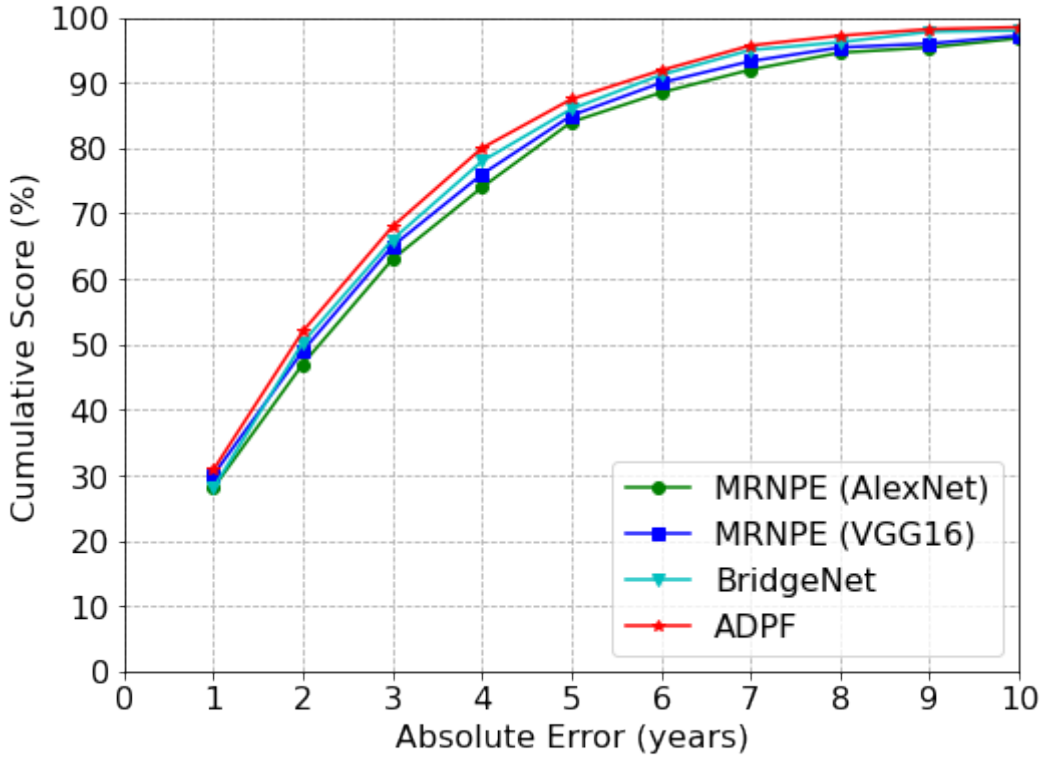


Figure 4.5: CS curves for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting II.

under the CS metric for the FG-NET dataset. It can be seen from Table 4.4 that ADPF achieves an MAE value under 3.00, which shows that it can perform well even with small datasets.

### 4.3.4 Evaluations on the CACD

Evaluation results for the CACD under the MAE metric are tabulated in Table 4.5. ADPF achieves the best performance when trained on the validation dataset but only achieves the third best performance when trained on the training set. This may due to the age labels in the training set not being accurate. Since the input to the FusionNet of ADPF is sixfold, i.e., it includes one facial image and five patches, compared to other single-input networks, inaccurate labels may confuse the model due to mis-information.

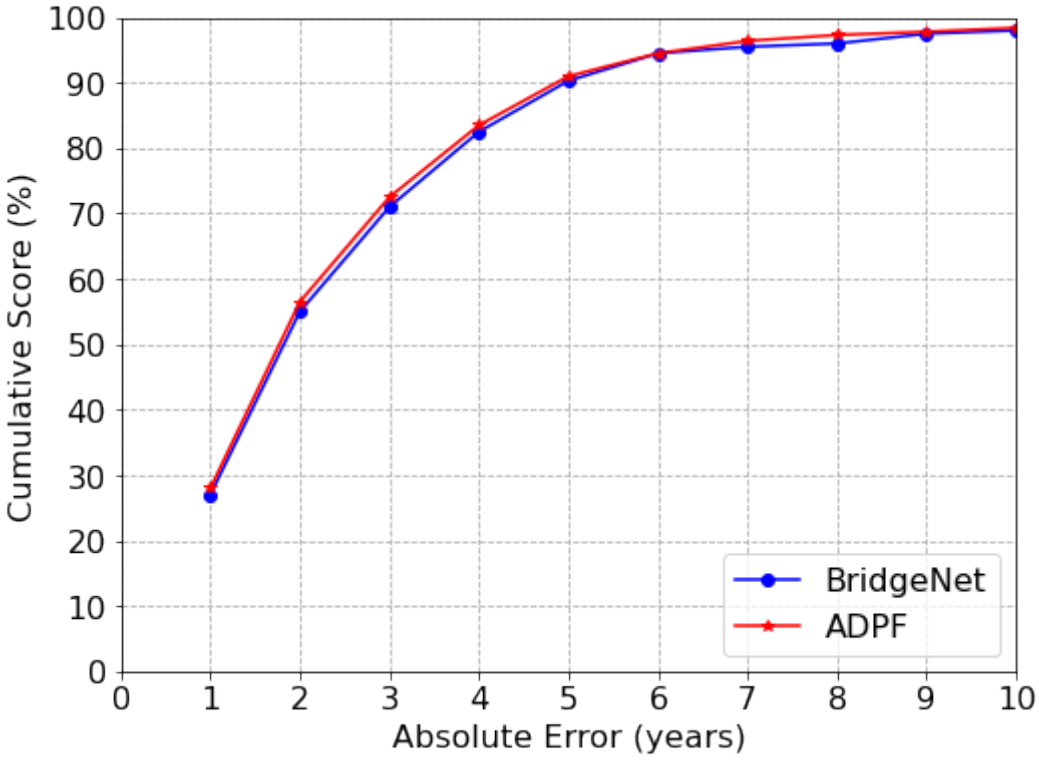


Figure 4.6: CS curves for several state-of-the-art Face-based Age Estimation Methods on the MORPH II Dataset under Setting III.

### 4.3.5 Ablation Study

We conduct ablation experiments to demonstrate the effectiveness of each component of ADPF. Specifically, we aim to demonstrate that: 1) the hybrid attention mechanism is more effective than the self-attention mechanism when discovering age-specific patches; 2) the ranking operation in RMHHA is beneficial for feature learning in the FusionNet; 3) the effectiveness of the diversity loss; and 4) the importance of combining the FusionNet and the AttentionNet in a single framework. To this end, we design several baseline models as follows:

- *ADPF w/SA*: ADPF with the self-attention mechanism instead of the hybrid attention mechanism in the AttentionNet. The single channel feature maps are then generated by performing summation along the channel axis of the self-attention maps.

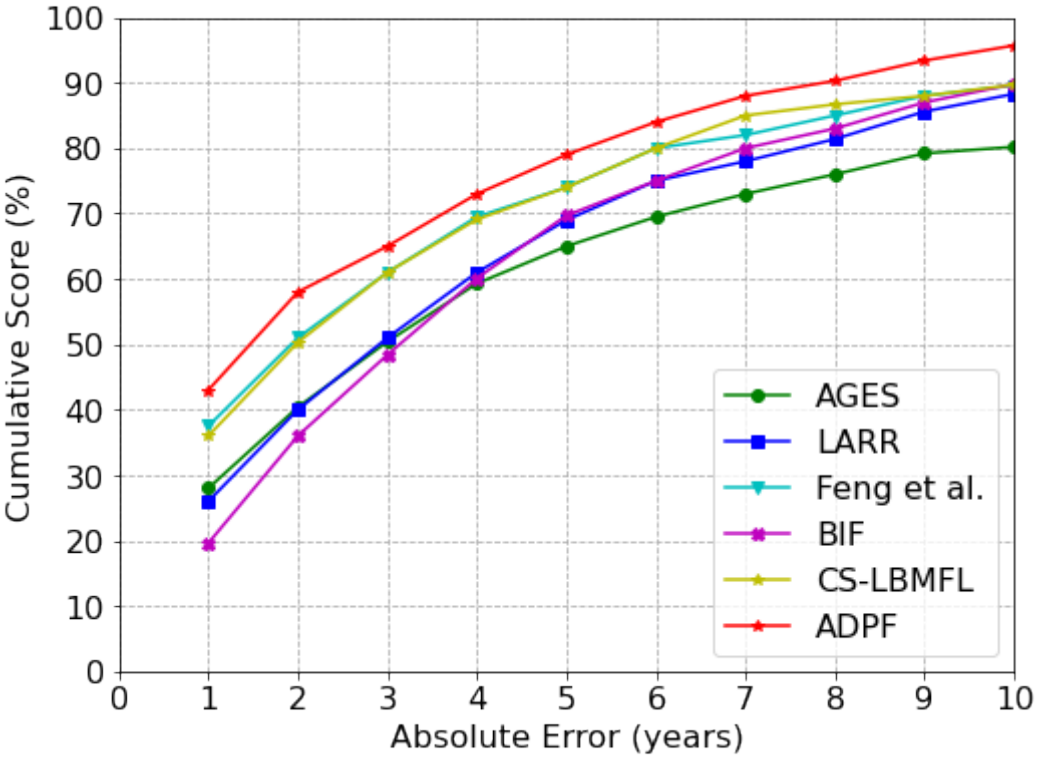


Figure 4.7: CS curves for several state-of-the-art Face-based Age Estimation Methods on the FG-NET Dataset.

- *ADPF w/o ranking*: ADPF without the ranking operation for age-specific patches.
- *ADPF w/o diversity*: ADPF without the diversity loss.
- *AttentionNet*: ADPF with no FusionNet.

The evaluation results on the MORPH II dataset, Setting I, for the aforementioned baseline models and ADPF are tabulated in Table 4.6. The attention maps computed by the *ADPF w/SA* baseline model are shown in Fig. 5.7. As shown in this figure, although *ADPF w/SA* can reveal key regions for age estimation, it may also reveal non-important regions, including sections of the background, which may be treated as noise during the feature learning process and eventually hinder the performance. In *ADPF w/o ranking*, we feed the patches into the FusionNet

based on their original order in the input tensor along the channel axis as produced by RMHHA. This feeding strategy cannot guarantee that the learning path for the most informative patch is long enough to extract meaningful features.

To demonstrate the effectiveness of the proposed diversity loss, we visualise the attention maps learned on the MORPH II dataset, Setting I, by ADPF and the baseline model *ADPF w/o diversity*. As shown in Fig. 4.9, in the *ADPF w/o diversity* baseline model, the two attention maps overlap in the highlighted nose region, which leads to redundant input information to the network. With the aid of the diversity loss, these key regions detected by these two attention maps are forced to move in opposite directions resulting in two attention maps with negligible overlap.

MAE values tabulated in Table 4.6 confirm the importance of combining the AttentionNet and the FusionNet in a single framework instead of using the AttentionNet exclusively. As we can see from this table, the performance of the *AttentionNet* baseline model significantly drops compared to that of ADPF. This is mainly due to the limited number of feature maps available to the FC layer in the AttentionNet. With such a limited number of feature maps, the estimator cannot get enough information from the feature extractor. However, implementing the AttentionNet in this way is essential to learn and rank multiple single-channel attention maps, which shows the importance of combining the AttentionNet and the FusionNet in a single framework.

### 4.3.6 Discussions

#### Training Efficiency

We compare the training time required by our prior work [146] and the ADPF on the MORPH II dataset with *Setting I*. The training times are tabulated in Table 4.7. Note that it takes about 70 hours to train the whole method in [146] out of which 60 hours are required to compute and rank BIF-based patches and 10 hours

Table 4.6: MAE values for several baseline models and the complete ADPF framework on the MORPH II Dataset under Setting I.

| Method | MAE |
|---|---|
| ADPF w/SA | 2.90 |
| ADPF w/o ranking | 2.74 |
| ADPF w/o diversity | 2.65 |
| AttentionNet | 3.31 |
| ADPF | **2.54** |

Figure 4.8: Attention maps computed by (upper row) the ADPF framework and (bottom row) the *ADPF w/SA* baseline model.

to train the CNN. Thanks to the proposed RMHHA mechanism, ADPF only takes about one third of this time to converge with significantly boosted performance (see MAE values). In addition, the process of acquiring patches and training the CNN can only be done separately in [146]. On the contrary, in ADPF, the training of the FusionNet can be done directly after the AttentionNet converges, which further makes the training process more time-efficient.

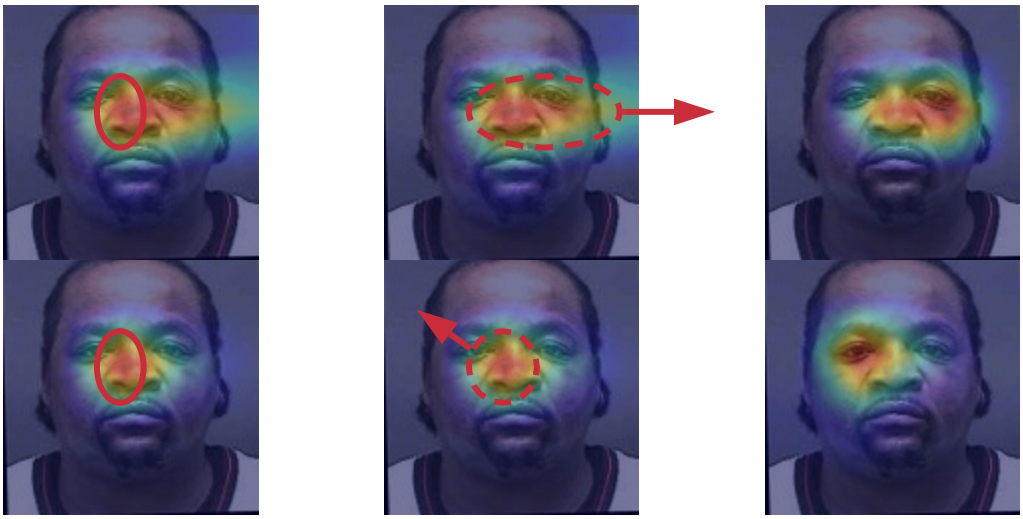

Figure 4.9: **Left**: Two attention maps overlap in the annotated area with out the supervision from the diversity loss. **Middle**: By minimising the diversity loss, the two attention maps are forced to move in opposite directions. **Right**: attention maps generated by using the diversity loss.

Table 4.7: The time it costs when the FusionNet and ADPF converges.

| Method | Training time (hours) | MAE |
|---|---|---|
| BIF + FusionNet [146] | 70 | 2.76 |
| ADPF | **25** | **2.54** |

### Robustness of Age-Specific Patches

We visually compare the patches computed by the BIF and Adaboost algorithms used in [146] and those computed by RMHHA. This comparison is conducted on the CACD dataset as the facial images in this dataset contain PIE variations. Fig. 4.10 depicts sample patches, where the most informative patches computed by [146] are marked with red boxes. It is clear that the location and shape of each patch computed by [146] are identical for all the images. On the contrary, the location and shape of the patches computed by the RMHHA vary from image to image. For example, in the bottom row, the patch capturing the right laughline is larger than

Figure 4.10: Sample age-specific patches computed by our prior work [146] and the ADPF framework. The left column depicts the original facial images with patches computed by [146] highlighted in red. The five patches computed by the ADPF framework are depicted in the last five columns. Within these columns, the patches are depicted from left to right in descending order in terms of their importance.

that of the other two images, which allows capturing the complete skin texture of this key region.

### Number of Heads

The performance of ADPF with different number of attention heads is tabulated in Table 4.8. We can see that the best performance can be achieved when 5 or 6 attention heads are implemented. This may due to the fact that with less heads, some age-specific patches may remain undiscovered. Moreover, since most of the facial regions are already revealed when 5 attention heads are used, adding more heads only forces the framework to attend to irrelevant regions like the background, which as discussed previously, can be treated as noise and degrade the performance. Since 6 heads requires more time to train with no significant performance gains, 5 is an appropriate number to be used by ADPF.

Table 4.8: Performance of ADPF with different number of attention heads on the MORPH II dataset under Setting I.

| # Heads | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| MAE | 2.77 | 2.62 | **2.54** | **2.54** | 2.55 | 2.61 |

## 4.4 Conclusion

In this chapter, we proposed the ADPF framework to improve the performance of the face-based age estimation task. Our framework merges an AttentionNet and a FusionNet. The AttentionNet includes a novel hybrid attention mechanism, namely RMHHA, which allows learning multiple single-channel attention maps to reveal age-specific patches. After ranking them, these patches are used by the FusionNet, along with the facial image to compute the final age prediction. Based on evaluations on several benchmark datasets, ADPF significantly improves prediction accuracy compared to several state-of-the-art methods. ADPF also outperforms our previous work, both in terms of accuracy and training times. Since this work focuses on building customised feature extractors, in the future, we will investigate the design of customised estimators to further boost performance by, for example, considering the ordinal information among ages and further minimising the distance between label distributions and feature distributions.

# Chapter 5

# Age-Oriented Face Synthesis with Conditional Discriminator Pool and Adversarial Triplet Loss

## 5.1 Introduction

AOFS is a generative task aiming to generate older and younger faces by rendering facial images with natural ageing and rejuvenating effects. An efficient AOFS method can be integrated into a wide range of forensic and commercial applications, e.g., tracking persons of interest like suspects or missing children over a long time span, predicting the outcomes of a cosmetic surgery, and generating special visual effects on characters of video games, films and dramas [40, 88]. The synthesis in recent works [92, 151, 160, 170] is usually conducted among age categories (e.g., the 30s, 40s, 50s) rather than specific ages (e.g., 32, 35, 39) since there is no noticeable visual change of a face over a few years.

The vanilla GAN [50] is commonly used as the backbone of several state-of-

the-art AOFS methods [5, 47, 92, 118, 171]. One of the biggest advantages of the vanilla GAN over other generative methods, like the Variational Autoencoder [79], is that it can generate sharp and realistic images by playing a minimax game between the generator and the discriminator. However, the vanilla GAN suffers from the mode collapse issue caused by the vanishing gradient due to the involvement of the negative log-likelihood loss [6]. Specifically, once the discriminator converges, the loss does not penalise the generator any further [17]. This allows the generator to find a specific mode (i.e., a distribution) that can easily fool the discriminator [10]. The mode collapse issue may also occur in the AOFS task, where a mode is represented by an age category. Within this context, the vanilla GAN may generate faces with limited variations, resulting in poor synthesis accuracy.

To boost the state-of-the-art performance in the AOFS task, this work proposes an AOFS method that includes two novel components. Namely, a CDP and an Adversarial Triplet loss. The proposed CDP helps to achieve a high synthesis accuracy by alleviating the mode collapse issue. Specifically, it allows learning multiple modes (i.e., age categories) explicitly and independently to generate realistic faces with a wide range of variations. Our CDP comprises multiple feature-level discriminators that learn the transformations from the source age category to the target age category. For each transformation, only the feature-level discriminator associated with the target age category is used. As a result, each feature-level discriminator only needs to learn one age category throughout the entire training process. The proposed Adversarial Triplet loss helps to preserve the identity information in the synthesised faces. This loss, which improves the Triplet loss [53], uses an additional ranking operation that can further optimise the distances within a triplet of feature embeddings comprising an *anchor*, a *positive* and a *negative*. Specifically, it helps to bring the *positive* much closer to the *anchor*, while guaranteeing that the distance between the *anchor* and the *negative* is larger than that between the *anchor* and the *positive*. The additional ranking operation forces the triplets to a play zero-sum game [5] during training. As a result, our Adversarial Triplet loss yields high-density

clusters with dramatically reduced intra-class variances in the feature space.

## 5.2 Proposed AOFS Method

In this section, we explain in detail our proposed method by first formulating the problem and explaining the pre-trained MTFE used to extract age-specific and identity-specific features. We then present the proposed CDP and the Adversarial Triplet loss. Finally, we explain the overall loss used to train our method.

### 5.2.1 Problem Formulation

Since the transformation is conducted among age categories rather than specific ages, following the prior work in [92, 101, 160], we divide the data into four categories according to the following age ranges: $30^{-}$, $31-40$, $41-50$, and $51^{+}$. Each category is denoted by $C_i$, where $i \in [1, 4]$.

To render ageing and rejuvenating effects, the proposed AOFS method takes two faces, $x \in C_X$ and $y \in C_Y$, and the age label of $y$, $l^y_{age}$, as the inputs, where $X \neq Y$. Specifically, $x$ is the face that is to be aged or rejuvenated and $y$ carries the desired age information. Our method aims to generate an aged or rejuvenated $x$, denoted by $\tilde{x}$, which is expected to belong to the same age category as $y$. Moreover, to ensure that the identity information is effectively preserved in $\tilde{x}$, our method also uses other images in the same batch, $\{x'\}$, to compute the Adversarial Triplet loss. It is worth noting that both $x'$ and $y$ do not share the same identity information of $x$.

In summary, the proposed method achieves three goals simultaneously: 1) To generate realistic aged and rejuvenated faces; 2) to force the synthesised faces to be within the target age category; and 3) to preserve the identity information in the synthesised image. The architecture of our proposed AOFS method is illustrated in Fig. 5.1.

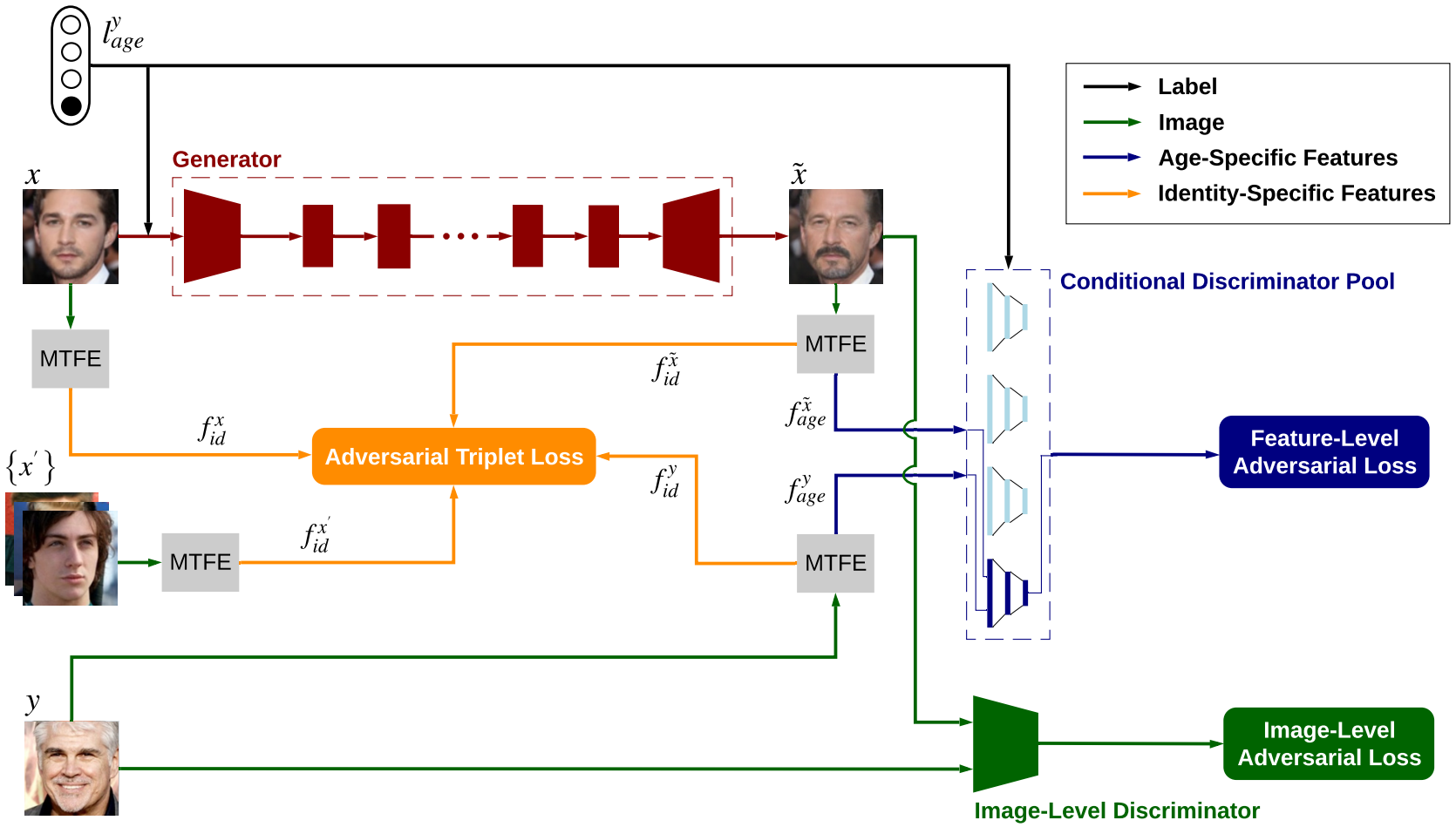


Figure 5.1: Architecture of the proposed AOFS method. It consists of a generator with residual blocks (red rectangles), an image-level discriminator, and a CDP that contains several feature-level discriminators. The number of feature-level discriminators equals the number of age categories that the method should learn. Two adversarial losses are used to synthesise realistic aged and rejuvenated faces. To further optimise the identity features in the synthesised image, $\tilde{x}$, we leverage additional input images, $\{x'\}$, that are within the same age category as the source image, $x$. Image $y$ carries the target age information for $\tilde{x}$.

### 5.2.2 Multi-Task Feature Extractor

The CDP and the Adversarial Triplet loss of the proposed AOFS method use age-specific and identity-specific features from input images and synthesised images. To extract and disentangle these features, we use the decomposition method proposed in [150]. Specifically, we use a ResNet-50 [58] as the backbone. The architecture of this feature extractor is depicted in Fig. 5.2. This model decomposes all the features extracted from a facial image into two components based on a spherical coordinate system, which is formulated as:

$$f_{sphere} := \{r; \boldsymbol{theta}\}, \tag{5.1}$$

where the $f_{sphere}$ is the set of features after the decomposition in which the angular component $\boldsymbol{theta} = \{\theta_1, \theta_2, ..., \theta_k\}$ indicate the identity-specific features for $k$ identities, and the radial component $r$ encodes the age-specific features.

We replace the regression loss used to learn age-specific features in [150] with an age regression model [130, 146] to supervise the age-specific learning process, which has been shown to achieve better performance for the age estimation task. We observe that feature extractors trained in this multi-tasking manner can achieve higher accuracy on both the age category classification and identity classification tasks than single-task networks. Additionally, we use our proposed Adversarial Triplet loss to learn identity-specific features.

### 5.2.3 Conditional Discriminator Pool

In the vanilla GAN with a single image-level discriminator, the loss function for face synthesis is usually formulated as:

$$\begin{aligned} \mathcal{L}_{adv} =& \mathbb{E}_y[logD(y)] \\ &+ \mathbb{E}_x[log(1 - D(G(x))], \end{aligned} \tag{5.2}$$

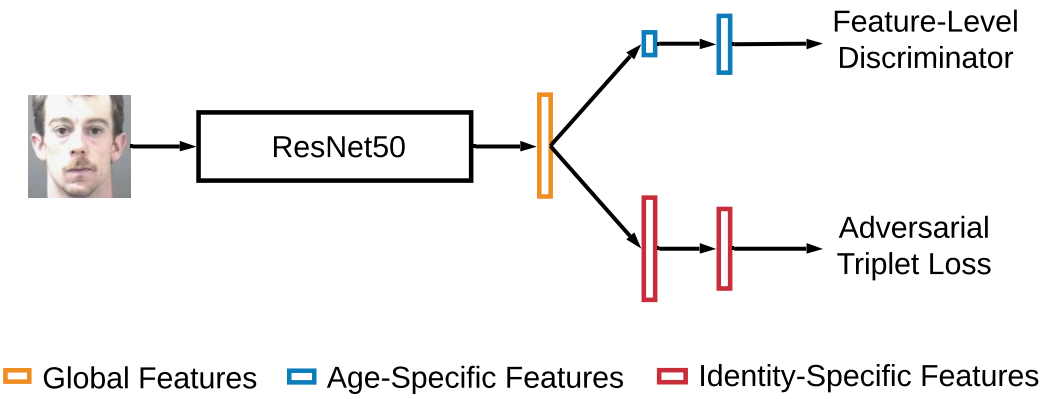


Figure 5.2: Architecture of our MTFE. After the decomposition, we resize each set of task-specific features to be used by the corresponding feature-level discriminator of the CDP or the Adversarial Triplet loss.

where $G$ is the generator trying to minimise the loss, and $D$ is the discriminator trying to maximise the loss. As mentioned before, GANs based on this loss function suffer from the mode collapse issue. To force the network to learn each mode independently and thus alleviate this issue, one can add more discriminators directly. However, such an strategy may lead to a high computational complexity and redundancy during training, as not all the discriminators are expected to back-propagate the loss during each transformation. Therefore, we propose a mechanism to select the corresponding discriminator for each transformation based on the input label that represents the target age information. Let us recall that our proposed AOFS method treats each age category as a mode, which results in four modes in total. We use the input label, $l^y_{age}$, to select the corresponding discriminator that learns the target age category. Our proposed method implements this mechanism on discriminators at the feature level, which are used to synthesise ageing and rejuvenating effects. Therefore, we assemble four feature-level discriminators with an identical architecture to form our CDP. Each feature-level discriminator targets one mode. Our method additionally uses an image-level discriminator to remove artificial effects from the synthesised faces. As illustrated in Fig. 5.1, in each transformation, our method leverages the selected feature-level discriminator alongside the image-level discriminator.

It is important to note that an alternative way to select the feature-level discriminator is by employing an additional classifier. However, within the context of AOFS, the accuracy of classifying age categories may be very low, from 25% to 60% depending on the specific age category in different AOFS benchmark datasets [101, 151]. Employing such a low-accuracy classifier may result in a selecting a discriminator that learns an incorrect mode. Instead, we directly use $l^y_{age}$ to select discriminators, which guarantees that, in each transformation, the discriminator associated with the target mode is used. We then formulate the feature-level adversarial loss as follows:

$$\begin{aligned}\mathcal{L}_{adv_{feature}} &= \mathbb{E}_{f^y_{age}}[log(FD_{C_i}(f^y_{age})|l^y_{age})] \\ &+\mathbb{E}_{f^{\tilde{x}}_{age}}[log(1-(FD_{C_i}(f^{G(x|l^y_{age})}_{age})|l^y_{age}))],\end{aligned} \tag{5.3}$$

where $FD_{C_i}$ is the selected feature-level discriminator trying to maximise the loss; $f^y_{age}$ denotes the age-specific features extracted from the target image, $y$; and $f^{G(x|l^y_{age})}_{age}$ denotes the age-specific features extracted from the synthesised image, $\tilde{x}$, where $G(x|l^y_{age})$ is the generator that produces $\tilde{x}$ conditioned on $l^y_{age}$. Finally, $l^y_{age}$ is a one-hot encoded vector indicating the label for the target age category, $C_i$.

### 5.2.4 Adversarial Triplet Loss

The Triplet loss [131] with three feature embeddings is formulated as:

$$\mathcal{L}_{Triplet}(a,p,n) = \sum_{a,p,n}[m + Dist_{a,p} - Dist_{a,n}]_+, \tag{5.4}$$

where $Dist_{j,k}$ indicates the Euclidean distance between embeddings $j$ and $k$ in the feature space and $a, p, n$ are the indices of the *anchor*, the *positive* and the *negative*, respectively. This loss forces $Dist_{a,n}$ to be larger than $Dist_{a,p}$ by at least a margin $m$. However, once this criterion is satisfied, $Dist_{a,p}$ cannot be further minimised, which may lead to large intra-class variances. To overcome this problem, we add another ranking operation to Eq. (5.4), which forces $Dist_{a,n}$ to be larger than the

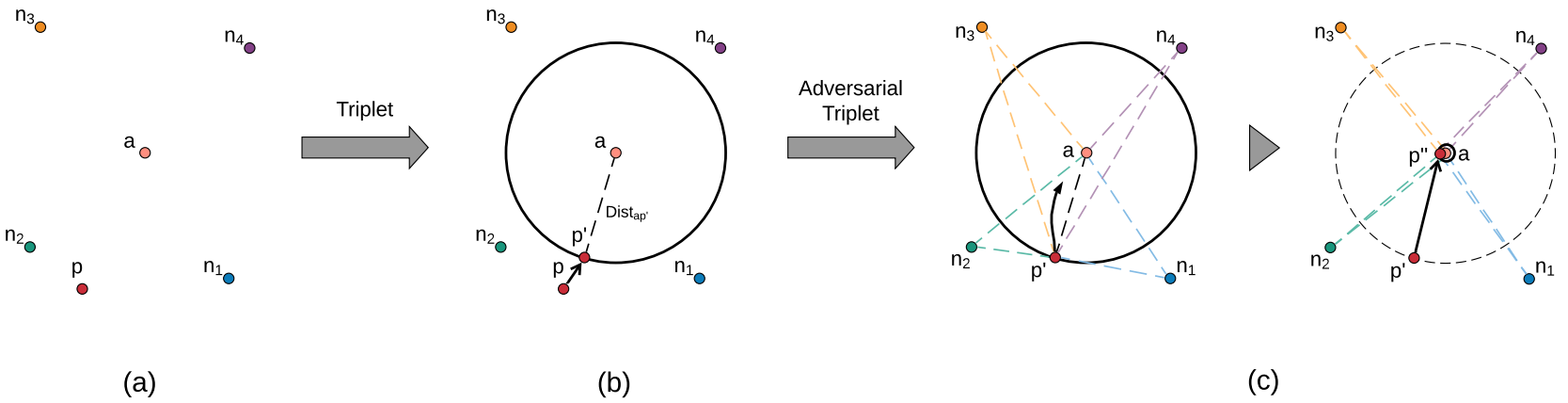


Figure 5.3: An example showing how the Adversarial Triplet loss works. $a$ (*anchor*) and $p$ (*positive*) are feature embeddings representing the same class. The *negatives* $n_1$, $n_2$, $n_3$, and $n_4$ indicate feature embeddings from other classes, each one from a distinct class. (a) Original positions of these feature embeddings. (b) By using the Triplet loss, $p$ can move towards $p'$ when minimising Eq. (5.4). (c) Our Adversarial Triplet loss guarantees that for each $n_i$ where $i \in [1, 2, 3, 4]$, $Dist_{an_i} \approx Dist_{n_i p}$ by adding an additional operation as formulated in Eq. (5.5). In this case, $p'$ may continue moving towards $a$ and end up at a location which is extremely close to it, i.e., $p''$.

distance between $n$ and $p$, $Dist_{n,p}$. This additional operation helps to further bring $p$ closer to $a$ by forcing different triplets with the same $a$ and $p$ but different $n$ to play a zero-sum game:

$$\begin{aligned}\mathcal{L}_{AT}(a,p,n) = \sum_{a,p,n} [m + Dist_{a,p} - Dist_{a,n}]_+ \\ + [Dist_{n,p} - Dist_{a,n}].\end{aligned} \tag{5.5}$$

Let us assume there are several triplets with the same $a$ and $p$, but different $n$, where each distinct $n$ is denoted by $n_i$. Under this assumption, the Triplet loss in Eq. (5.4) can be minimised as long as $Dist_{a,n_i} > Dist_{a,p} + m$, which may result in clusters with large intra-class variances. To reduce such variances, $Dist_{a,n_i}$ should be larger than $Dist_{n_i,p}$. Let us take the triplets $a - p - n_1$ and $a - p - n_3$ in Fig. 5.3 as an example, where $n_1$, $n_2$, $n_3$, and $n_4$ are all from different classes. In this example, both $n_1$ and $n_3$ should maintain their relative position with respect to the $a - p$ cluster in order to also be far from other neighbouring clusters. In other words, $n_1$ and $n_3$ should not move towards either $n_2$ or $n_4$. In this case, $\mathcal{L}_{AT}(a,p,n_1)$ tries to pull $p$ towards $n_1$ and minimise $Dist_{n_1,p}$, while $\mathcal{L}_{AT}(a,p,n_3)$ tries to pull $p$ towards $n_3$ and minimise $Dist_{n_3,p}$. Therefore, $\mathcal{L}_{AT}(a,p,n_1)$ and $\mathcal{L}_{AT}(a,p,n_3)$ play a zero-sum game as minimising one loss increases the other. This is also true for $\mathcal{L}_{AT}(a,p,n_2)$ and $\mathcal{L}_{AT}(a,p,n_4)$. In order to minimise all losses in this example, i.e., to have a total loss equal to zero, $p$ should be in the same position as $a$ so that $Dist_{a,n_i} = Dist_{n_i,p}$. In practice, however, our Adversarial Triplet loss pulls $p$ to a position very close to $a$ so that $Dist_{a,n_i} \approx Dist_{n_i,p}$.

Fig. 5.4 demonstrates the performance of the Adversarial Triplet loss on a real dataset. In this example, the feature distribution of the MNIST dataset for classification is presented. To this end, we employ an Alexnet [83] as the deep network, but replace all the fully-connected layers, except the output layer, by a single linear layer with two neurons for visualisation purposes. From the figure, we can observe that the features learned by the Adversarial Triplet loss dramatically

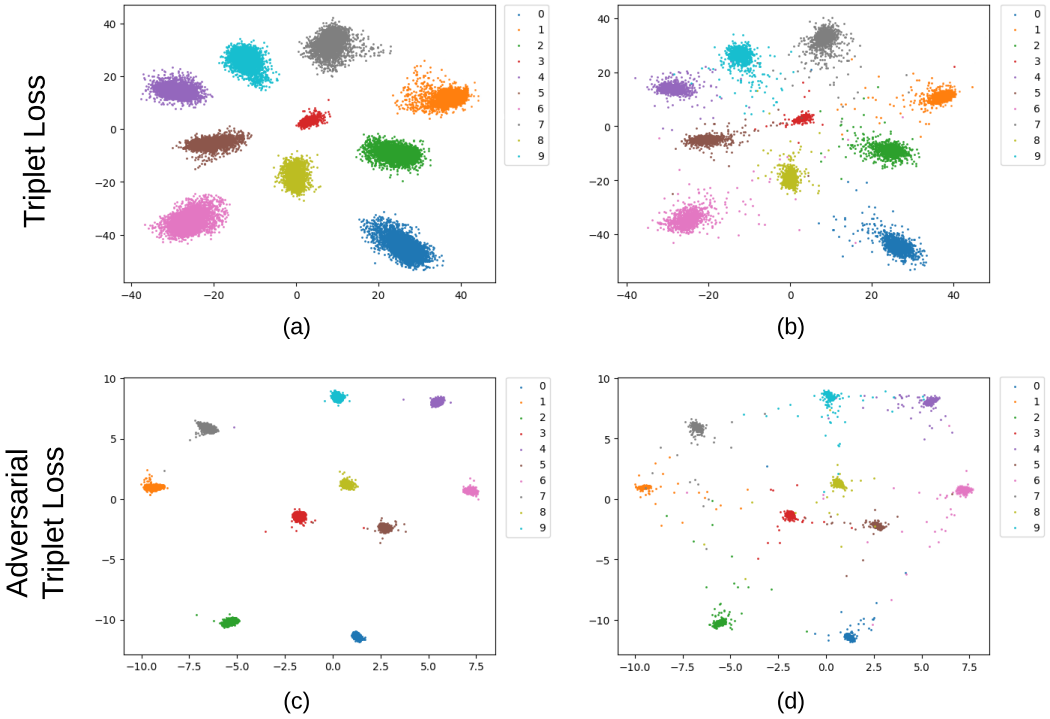


Figure 5.4: Feature distribution of the MNIST dataset for classification on (a),(c) the training set and (b),(d) the test set when the Triplet loss and the Adversarial Triplet loss are used.

Table 5.1: Classification accuracy (%) n on the MNIST dataset.

| Loss | Triplet | Adversarial Triplet |
|---|---|---|
| Accuracy | 99.43 | **99.67** |

reduce the intra-class variances compared to the features learned by the Triplet loss. The classification accuracy attained by each loss is tabulated in Table 5.1.

One of the most critical issues in the Triplet loss is that as the number of triplets grows, many triplets can easily satisfy the constraint in Eq. (5.4), which in turn may lead to poor convergence [131]. To overcome this issue in the Adversarial Triplet loss, we adopt a hard negative mining strategy [62]. Specifically, we use an online hard sample mining method in which each batch consists of samples from $T$ classes, and each class has $S$ samples within one batch, for a batch size of $B = TS$. In this method, each sample in a batch acts as the *anchor* for one triplet, thus, there are a total of $B$ triplets within one batch. For each *anchor*, a hardest *positive* sample with the largest distance and a hardest *negative* sample with the smallest distance are selected to form a triplet. This method does not require pre-defining the triplets and can generate hard triplets in an online manner. After incorporating this hard sample mining strategy, our Adversarial Triplet loss in Eq. (5.5) is as follows:

$$\begin{aligned}\mathcal{L}_{AT}(a,p,n) = \sum_{t=1}^{T}\sum_{s=1}^{S}[m + \max_{p} Dist_{a,p} - \min_{n} Dist_{a,n}]_+ \\ + [Dist_{n,p} - \min_{n} Dist_{a,n}],\end{aligned} \tag{5.6}$$

where $t$ is the class index and $s$ is the image index for each class in one batch.

Since we are trying to optimise the identity-specific features on the synthesised faces when training our AOFS method, we use the identity-specific features, $f_{id}^{x}$, from the source image as the *anchor* and the identity-specific features, $f_{id}^{\tilde{x}}$, from the synthesised image as the *positive*. In addition, we use all other images in the same batch that do not share the same identity with the source image as the *negatives*.

The Adversarial Triplet loss of our AOFS method with the hard sample mining strategy is then formulated as:

$$\begin{aligned}\mathcal{L}_{AT}(f_{id}^{x}, f_{id}^{\tilde{x}}, \left\{f_{id}^{x'}, f_{id}^{y}\right\}) = \sum_{t=1}^{T}\sum_{s=1}^{S} \\ [m + Dist_{f_{id}^{x}, f_{id}^{\tilde{x}}} - \min_{\left\{f_{id}^{x'}, f_{id}^{y}\right\}} Dist_{f_{id}^{x}, \{f_{id}^{x'}, f_{id}^{y}\}}]+ \\ + [Dist_{\{f_{id}^{x'}, f_{id}^{y}\}, f_{id}^{\tilde{x}}} - \min_{\left\{f_{id}^{x'}, f_{id}^{y}\right\}} Dist_{f_{id}^{x}, \{f_{id}^{x'}, f_{id}^{y}\}}],\end{aligned} \tag{5.7}$$

where $\left\{f_{id}^{x'}\right\}$ are the identity-specific features of images within the same age category as the source image but carrying different identity information, and $f_{id}^{y}$ are the identity-specific features of images within the target age category. It is worth noting that the above equation do not have the *max* operation as in Eq. (5.6) since the *positive* in this case, $f_{id}^{\tilde{x}}$, is synthesised thus cannot be selected.

### 5.2.5 Overall Loss

The image-level adversarial loss in our AOFS method is formulated as:

$$\begin{aligned}\mathcal{L}_{adv_{image}} =& \mathbb{E}_{y}[logD(y)] \\ &+ \mathbb{E}_{x}[log(1 - D(G(x|l_{age}^{y}))].\end{aligned} \tag{5.8}$$

The overall loss function, $\mathcal{L}_{overall}$, to train our method is a weighted summation of several losses, with $\mathcal{L}_{adv_{image}}$ removing ghost artifacts, $\mathcal{L}_{adv_{feature}}$ synthesising ageing and rejuvenating effects and attaining a high synthesis accuracy, and $\mathcal{L}_{AT}$ preserving the identity information:

$$\begin{aligned}\mathcal{L}_{overall} =& \mathcal{L}_{adv_{image}} + \lambda_{adv_{feature}}\mathcal{L}_{adv_{feature}} \\ &+ \lambda_{AT}\mathcal{L}_{AT},\end{aligned} \tag{5.9}$$

where $\lambda_{adv_{feature}}$ and $\lambda_{AT}$ control the relative importance among learning objectives.

## 5.3 Experiments

In this section, we first briefly describe the two AOFS benchmark datasets used in our experiments followed by the implementation details of our method. Then, we compare our method with state-of-the-art methods and conduct ablation studies, both qualitatively and quantitatively, to show that our method can achieve a high synthesis accuracy while preserving the identity information on the synthesised facial images.

### 5.3.1 Experimental Settings

All images are cropped to 128×128 pixels and aligned based on the location of the eyes. Since not all images can be aligned by using this technique, in the end, 55,062 images from the MORPH II dataset and 159,226 images from the CACD are used in our experiments. For each dataset, we use 80% of the images for training and the remaining 20% for testing. The number of training images for each age category in the MORPH dataset is 19,949, 12,496, 8,982, and 2,622, for the categories $\{30^{-}, 31-40, 41-50, 51^{+}\}$, respectively. For the CACD, the number of training images of each age category is 39,416, 33,742, 30959, and 23,262, respectively. There is no identity overlap between the training and test sets.

Follow previous works [151, 160], we conduct a five-fold cross validation for all our experiments. For the MORPH II dataset, each fold has about 2,550 subjects with 3,989, 2,499, 1,796, and 524 images within each age category, respectively. For the CACD, each fold contains about 400 subjects with 7,883, 6,748, 6,191 and 4,652 images within each age category, respectively.

To evaluate our method and demonstrate its robustness, we use another two large-scale benchmark datasets to train two separate validation networks, one for each criterion. In particular, we use the AgeDB dataset [111], which is widely used for age estimation, to train the network that evaluates the synthesis accuracy and a face recognition benchmark dataset, the VGGFace2 dataset [13], to train the

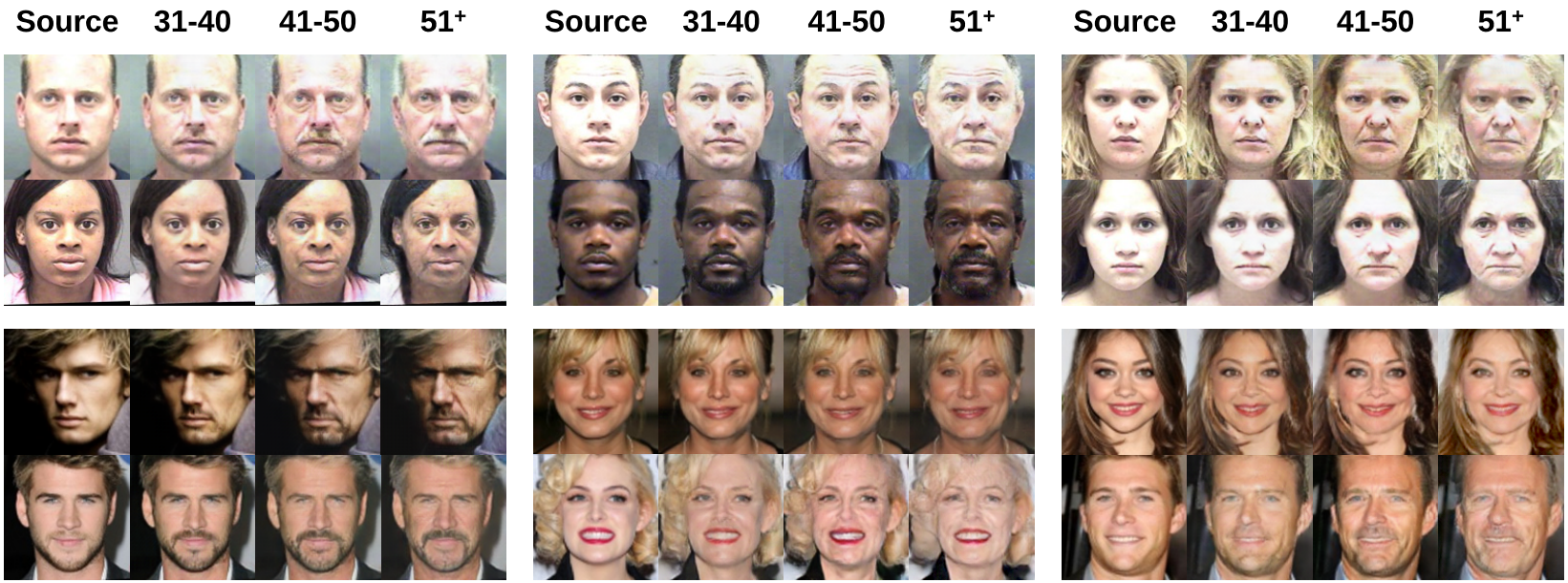


Figure 5.5: Ageing results. The top five rows show the synthesised results on the MORPH II dataset, and the bottom five rows show the synthesised results on the CACD.

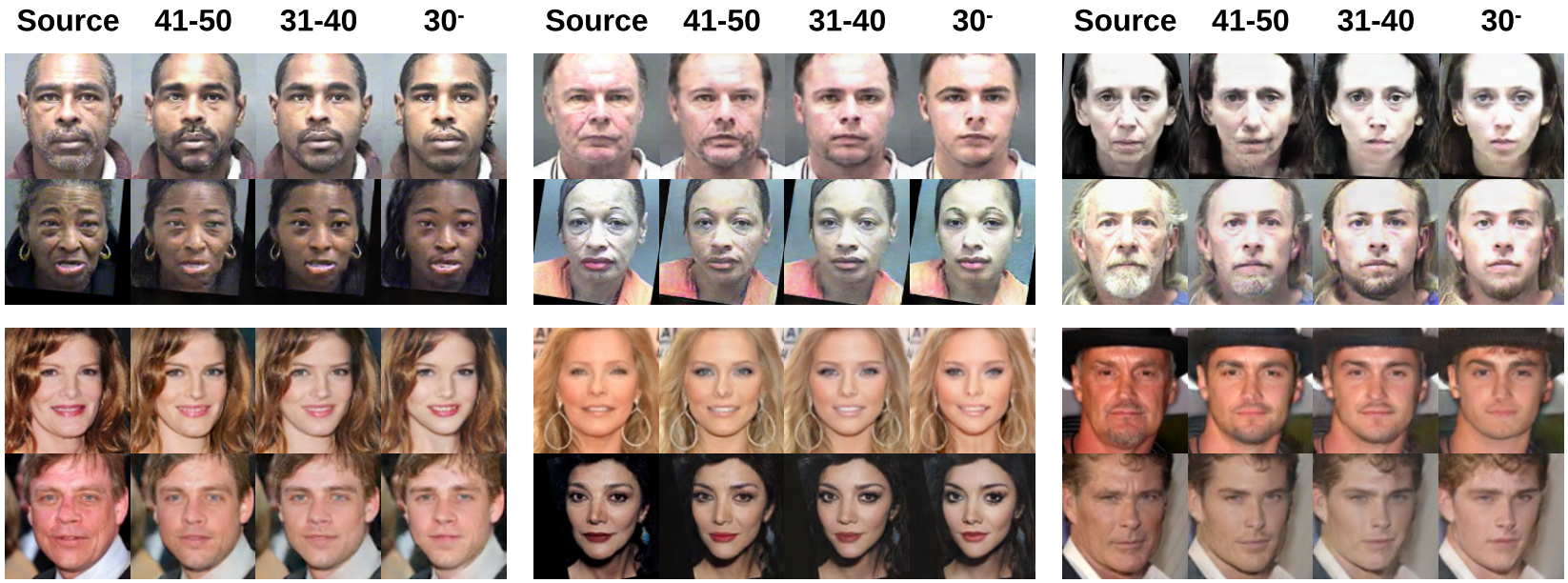


Figure 5.6: Rejuvenating results. The top five rows show the synthesised results on the MORPH II dataset, and the bottom five rows show the synthesised results on the CACD.

Table 5.2: Architecture of the generator.

| Encoder | | | |
|---|---|---|---|
| #Layer | Convolution | Normalisation | Non-linear |
| 1 | *k=7, s=1, p=1* | Instance | ReLU |
| 2 | *k=3, s=2, p=1* | Instance | ReLU |
| Residual Block (× 6) | | | |
| #Layer | Convolution | Normalisation | Non-linear |
| 1 | *k=3, s=2, p=1* | Instance | ReLU |
| 2 | *k=3, s=2, p=1* | Instance | ReLU |
| Decoder | | | |
| #Layer | Deconvolution | Normalisation | Non-linear |
| 1 | *k=3, s=2, p=1* | Instance | ReLU |
| 2 | *k=3, s=2, p=1* | Instance | Tanh |

Table 5.3: Architecture of the discriminators.

| Feature-Level (× 4) | | | |
|---|---|---|---|
| #Layer | Fully-Connected | Normalisation | Non-linear |
| 1 | *128* | Instance | LeakyReLU |
| 2 | *64* | Instance | LeakyReLU |
| 3 | *32* | Instance | LeakyReLU |
| 4 | *16* | Instance | LeakyReLU |
| 5 | *1* | - | - |
| Image-Level | | | |
| #Layer | Convolution | Normalisation | Non-linear |
| 1 | *k=3, s=2, p=1* | Instance | LeakyReLU |
| 2 | *k=3, s=2, p=1* | Instance | LeakyReLU |
| 3 | *k=3, s=2, p=1* | Instance | LeakyReLU |
| 4 | *k=3, s=2, p=1* | Instance | LeakyReLU |
| 5 | *k=3, s=1, p=1* | - | - |

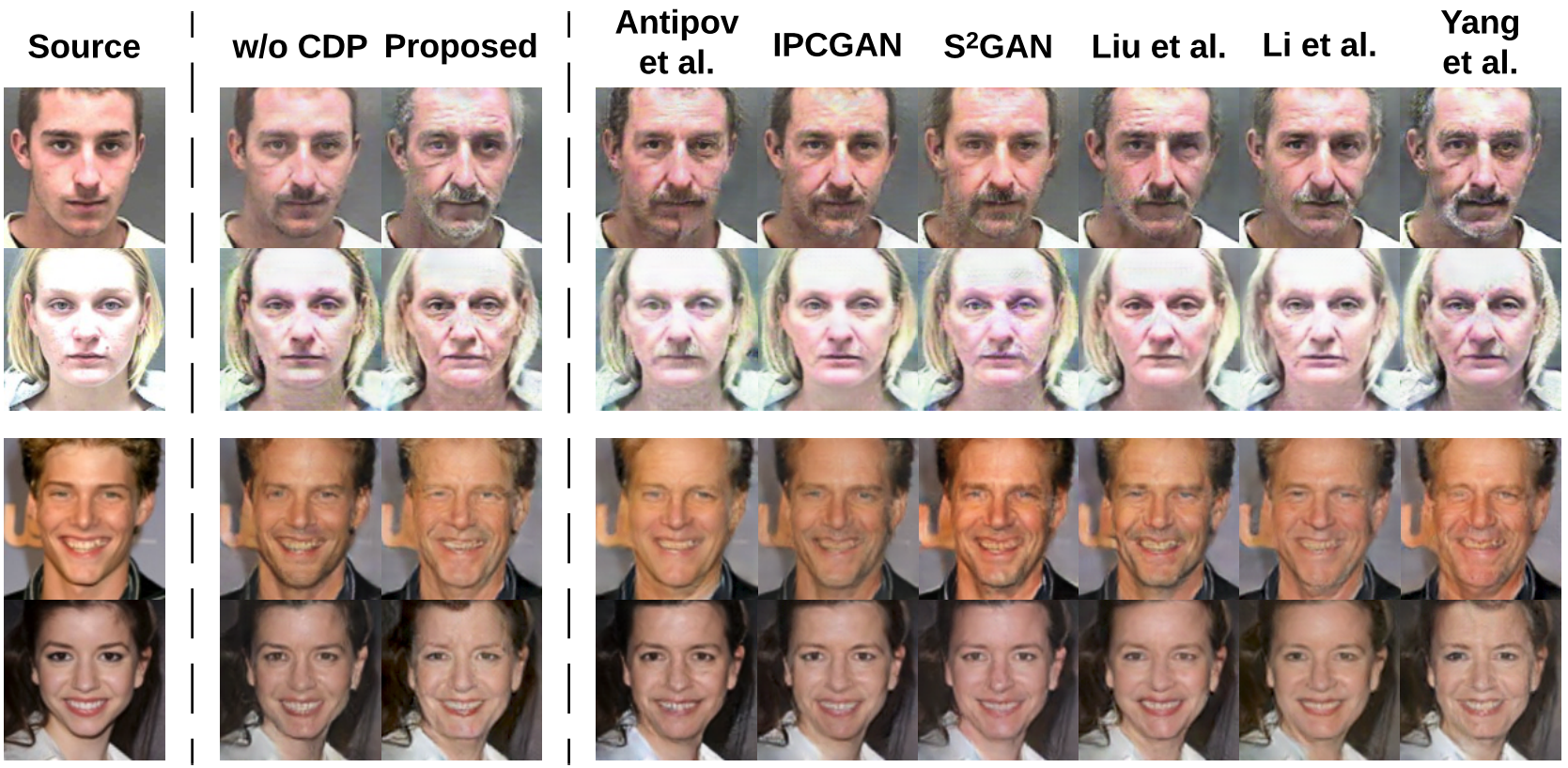


Figure 5.7: Visual comparison of a baseline model, six state-of-the-art works, and our proposed method on two benchmarks. The top two rows show the results on the MORPH II dataset and the bottom two rows show the results on the CACD. The input image is within the youngest group and the results are expected to be within the eldest group.

network that evaluates the identity permanence capabilities. In addition, we use the commonly used ResNet-50 as the backbone for both evaluation networks.

### 5.3.2 Network architecture

We employ the architecture from [174] for our generator. The generator has six residual blocks and each convolutional and deconvolutional layer is followed by an instance normalization and a ReLU function. For the image-level discriminator, we implement a patch discriminator [73] with five convolutional layers, each followed by an instance normalization and a LeakyReLU function. Each feature-level discriminator has the same architecture as that of the image-level discriminator but consists of fully-connected layers.

The details of the architectures of the generator and discriminators in our AOFS method are tabulated in Tables 5.2 and 5.3, respectively. In both tables, for each convolutional and deconvolutional layer, $k$ indicates the kernel size, $s$ indicates the stride, and $p$ indicates the padding size. In Table 5.3, the second column for the feature-level discriminators tabulates the dimensions of the corresponding layer.

### 5.3.3 Data augmentation

When training the MTFE and validation networks, we use a combination of rotation, flip, and crop operations to augment the data. Specifically, we first randomly rotate each image by a angle between +10 deg. and -10 deg., and then randomly flip the rotated image with a probability of 0.5. Finally, we pad the image on all sides with 10 pixels and crop the padded image at a random location to the original image size (i.e. $128 \times 128$ pixels). When training the proposed AOFS method, in order to increase the size of the training set without introducing additional variance to the dataset, we only use the flip operation.

Table 5.4: Age category classification accuracy (%) on the images synthesised for the MORPH II dataset and the CACD for the ageing process.

| | MORPH II | | | CACD | | |
|---|---|---|---|---|---|---|
| Age Category | 31-40 | 41-50 | $51^{+}$ | 31-40 | 41-50 | $51^{+}$ |
| Natural Faces | 59.04 ± 2.42 | 58.68 ± 2.18 | 58.83 ± 2.23 | 37.91 ± 5.09 | 37.34 ± 4.79 | 34.46 ± 4.92 |
| Antipov *et al.* [5] | 39.56 ± 2.28 | 39.79 ± 2.10 | 35.22 ± 2.50 | 20.29 ± 4.58 | 20.49 ± 5.04 | 18.43 ± 5.40 |
| IPCGAN [151] | 44.67 ± 2.25 | 44.70 ± 2.43 | 41.84 ± 1.77 | 24.90 ± 4.29 | 27.70 ± 4.25 | 28.49 ± 5.00 |
| $S^2$GAN [61] | 52.97 ± 2.65 | 52.46 ± 1.84 | 51.30 ± 1.98 | 29.25 ± 4.88 | 29.05 ± 4.62 | 26.33 ± 4.81 |
| Liu *et al.* [101] | 52.12 ± 1.97 | 53.85 ± 1.92 | 54.82 ± 1.45 | 29.31 ± 5.16 | 31.87 ± 4.95 | 32.79 ± 4.88 |
| Li *et al.* [92] | 51.22 ± 2.15 | 53.60 ± 1.74 | 54.61 ± 1.97 | 28.61 ± 4.41 | 31.02 ± 4.19 | 32.46 ± 4.75 |
| Yang *et al.* [160] | 53.24 ± 1.67 | 53.23 ± 2.86 | 53.20 ± 1.73 | 30.68 ± 4.12 | 30.85 ± 4.43 | 31.64 ± 4.38 |
| w/o CDP | 43.52 ± 1.73 | 41.53 ± 1.82 | 41.93 ± 1.45 | 25.01 ± 5.52 | 25.06 ± 4.89 | 25.55 ± 5.18 |
| Proposed | **56.60 ± 1.91** | **55.42 ± 1.80** | **54.63 ± 1.98** | **33.73 ± 3.91** | **33.77 ± 4.32** | **32.54 ± 4.61** |

### 5.3.4 Hyper-parameter setting

When training the MTFE, we set the batch size to 128 and the initial learning rate to 0.002 for both datasets. We train it for 500 epochs while decreasing the learning rate by 0.1 every 150 epochs. When training the AOFS method, we set the batch size to 8 and the initial learning rate to 0.0002. The learning rate decreases linearly after the first 25 epochs. We empirically set $\lambda_{adv_{feature}}$ to 1 and $\lambda_{AT}$ to 0.001. The margin hyper-parameter, $m$ in Eq. (5.7), is set to 0.3. We use the PyTorch framework [122] for the implementation and run each experiment for 50 epochs. All experiments are run on a single NVIDIA GTX2080Ti GPU.

### 5.3.5 Synthesis accuracy

We first qualitatively evaluate the synthesised facial images based on their visual quality. We then present quantitative results based on age category classification accuracy, image quality and the degree of mode collapse. We perform these evaluations for our AOFS method and several state-of-the-art methods. Note that, except for the IPCGAN, we tried our best to re-implement existing methods and obtained the results from our implementations.

#### Visual Quality

Fig. 5.5 and 5.6 show some sample images synthesised by our AOFS method. Fig. 5.5 shows ageing results for 6 subjects from the MORPH II dataset and 6 from the CACD using a source image from the youngest category ($30^{-}$). We can see that our method turns hair gray or white, introduces forehead wrinkles and nasolabial folds, and makes the skin to appear rough. Fig. 5.6 shows rejuvenating results for 6 subjects from each dataset using a source image from the oldest category ($51^{+}$). We can see that for these cases, our method removes wrinkles and gray/white hair.

We also evaluate six state-of-the-art methods, namely the method by Antipov *et al.* [5], the IPCGAN [151], the $S^2$GAN [61], and the methods by Liu *et al.* [101],

Table 5.5: Age category classification accuracy (%) on the images synthesised for the MORPH II dataset and the CACD for the rejuvenating process.

| | MORPH II | | | CACD | | |
|---|---|---|---|---|---|---|
| Age Category | $30^{-}$ | 31-40 | 41-50 | $30^{-}$ | 31-40 | 41-50 |
| Natural Faces | 63.08 ± 1.81 | 59.04 ± 2.42 | 58.68 ± 2.18 | 43.82 ± 4.06 | 37.91 ± 5.09 | 37.34 ± 4.79 |
| Antipov *et al.* [5] | 50.55 ± 2.32 | 44.71 ± 2.45 | 44.77 ± 1.84 | 28.41 ± 3.92 | 26.36 ± 5.87 | 26.17 ± 4.71 |
| IPCGAN [151] | 57.33 ± 1.82 | 52.03 ± 1.79 | 52.32 ± 2.21 | 32.67 ± 4.43 | 31.89 ± 4.50 | 31.41 ± 5.08 |
| $S^2$GAN [61] | 58.18 ± 1.83 | 54.11 ± 2.04 | 54.24 ± 1.43 | 33.36 ± 4.01 | 32.30 ± 4.38 | 32.63 ± 3.89 |
| Liu *et al.* [101] | 59.06 ± 2.41 | 55.33 ± 1.61 | 55.54 ± 2.01 | 36.65 ± 4.31 | 34.25 ± 4.34 | 34.26 ± 4.69 |
| Li *et al.* [92] | 58.87 ± 2.30 | 55.21 ± 2.18 | 55.06 ± 1.94 | 37.84 ± 4.66 | 34.95 ± 4.86 | 34.30 ± 4.26 |
| Yang *et al.* [160] | 60.79 ± 2.21 | 56.99 ± 2.17 | **56.65 ± 2.39** | 39.09 ± 4.72 | 35.62 ± 4.83 | 35.89 ± 4.61 |
| w/o CDP | 53.67 ± 2.35 | 51.41 ± 2.33 | 51.96 ± 2.45 | 29.17 ± 5.05 | 28.42 ± 5.39 | 28.67 ± 5.31 |
| Proposed | **61.20 ± 1.41** | **57.12 ± 1.36** | 56.55 ± 2.23 | **41.24 ± 4.12** | **36.84 ± 4.10** | **36.59 ± 4.81** |

Table 5.6: ResNet Score and Fréchet ResNet Distance on the MORPH II dataset.

| Model | RS | FRD |
|---|---|---|
| Antipov *et al.* [5] | 27.83 ± 1.34 | 31.72 ± 0.60 |
| IPCGAN [151] | 36.70 ± 1.18 | 28.08 ± 0.44 |
| $S^2$GAN [61] | 38.92 ± 1.14 | 25.64 ± 0.32 |
| Liu *et al.* [101] | 39.14 ± 1.23 | 25.57 ± 0.42 |
| Li *et al.* [92] | 39.26 ± 1.22 | 25.51 ± 0.41 |
| Yang *et al.* [160] | 43.35 ± 1.36 | 22.30 ± 0.59 |
| w/o CDP | 30.19 ± 1.26 | 28.62 ± 0.49 |
| Proposed | **44.04 ± 1.25** | **21.93 ± 0.46** |

Table 5.7: ResNet Score and Fréchet ResNet Distance on the CACD.

| Model | RS | FRD |
|---|---|---|
| Antipov *et al.* [5] | 24.71 ± 2.04 | 33.83 ± 0.95 |
| IPCGAN [151] | 33.21 ± 1.82 | 30.18 ± 0.79 |
| $S^2$GAN [61] | 34.24 ± 1.75 | 27.01 ± 0.61 |
| Liu *et al.* [101] | 34.54 ± 1.86 | 26.99 ± 0.63 |
| Li *et al.* [92] | 35.00 ± 1.91 | 26.91 ± 0.67 |
| Yang *et al.* [160] | 37.39 ± 2.09 | 24.62 ± 0.87 |
| w/o CDP | 30.87 ± 1.87 | 30.71 ± 0.82 |
| Proposed | **38.55 ± 1.90** | **23.98 ± 0.73** |

Table 5.8: Degree of mode collapse as measured by the KL divergence.

| Model | MORPH II | CACD |
|---|---|---|
| Antipov *et al.* [5] | 1.86 ± 0.10 | 1.93 ± 0.13 |
| IPCGAN [151] | 0.64 ± 0.15 | 0.68 ± 0.21 |
| $S^2$GAN [61] | 0.59 ± 0.08 | 0.62 ± 0.11 |
| Liu *et al.* [101] | 0.55 ± 0.09 | 0.57 ± 0.13 |
| Li *et al.* [92] | 0.55 ± 0.11 | 0.58 ± 0.14 |
| Yang *et al.* [160] | 0.49 ± 0.04 | 0.52 ± 0.05 |
| w/o CDP | 1.19 ± 0.09 | 1.30 ± 0.14 |
| Proposed | **0.37 ± 0.04** | **0.42 ± 0.07** |

Li *et al.* [92], and Yang *et al.* [160]. To have a fair comparison, we replace the feature extractors in these methods with our pre-trained MTFE and use the same number of residual blocks in their generator expect for the method in [5], as there is no residual block originally involved in this particular method.

Since the synthesis accuracy of our AOFS method depends on the CDP, we also evaluate a baseline model without the CDP (hereinafter called *w/o CDP*) as part of an ablation study. The *w/o CDP* model replaces the CDP with a simple feature-level discriminator, which makes this model similar to a vanilla GAN but with two discriminators, one at the feature level and the other at the image level.

Fig. 5.7 depicts the visual results of these evaluations. Note that it is visually evident that the results generated by the *w/o CDP* model do not contain much ageing and rejuvenating effects as this model suffers from the mode collapse issue. On the contrary, our proposed method can synthesise the ageing and rejuvenating effects realistically. Among all state-of-the-art methods, Yang *et al.* [160] is able to synthesise the most realistic effects due to the use of a multi-level feature discriminator.

Table 5.9: Face verification results in terms of accuracy (%) for the MORPH II dataset and the CACD. The query images are the original facial images, and the gallery images are the synthesised images generated by each corresponding model.

| Gallery Image | | Aging | | | Rejuvenating | | |
|---|---|---|---|---|---|---|---|
| | | S31-40 | S41-50 | S51$^{+}$ | S41-50 | S31-40 | S30$^{-}$ |
| MORPH II | Antipov *et al.* [5] | 94.46 ± 0.16 | 93.57 ± 0.12 | 91.24 ± 0.20 | 95.33 ± 0.16 | 93.54 ± 0.13 | 92.48 ± 0.27 |
| | IPCGAN [151] | 94.56 ± 0.23 | 93.87 ± 0.19 | 91.63 ± 0.22 | 94.91 ± 0.28 | 93.83 ± 0.20 | 92.21 ± 0.27 |
| | S$^2$GAN [61] | 94.88 ± 0.09 | 93.65 ± 0.17 | 91.44 ± 0.12 | 95.50 ± 0.11 | 94.72 ± 0.19 | 92.54 ± 0.18 |
| | Liu *et al.* [101] | 94.22 ± 0.28 | 93.49 ± 0.26 | 91.28 ± 0.21 | 95.63 ± 0.22 | 94.84 ± 0.23 | 93.23 ± 0.27 |
| | Li *et al.* [92] | 95.08 ± 0.11 | 93.99 ± 0.14 | 91.87 ± 0.15 | 95.40 ± 0.14 | 94.05 ± 0.16 | 92.52 ± 0.17 |
| | Yang *et al.* [160] | 94.29 ± 0.22 | 93.34 ± 0.27 | 91.18 ± 0.28 | 95.76 ± 0.21 | 94.40 ± 0.22 | 93.76 ± 0.29 |
| | Triplet | 97.87 ± 0.07 | 97.01 ± 0.09 | 94.86 ± 0.17 | 98.14 ± 0.06 | 98.23 ± 0.11 | 97.71 ± 0.14 |
| | Proposed | **99.06 ± 0.03** | **98.73 ± 0.06** | **95.58 ± 0.11** | **99.61 ± 0.03** | **99.39 ± 0.08** | **97.85 ± 0.09** |
| CACD | Antipov *et al.* [5] | 92.06 ± 0.27 | 88.46 ± 0.35 | 85.40 ± 0.56 | 92.67 ± 0.23 | 89.30 ± 0.28 | 86.24 ± 0.42 |
| | IPCGAN [151] | 92.29 ± 0.30 | 88.77 ± 0.33 | 85.22 ± 0.57 | 93.93 ± 0.25 | 89.32 ± 0.32 | 85.35 ± 0.50 |
| | S$^2$GAN [61] | 92.39 ± 0.35 | 88.94 ± 0.55 | 85.87 ± 0.59 | 93.32 ± 0.33 | 89.60 ± 0.42 | 86.29 ± 0.54 |
| | Liu *et al.* [101] | 92.25 ± 0.26 | 88.51 ± 0.32 | 85.46 ± 0.48 | 93.21 ± 0.23 | 89.50 ± 0.32 | 85.02 ± 0.47 |
| | Li *et al.* [92] | 93.33 ± 0.24 | 89.04 ± 0.38 | 85.91 ± 0.45 | 94.52 ± 0.21 | 89.47 ± 0.36 | 85.31 ± 0.39 |
| | Yang *et al.* [160] | 92.24 ± 0.29 | 88.58 ± 0.48 | 85.54 ± 0.57 | 92.80 ± 0.20 | 89.07 ± 0.39 | 86.91 ± 0.42 |
| | Triplet | 93.89 ± 0.17 | 92.73 ± 0.21 | 89.15 ± 0.24 | 94.79 ± 0.15 | 93.46 ± 0.17 | 90.31 ± 0.23 |
| | Proposed | **94.98 ± 0.10** | **94.16 ± 0.14** | **90.77 ± 0.18** | **95.08 ± 0.11** | **94.56 ± 0.14** | **91.68 ± 0.15** |

### Age category classification accuracy

Table 5.4 and 5.5 tabulate the age category classification accuracies of various methods on the synthesised images when images from the $30^-$ and $51^+$ categories are used as source images, respectively. In these tables, the *Natural Faces* row tabulates the accuracy attained when using the original facial images. Since [5] uses a relatively shallow generator compared to other works, its performance is hence below others by a significant margin. IPCGAN uses the age labels as conditions in the GAN learning process and incorporates an age category classification loss. However, due to the fact that the classification error is high (the classifier is noisy), the gradient for the age information is not accurate. As a result, although its performance is higher than that of [5], it is still lower than the one attained on the original facial images by a large margin. The recently proposed $S^2$GAN attains a higher accuracy by implementing a customised generator where each age category is associated with a decoder. The methods of Liu *et al.* [101] and Li *et al.* [92] achieve similar accuracy since both use the Wavelet transform. Among all the other evaluated methods, the one proposed by Yang *et al.* [160] achieves the best performance by using a multi-level feature discriminator. By adding a feature-level discriminator to the vanilla GAN, the baseline *w/o CDP* model achieves a comparable performance to that achieved by IPCGAN. Our proposed AOFS method outperforms all evaluated methods for the majority of age categories.

### Image Quality

The synthesis accuracy is also related to the quality of the generated images [151]. The quality and diversity of the synthesised images are usually measured in terms of the IS and the FID. IS measures the image quality and diversity by computing the KL divergence between the real and the generated class distributions. On the other hand, FID uses a multivariate Gaussian distribution to model the data distribution and the mean and the covariance from two distributions to compute their distance. Since

we use a ResNet-50 to evaluate the identity permanence capabilities (see Section 5.3.7), we rename these two metrics as the RS and the FRD. The RS and FRD are tabulated in Table 5.6 and Table 5.7, respectively, for our AOFS method and several state-of-the-art methods. Since our AOFS method can render more realistic ageing and rejuvenating effects than other evaluated methods and has stronger identity permanence capabilities, it achieves the best performance for both metrics, especially for the FRD, which is sensitive to the mode collapse issue.

#### Degree of Mode Collapse

Since our method tackles the AOFS task from the aspect of mode learning, we also measure the degree of mode collapse by computing the KL-divergence between the distribution of the synthesised images and the expected distribution. We compute this divergence for all synthesised images within each fold.

As shown in Table 5.8, the proposed AOFS method significantly outperforms the baseline model and the method in [5], which use the negative log-likelihood loss from the vanilla GAN. By using different discriminators to learn different modes, our method also achieves a lower divergence value compared to other methods that leverage the least square loss from the LSGAN.

### 5.3.6 Identity permanence

To evaluate the identity permanence on the synthesised images, we design a new baseline, the *Triplet* model. Specifically, in the *Triplet* model, we replace the Adversarial Triplet loss with the original Triplet loss to directly compare these two loss functions. The identity permanence capabilities are measured in terms of the face verification accuracy, i.e.. whether the synthesised image and the original image depict the same person. To this end, we define three input settings based on three different target age categories for each synthesis process. Specifically, the query images are the original facial images from the datasets, while the gallery images are the synthesised images that are expected to be within the target age category, as

tabulated in Table 5.9 with the column headings *S31-40*, *S41-50*, and $S51^{+}$ for the ageing process and headings *S41-50*, *S31-40*, and $S30^{-}$ for the rejuvenating process. For example, *S31-40* refers to the synthesised images expected to be within the $31-40$ category. We use the *cosine similarity* to measure the distance of each pair of query and gallery images.

As tabulated in Table 5.9, all the state-of-the-art methods achieve a similar accuracy since they all use a similar strategy, namely, minimising the distance between two identity-specific features using the L1 or L2 loss. Li *et al.* [92] slightly outperforms other methods as it uses a combination of these two losses. The subtle difference in accuracy among these methods may also be due to the quality of the images, since the identity information may be distorted in images of poor quality. By replacing the L1 or L2 loss with the Triplet loss, the identity permanence capability can be remarkably boosted by about 3 % on both datasets. Our AOFS method, which uses the Adversarial Triplet loss, reduces intra-class variances within each age category in the feature space. Consequently, it achieves the highest accuracy among all evaluated methods.

## 5.4 Conclusion

In this chapter, we tackle the Age-Oriented Face Synthesis task from the aspect of the mode learning. Specifically, we present an AOFS method that incorporates a novel Conditional Discriminator Pool to alleviate the mode collapse issue in the vanilla GAN. Our method also incorporates a novel Adversarial Triplet loss to attain strong identity permanence capabilities. By using the proposed CDP, only the target feature-level discriminator that learns the current mode is deployed, which does not increase the computational complexity during training. Our CDP then allows learning multiple modes explicitly and independently. As a result, our proposed AOFS method outperforms several state-of-the-art methods on AOFS benchmark datasets. In the future, we will investigate into improving the ageing and rejuvenating

effects by including the synthesis and removal of wrinkles and face shape manipulation among different age categories. Improving these aspects of the synthesis process is expected to further boost the synthesis accuracy and have the potential to simulate a more personalised ageing and rejuvenating process.

# Chapter 6

# Unsupervised Age-Invariant Face Recognition with Disentangled Contrastive Learning

## 6.1 Introduction

AIFR aims to recognise the identity of subjects regardless of their age and is an important yet less studied topic compared to other sub-problems of face recognition. Different from the conventional face recognition problem, AIFR needs to consider the intra-class variance caused by the age information. A robust solution for AIFR can be used in various biometrics and forensics applications like tracking a person-of-interest, such as missing children, people with dementia, or suspects over several years span [148].

Most existing AIFR methods try to solve the AIFR problem under supervised settings. However, cross-age facial images of the same subject are extremely hard to collect, as a result existing noise free age-oriented face datasets are of small size with

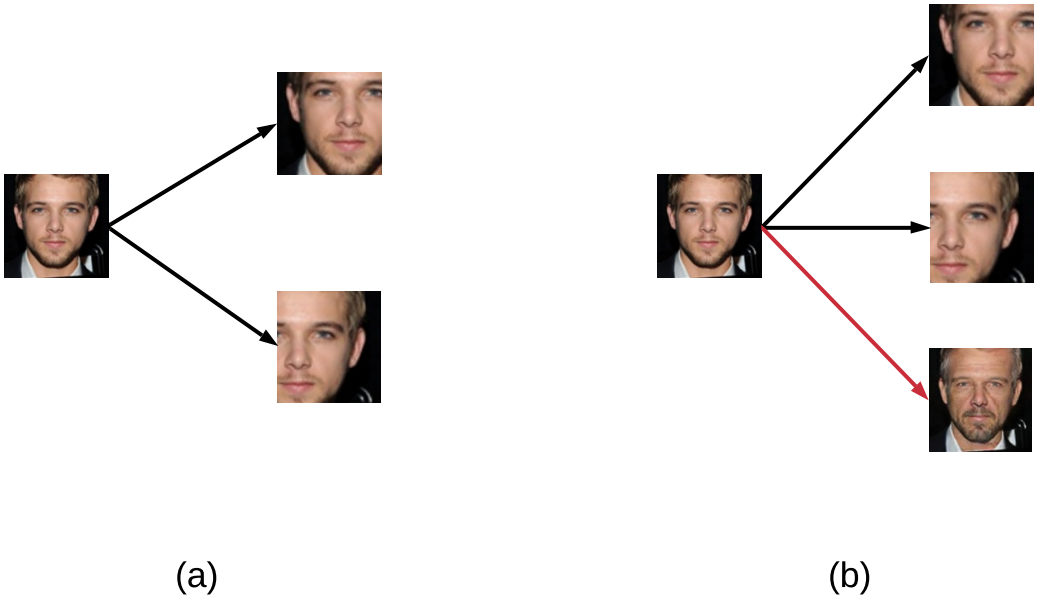


Figure 6.1: Data augmentation strategy used in (a) conventional contrastive learning, where two augmented samples are used to learn the shared features representing the identity within the input image and (b) DCL, where the additional sample is synthesised by a GAN model and used to learn age-invariant features.

limited samples per subject [28, 126]. Moreover, in real scenarios, images of the same subject at different ages are usually hard or even impossible to obtain, which yields insufficient supervised information and limits the versatility of supervised models.

Generally, existing AIFR methods can be categorised as either discriminative models or generative models [94, 152]. Discriminative models [49, 132, 152, 159] aim to learn and extract age-invariant features directly from input images while generative models [88, 119] synthesise samples that match the target age before the feature extraction. The research community usually favours the discriminative approach in light of the fact that traditional generative models are time-consuming to train, and the quality of the synthesised samples is usually unsatisfactory. Recently, many works [5, 147, 160] have demonstrated that GANs [50] can synthesise high-realistic images of subjects within different age groups, which brings researchers' attention back to the generative approach [171, 172].

In this chapter, we combine these two approaches and propose a novel method

called DCL to tackle the unsupervised AIFR problem. Specifically, we adopt the idea of contrastive learning [55] to maximise the similarity between features extracted from a pair of augmented samples from the same input image. Different from conventional contrastive learning methods [23], we use a generative model to synthesise an additional augmented sample within a different age group. By maximising the similarity among features from samples derived from the same image but within different age groups, disentangled identity features can be learned. Examples of augmented samples in conventional contrastive learning and DCL are depicted in Fig. 6.1. We further modify the conventional contrastive loss to fit this three-sample setting. The modified contrastive loss can simultaneously maximise the similarity among the set of three features and minimise the similarity between them and other samples from different images.

## 6.2 Disentangled Contrastive Learning

In this section, we explain in detail the proposed DCL by first formulating the contrastive AIFR. Then, we discuss the data augmentation process involved in our method, followed by the modified contrastive loss.

### 6.2.1 Problem Formulation

Since we aim to tackle unsupervised AIFR, we use no labels associated with input images. Given an input image $x$, we aim to obtain its disentangled identity features that are not affected by the age variation.

As shown in Fig. 6.2 (a), by using contrastive learning, a pair of augmented samples are generated through a stochastic data augmentation process, $\mathcal{T}$. The two augmented samples are considered as a positive pair and denoted as $\tilde{x}_i$ and $\tilde{x}_j$. Then, a feature extractor $f(\cdot)$ produces multi-dimensional features $h_i$ and $h_j$ from the two augmented samples. $h_i$ and $h_j$ are further fed into a projection head $g(\cdot)$ that is used to produce feature vectors $z_i$ and $z_j$. This procedure is summarised by the following

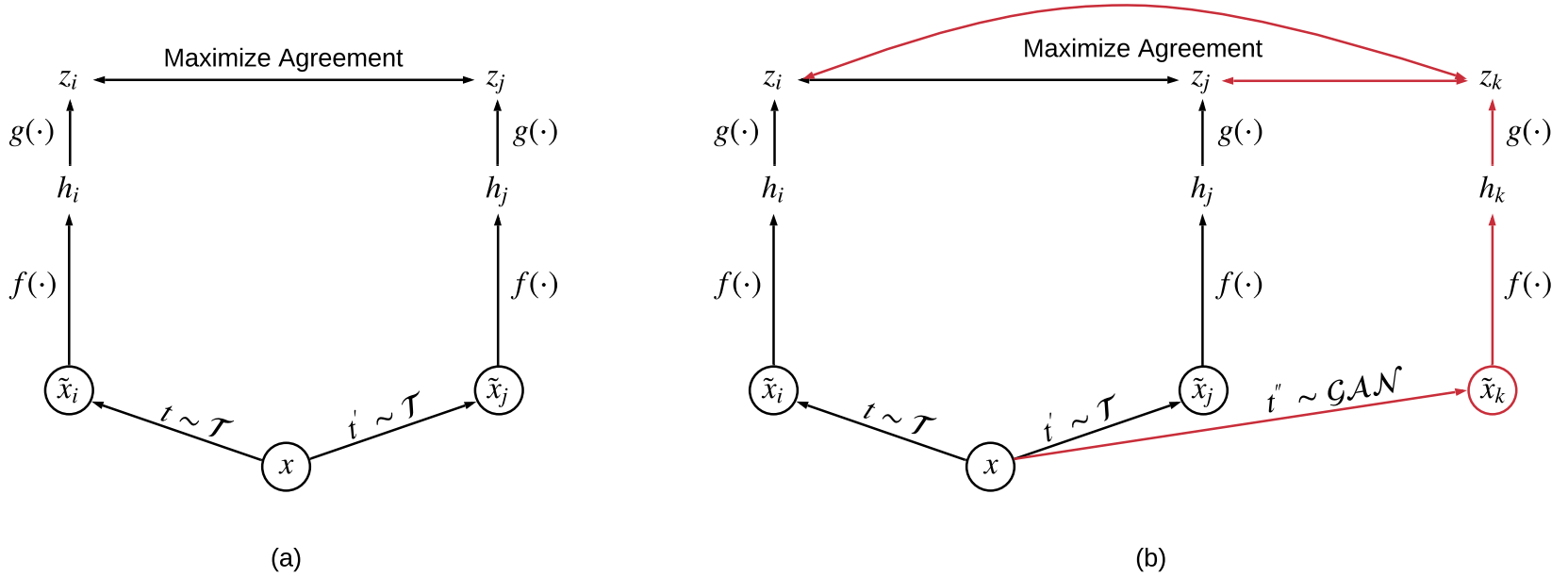


Figure 6.2: Comparison between (a) conventional contrastive learning [23] and (b) DCL. Build upon the conventional framework, DCL has an additional path (highlighted in red) used to learn disentangled identity features. Face recognition is performed using $h_i$, $h_j$, and, $h_k$ as they preserves the spatial information of the input image.

equations [23]:

$$z_i = g(f(\tilde{x}_i)), \tag{6.1}$$

and

$$z_j = g(f(\tilde{x}_j)), \tag{6.2}$$

where $f(\tilde{x}_i)$ is equivalent to $h_i$ and $f(\tilde{x}_j)$ is equivalent to $h_j$. To extract multi-dimensional features, $f(\cdot)$ is usually formulated as CNNs. To produce one dimensional feature vectors, $g(\cdot)$ is usually formulated as a stack of fully-connected layers.

The contrastive learning can learn robust identity features for a conventional face recognition task, where age variation is not considered [27]. To learn disentangled identity features, the model needs to disentangle the age features from the identity features. To this end, we leverage additional augmented samples. By maximising the similarities among features that represent the same subject but within different age groups, the model can gain disentangle capabilities and learn age-invariant features.

As shown in Fig. 6.2 (b), in DCL, the third augmented sample, $\tilde{x}_k$, is synthesised from a GAN model with features $h_k$ and $z_k$ produced by corresponding networks. The maximisation is then performed among a set of three features: $z_i$, $z_j$, and $z_k$.

### 6.2.2 Data Augmentation

For $x_i$ and $x_j$, we follow the stochastic data augmentation process in [23]. Specifically, the process consists of random cropping, a resizing operation to make the spatial dimension of the cropped image the same as the original one, random colour distortions, and random Gaussian blur. As demonstrated in [23], random cropping and random colour distortion are crucial for contrastive learning to achieve good results.

For $\tilde{x}_k$, we adopt the GAN model from [147] as depicted in Fig.6.3. The label, $\tilde{l}$, used for the GAN is randomly generated so that the DCL can utilise images within different age groups. In addition, we allow each age group to span 5 years rather

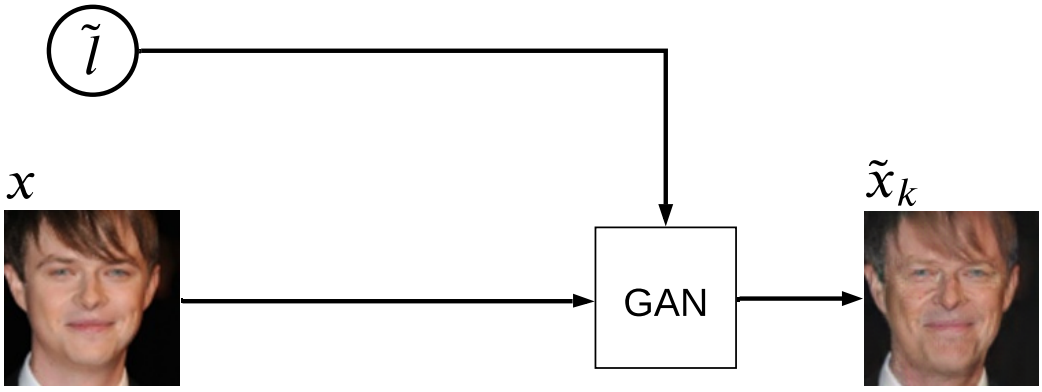


Figure 6.3: Data augmentation by using a GAN model. $\tilde{l}$ is randomly generated for the GAN model to synthesise faces within a random age group.

than the four groups used in [147]. This finer age group granularity can further increase the disentangle capabilities of the model.

### 6.2.3 Modified Contrastive Loss

As aforementioned, given a pair of feature vectors $z_i$ and $z_j$, the contrastive loss tries to maximise the similarity between them. The similarity between a pair of features is computed as:

$$sim(\tilde{z}_i, \tilde{z}_j) = \frac{\tilde{z}_i^T \cdot \tilde{z}_j}{||\tilde{z}_i||||\tilde{z}_j||}, \tag{6.3}$$

where $\cdot$ indicates dot product. $||\tilde{z}_i||$ and $||\tilde{z}_j||$ are L2 normalized feature vectors. Instead of only maximising the similarity between features representing the same subject, we also want to minimise the similarity between features extracted from other images. To this end, the normalized temperature-scaled cross-entropy loss (NT-Xent) is employed in previous works [23, 155]. The NT-Xent loss for a pair of features is formulated as:

$$\mathcal{L}_{NT-Xent}(i,j) = -log\frac{exp(\frac{sim(z_i,z_j)}{\tau})}{\sum_{b=1}^{2B} \mathbf{1}\{b \neq i\} exp(\frac{sim(z_i,z_b)}{\tau})}, \tag{6.4}$$

where $\mathbf{1}_{[n \neq i]}$ equals to 1 iff $n \neq i$, otherwise 0. $\tau$ indicates the temperature parameter [155] and $z_b$ indicates an augmented sample from other images within the same batch. Given a batch size of $B$, there are $2B$ augmented samples in conventional contrastive learning. The conventional contrastive loss can then be formulated as [23]:

$$\begin{aligned} \mathcal{L}_{contrastive} = \frac{1}{2B} \sum_{b=1}^{B} [\mathcal{L}_{NT-Xent}(2b-1, 2b) + \\ \mathcal{L}_{NT-Xent}(2b, 2b-1)]. \end{aligned} \tag{6.5}$$

Given three augmented samples, Eq. 6.4 can be modified as:

$$\mathcal{L}_{NT-Xent}(i,j,k) = -log \frac{exp(\frac{sim(z_i, z_j)}{\tau}) + exp(\frac{sim(z_i, z_k)}{\tau})}{\sum_{b=1}^{2B} \mathbf{1}\{b \neq i\} exp(\frac{sim(z_i, z_b)}{\tau})}, \tag{6.6}$$

which simultaneously maximises the similarity among a set of feature vectors, $z_i$. $z_j$. $z_k$.

With three augmented samples, there will be $3B$ samples in total in a batch and the contrastive loss in Eq.5 can therefore be modified as:

$$\begin{aligned} \mathcal{L}_{contrastive} = \frac{1}{3B} \sum_{b=1}^{B} [\mathcal{L}_{NT-Xent}(3b-2, 3b-1, 3b) + \\ \mathcal{L}_{NT-Xent}(3b-1, 3b, 3b-2) + \\ \mathcal{L}_{NT-Xent}(3b, 3b-2, 3b-1)]. \end{aligned} \tag{6.7}$$

## 6.3 Experiments

### 6.3.1 Experiment Settings

**Data Pre-processing**. We use the open-source computer vision library dlib [63] for image pre-processing. Specifically, 68 facial points are detected in each facial image to crop images based on the location of the eyes to a size of $128 \times 128$ pixels.

**Data Partition**. For the FG-Net dataset, we use the leave-one-image-out strategy as in previous works [157]. For homogeneous dataset evaluation, 1 image is

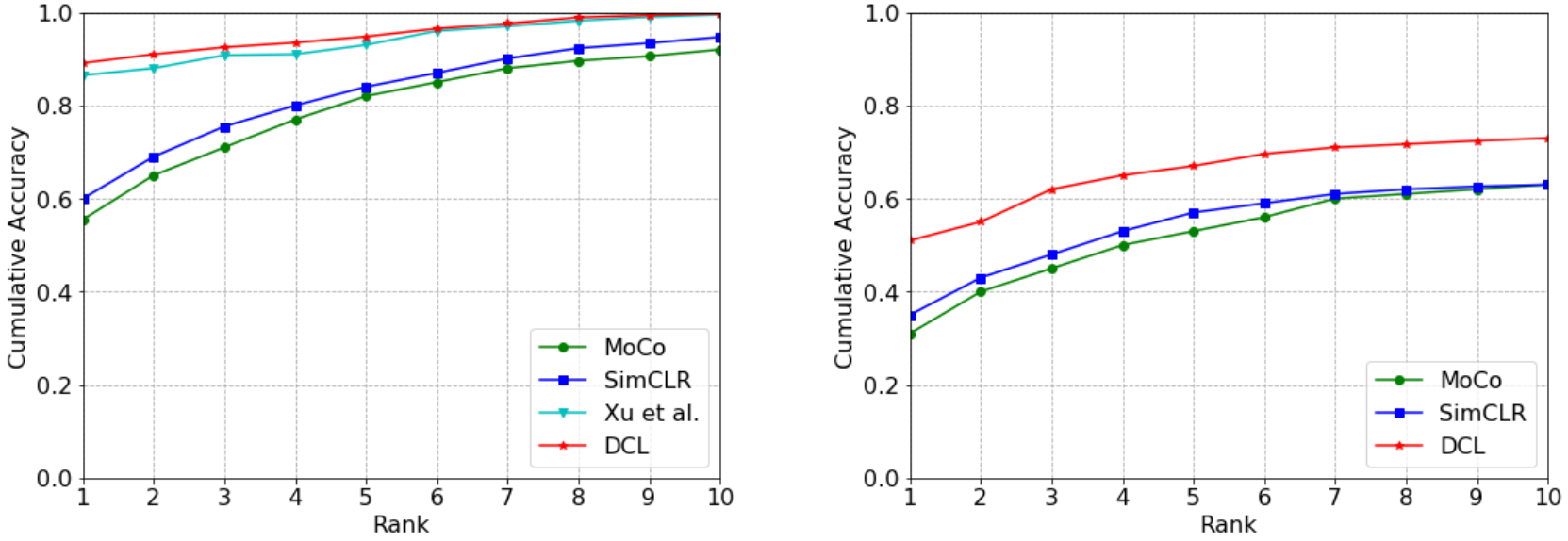


Figure 6.4: CMC curve for the FG-NET dataset. The left plot depicts the results for homogeneous-dataset evaluation, and the right plot depicts the results for cross-dataset evaluation.

Table 6.1: Rank-1 accuracy and mAP value for state-of-the-art methods on the FG-NET dataset for homogeneous-dataset evaluations.

| Method | Rank-1 | mAP |
|---|---|---|
| MoCo [59] | 55.6 | 49.2 |
| SimCLR [23] | 59.8 | 52.5 |
| Xu *et al.* [157] | 86.5 | 80.3 |
| DCL | **90.1** | **82.7** |

Table 6.2: Rank-1 accuracy and mAP value for state-of-the-art methods on the FG-NET dataset for cross-dataset evaluations.

| Method | Rank-1 | mAP |
|---|---|---|
| MoCo [59] | 31.6 | 24.3 |
| SimCLR [23] | 35.2 | 26.5 |
| DCL | **51.7** | **45.4** |

used for testing, and the remaining 1001 images are used for training. The whole process is repeated 1002 times, and the average is reported. For cross dataset evaluation, we also evaluate the model 1002 times and report the average result.

For the CACD-VS dataset, we follow the previous work [157] by preforming 10-fold cross-validation. For homogeneous-dataset evaluation, we use 9 folds for training and the remaining fold for testing.

For the MORPH II dataset, we use the partition strategy in [145, 150], where images of 10,000 subjects are used to construct the training set, and images of 3,000 subjects are used to construct the testing set. For the cross-dataset evaluation, only the testing set is used.

**Implementation Details**. We employ the ResNet-50 [58] as the function $f(\cdot)$ to extract comprehensive features from input images and a 3-layer fully-connected network to produce the features used by the NT-Xent loss. We use a batch size of 1,024 as [23] have argued that a large batch size is crucial for contrastive learning to achieve good performance except for the homogeneous-dataset evaluation on the

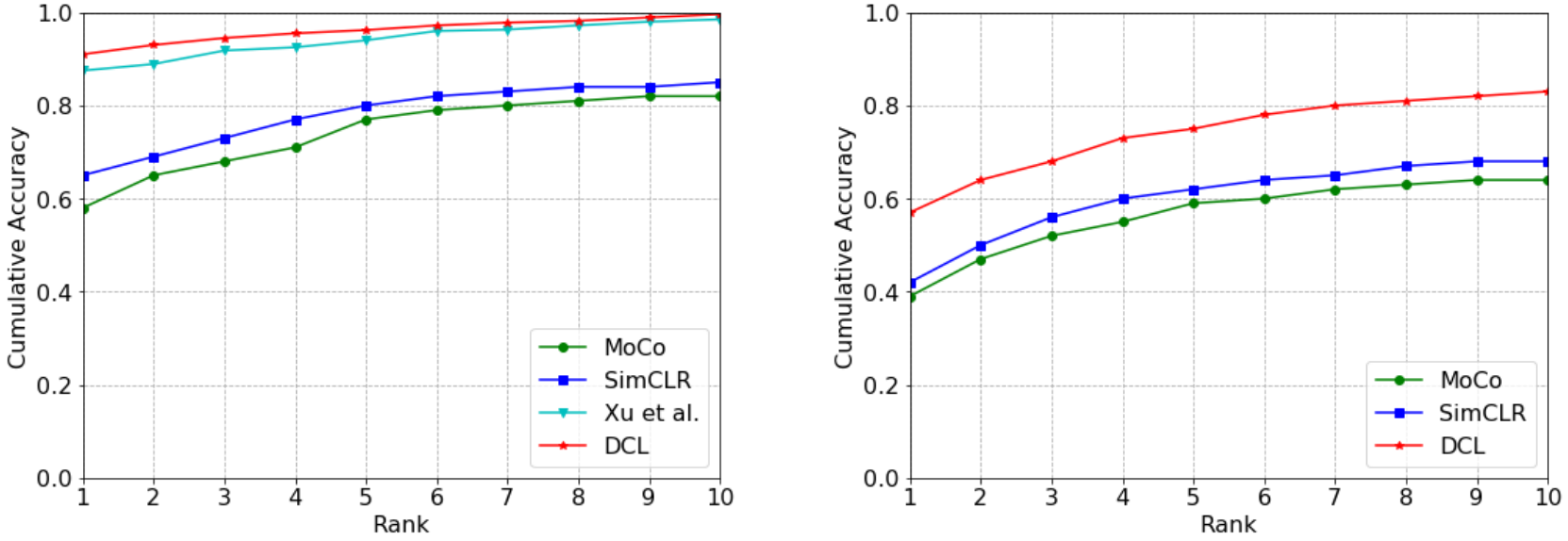


Figure 6.5: CMC curve for the MORPH II dataset. The left plot depicts the results for homogeneous dataset evaluation, and the right plot depicts the results for cross datasets evaluation.

FG-NET dataset. Additionally, the LARS optimiser [163] is utilised for multi-GPU training. In order to prevent overfitting, we use the AgeDB dataset [111] and the UTKFace dataset [170] to train the GAN model. When training DCL, parameters in the GAN model are fixed.

### 6.3.2 Comparison with State-of-the-Art Methods

The Rank-1 accuracy and the mAP value of the FG-NET dataset are tabulated in Table 1 and 2. Due to the limited number of works on unsupervised AIFR, we include two state-of-the-art unsupervised methods, MoCo [59] and SimCLR [23] for comparison. We can see that DCL dramatically outperforms these two unsupervised methods since they do not consider the age variation in facial images. Our method also outperforms the state-of-the-art unsupervised method [157] by a large margin under the two metrics. The CMC curve for homogeneous-dataset evaluations and cross-dataset evaluations are depicted in Fig. 6.4.

The Rank-1 accuracy and the mAP value of the MORPH II dataset is tabulated in Table 3 and 4. The CMC curves of the MORPH II dataset are depicted in Fig. 6.5. Again, thanks to the additional augmented sample synthesised by the GAN model, DCL can explicitly disentangle the age features representing different age groups from the identity features, which yields age-invariant features. We report the Rank-1 accuracy on the CACD-VS dataset in Table 5 with comparisons to the human performance. DCL outperforms the human average performance by about 8% and has a comparable performance with the human voting performance, where decisions from multiple participants are combined.

## 6.4 Conclusion

In this chapter, we proposed the DCL for unsupervised AIFR. Compared to previous contrastive learning works, our method utilises an additional augmented sample generated by a GAN to force the method to maximise the similarities among features

Table 6.3: Rank-1 accuracy and mAP value for state-of-the-art methods on the MORPH II dataset for homogeneous dataset evaluations.

| Method | Rank-1 | mAP |
|---|---|---|
| MoCo [59] | 58.9 | 44.8 |
| SimCLR [23] | 65.8 | 50.1 |
| Xu *et al.* [157] | 87.5 | 78.0 |
| DCL | **91.5** | **79.4** |

Table 6.4: Rank-1 accuracy and mAP value for state-of-the-art methods on the MORPH II dataset for cross datasets evaluations.

| Method | Rank-1 | mAP |
|---|---|---|
| MoCo [59] | 39.7 | 27.4 |
| SimCLR [23] | 42.2 | 29.3 |
| DCL | **57.6** | **48.1** |

Table 6.5: Rank-1 accuracy and mAP value for state-of-the-art methods on the CACD-VS dataset for homogeneous-dataset evaluations.

| Method | Rank-1 |
|---|---|
| Human, Average | 85.7 |
| Human, Voting (2015) | 94.2 |
| MoCo [59] | 83.3 |
| SimCLR [23] | 85.7 |
| Xu *et al.* [157] | 92.3 |
| DCL | **93.9** |

from the facial images of the same subject within different age groups. Differently from previous unsupervised AIFR methods, DCL merges a discriminative approach and a generative approach together for stronger feature disentangling capabilities. In addition, a modified contrastive loss for three augmented samples is proposed. Based on evaluations on several AIFR benchmark datasets, DCL dramatically outperforms both state-of-the-art unsupervised AIFR methods and contrastive learning methods. Since this work only focuses on the AIFR task, in the future, we will apply DCL to other tasks to explore the versatility of the method.

# Chapter 7

# Conclusions

In this thesis, we focus on developing deep learning-based methods for age-related facial analysis. We first discussed all three age-related facial analysis tasks and our motivations. We then analysed the shortcomings of existing methods and proposed methods from these perspectives. Specifically, we proposed two methods for age estimation that utilising age-specific facial patches while most existing works pay no attention to these informative regions. Then, we proposed a method for AOFS that aims to learn independent modes, which is not achievable by using a vanilla GAN that is widely used as the backbone model in existing AOFS works. Last but not least, we proposed a method to study the understudied but important unsupervised AIFR problem.

## 7.1 Contributions and conclusions

We summarise our main contributions as follows.

In Chapter 3, we proposed a customised CNN architecture called FusionNet for age estimation. Apart from the whole facial image, the FusionNet successively takes several age-specific facial patches as part of the input to emphasise the age-specific features. The age-specific facial patches are discovered by using the BIF and Adaboost algorithm. This work is the first deep learning-based method in which

the learning of age-specific features is enhanced. Experimental results showed that leveraging age-specific facial patches as inputs to the network is more robust than using dominant facial attributes like the eyes and nose that is widely adopted by existing works.

In Chapter 4, we proposed a modified method called ADPF for the age estimation problem. The ADPF aims to reduce the training complexity of the previous method by replacing the BIF and Adaboost algorithm with an AttentionNet that contains a novel hybrid attention mechanism. The hybrid attention leverages the merits from the multi-head self-attention mechanism and the channel-wise attention mechanism and produces multiple single-channel attention maps with each highlights one particular age-specific facial patches. As a result, the training time is reduced from 70 hours from the previous method to 25 hours with a boosted performance. We also conducted experiments on more datasets under additional settings to show the versatility of this method.

In Chapter 5, we proposed a GAN model with a CDP to achieve high synthesis accuracy for AOFS and an Adversarial Triplet loss to ensure the identity information is unaltered in the synthesised image. This method aims to alleviate the mode collapse issue in the vanilla GAN by using different discriminator to learn a particular mode. The discriminator is selected by the target label. Experimental results showed that CDP can alleviate the mode collapse issue to a great extent and achieve a higher synthesis accuracy than other state-of-the-art methods. The Adversarial Triplet loss aims to reduce the intra-class variations caused by age information in each identity cluster. A toy example on the MNIST dataset demonstrated that the Adversarial Triplet loss yields highly compact clusters with dramatically reduced intra-class variations. Experiments on age-oriented datasets also showed its superior performance compared to other identity preserving losses.

In Chapter 6, we proposed the DCL to tackle the understudied unsupervised AIFR problem. Most existing works study the supervised AIFR problem. However, cross-age facial images are not often collectable, which limits the implantation

and deployment of these supervised learning methods. Different from previous unsupervised AIFR method which requires input pairs for training, DCL only require a single image as the input and utilise advanced data augmentation processes to learn the constant features of the representation of the subject in the input image. In addition to the existing data augmentation methods, we use the GAN model that is proposed in Chapter 5 to synthesis facial images within different age groups to learn age-invariant features. Experimental results show that DCL outperforms both existing contrastive learning methods and unsupervised AIFR methods under widely used evaluation metrics on several benchmark datasets.

## 7.2 Future research directions

This thesis only tackles age-related facial analysis from some particular perspectives. Some other directions can be taken to further improve the deep learning-based methods. Some possible directions for each problem are as follows.

### 7.2.1 Age estimation

Although deep learning-based age estimators have achieved much better results than models that use traditional machine learning methods, there are still some issues that have not been addressed yet. First, existing age-oriented datasets like the MORPH II dataset and the FG-NET dataset involve other variations like PIE and occlusion. With these unexpected factors, extracting age-specific features is onerous. [3] shows that the expression can downgrade the performance of the age estimation models, and proposes a graphical model to tackle the expression-invariant age estimation problem. Such disentangled age estimation problem has not been studied by using a CNN yet, which could be a possible future research trend.

Another possible topic is to build large-scale noise-free datasets. Recent datasets for face recognition have several millions of training samples [13, 54]. However, the largest noise-free dataset for age estimation (the MORPH II dataset) has only

40,000 to 50,000 images for training based on different data partition strategies. Therefore, a larger noise-free dataset is needed to help to boost the age estimation performance further.

### 7.2.2 Age-Oriented Face Synthesis

The most important topic that none of the above works cover is standardising the evaluation methods of age synthesis models. Early attempts [5, 170] mainly use subjective evaluation methods by taking surveys. Recent works [151, 160] evaluate their model based on the two criteria mentioned in Section 3.2, but they use different evaluation models. Specifically, [160] uses a commercial face recognition and age estimation tool, while [151] uses their pre-trained face recognition and age estimation model. Such differences make related works hard to compare, which may hinder the development of further research.

Moreover, existing AOFS methods use a pre-trained face recognition model or an age estimation model to guide the training process. However, those models may be noisy. According to [151], the age estimation accuracy of their age estimator is only about 30%. Due to the fact that the classification error is high (the classifier is noisy), the gradient for the age information is not accurate. The performance can then be boosted by developing other methods to guarantee the synthesis accuracy and keep the identity information simultaneously. New methods could also make the whole training process end-to-end instead of pre-training several separate networks, which can save training time and computational resources.

### 7.2.3 Age-Invariant Face Recognition

Although recent AIFR models can attain good results, these results could be further improved if larger age-oriented datasets are available for training and testing. Instead of building the dataset from the ground up, age synthesis methods can be used to enlarge and augment existing datasets by generating the images of each subject at different ages or age groups. As a result, the training process could benefit from

more training samples, and higher accuracy could be achieved.

# Bibliography


[1] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1643–1652, 2017.

[2] S Akazaki, H Nakagawa, H Kazama, O Osanai, M Kawai, Y Takema, and G Imokawa. Age-related changes in skin wrinkles assessed by a novel three-dimensional morphometric analysis. *British Journal of Dermatology*, 147(4): 689–695, 2002.

[3] Fares Alnajar, Zhongyu Lou, José Manuel Álvarez, Theo Gevers, et al. Expression-invariant age estimation. In *British Machine Vision Conference (BMVC)*, 2014.

[4] Marcus Angeloni, Rodrigo de Freitas Pereira, and Helio Pedrini. Age estimation from facial parts using compact multi-stream convolutional neural networks. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 0–0, 2019.

[5] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017.

[6] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2017.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223, 2017.

[8] Yosuke Bando, Takaaki Kuratate, and Tomoyuki Nishita. A simple method for modeling wrinkles on human skin. In *10th Pacific Conference on Computer Graphics and Applications, 2002. Proceedings.*, pages 166–175. IEEE, 2002.

[9] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3286–3295, 2019.

[10] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[11] Laurence Boissieux, Gergo Kiss, Nadia Magnenat Thalmann, and Prem Kalra. Simulation of skin aging and wrinkles with cosmetics insight. In *Computer Animation and Simulation 2000*, pages 15–27. Springer, 2000.

[12] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

[13] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 67–74. IEEE, 2018.

[14] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Consistent rank logits for ordinal regression with convolutional neural networks. *arXiv preprint arXiv:1901.07884*, 2019.

[15] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *IEEE Conference on*

*Computer Vision and Pattern Recognition (CVPR)*, pages 585–592. IEEE, 2011.

[16] Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.

[17] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2017.

[18] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision (ECCV)*, pages 768–783. Springer, 2014.

[19] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2467–2474, 2013.

[20] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017.

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

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

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[24] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 403–412, 2017.

[25] Yiliang Chen, Zichang Tan, Alex Po Leung, Jun Wang, and Jianguo Zhang. Multi-region ensemble convolutional neural networks for high accuracy age estimation. In *British Machine Vision Conference*, 2017.

[26] Yiliang Chen, Shengfeng He, Zichang Tan, Chu Han, Guoqiang Han, and Jing Qin. Age estimation via attribute-region association. *Neurocomputing*, 367: 346–356, 2019.

[27] Yixian Cheng and Haiyang Wang. A modified contrastive loss method for face recognition. *Pattern Recognition Letters*, 125:785–790, 2019.

[28] T Cootes and A Lanitis. The fg-net aging database, 2008.

[29] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[30] Imre Csiszár, Paul C Shields, et al. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4): 417–528, 2004.

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

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

*Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[33] Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via $\chi$ upper bound minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2732–2741, 2017.

[34] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 766–774, 2014.

[35] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

[36] Gareth J Edwards, Andreas Lanitis, Christopher J Taylor, and Timothy F Cootes. Statistical models of face images-improving specificity. *Image and Vision Computing*, 16(3):203–211, 1998.

[37] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1–9, 2015.

[38] William Fedus, Mihaela Rosca, Balaji Lakshminarayanan, Andrew M Dai, Shakir Mohamed, and Ian Goodfellow. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

[39] Songhe Feng, Congyan Lang, Jiashi Feng, Tao Wang, and Jiebo Luo. Human facial age estimation by cost-sensitive label ranking and trace norm regularization. *IEEE Transactions on Multimedia*, 19(1):136–148, 2016.

[40] Yun Fu, Guodong Guo, and Thomas S Huang. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1955–1976, 2010.

[41] Dennis Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[42] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

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

[44] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM International Conference on Multimedia*, pages 307–316. ACM, 2006.

[45] Xin Geng, Zhi-Hua Zhou, and Kate Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 29(12):2234–2240, 2007.

[46] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. *IEEE transactions on pattern analysis and machine intelligence*, 35(10):2401–2412, 2013.

[47] Angelo Genovese, Vincenzo Piuri, and Fabio Scotti. Towards explainable face aging with generative adversarial networks. In *IEEE International Conference on Image Processing (ICIP)*, pages 3806–3810. IEEE, 2019.

[48] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser. A statistical model for synthesis of detailed facial geometry. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1025–1034. ACM, 2006.

[49] Dihong Gong, Zhifeng Li, Dahua Lin, Jianzhuang Liu, and Xiaoou Tang. Hidden factor analysis for age invariant face recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2872–2879, 2013.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[51] Guodong Guo and Guowang Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR 2011*, pages 657–664. IEEE, 2011.

[52] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[53] Guodong Guo, Guowang Mu, Yun Fu, and Thomas S Huang. Human age estimation using bio-inspired features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 112–119, 2009.

[54] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)*, pages 87–102. Springer, 2016.

[55] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742. IEEE, 2006.

[56] Hu Han, Charles Otto, Xiaoming Liu, and Anil K Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161, 2015.

[57] Hu Han, Anil K Jain, Fang Wang, Shiguang Shan, and Xilin Chen. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2597–2609, 2018.

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

[59] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[60] Yating He, Min Huang, Qinghai Miao, Haiyun Guo, and Jinqiao Wang. Deep embedding network for robust age estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 1092–1096. IEEE, 2017.

[61] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9440–9449, 2019.

[62] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[63] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[64] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[65] Hui-Lan Hsieh, Winston Hsu, and Yan-Ying Chen. Multi-task learning for face identification and attribute estimation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2981–2985. IEEE, 2017.

[66] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7132–7141, 2018.

[67] Zhenzhen Hu, Yonggang Wen, Jianfeng Wang, Meng Wang, Richang Hong, and Shuicheng Yan. Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097, 2016.

[68] Zhenzhen Hu, Yonggang Wen, Jianfeng Wang, Meng Wang, Richang Hong, and Shuicheng Yan. Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097, 2017.

[69] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5375–5384, 2016.

[70] Dong Huang, Longfei Han, and Fernando De la Torre. Soft-margin mixture of regressions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6532–6540, 2017.

[71] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[72] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

[73] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[74] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Soft biometric traits for personal recognition systems. In *Biometric authentication*, pages 731–738. Springer, 2004.

[75] Anil K Jain, Arun A Ross, and Karthik Nandakumar. *Introduction to biometrics*. Springer Science & Business Media, 2011.

[76] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *International Conference on Learning Representations (ICLR)*, 2019.

[77] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations (ICLR)*, 2018.

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

[79] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014.

[80] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 971–980, 2017.

[81] Brendan Klare and Anil K Jain. Face recognition across time lapse: On learning feature subspaces. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2011.

[82] Joshua C Klontz and Anil K Jain. A case study on unconstrained facial recognition using the boston marathon bombings suspects. *Michigan State University, Tech. Rep*, 119(120):1, 2013.

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[84] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[85] Young Ho Kwon and Niels da Vitoria Lobo. Age classification from facial images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 762–767, 1994.

[86] JM Lagarde, C Rouvrais, and D Black. Topography and anisotropy of the skin surface with ageing. *Skin Research and Technology*, 11(2):110–119, 2005.

[87] Michelle Lai, Ipek Oruç, and Jason JS Barton. The role of skin texture and facial shape in representations of age and identity. *Cortex*, 49(1):252–265, 2013.

[88] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[89] Gil Levi and Tal Hassner. Age and gender classification using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 34–42, 2015.

[90] Kai Li, Junliang Xing, Chi Su, Weiming Hu, Yundong Zhang, and Stephen Maybank. Deep cost-sensitive and order-preserving feature learning for cross-population age estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2018.

[91] Ling Li and Hsuan-Tien Lin. Ordinal regression by extended binary classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 865–872, 2007.

[92] Peipei Li, Yibo Hu, Ran He, and Zhenan Sun. Global and local consistent wavelet-domain age synthesis. *IEEE Transactions on Information Forensics and Security*, 2019.

[93] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145–1154, 2019.

[94] Zhifeng Li, Unsang Park, and Anil K Jain. A discriminative model for age invariant face recognition. *IEEE transactions on information forensics and security*, 6(3):1028–1037, 2011.

[95] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10): 4394–4412, 2006.

[96] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[97] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Group-aware deep feature learning for facial age estimation. *Pattern Recognition*, 66:82–94, 2017.

[98] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, 13(2):292–305, 2017.

[99] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Ordinal deep feature learning for facial age estimation. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 157–164. IEEE, 2017.

[100] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Label-sensitive deep metric learning for facial age estimation. *IEEE Transactions on Information Forensics and Security*, 13(2):292–305, 2018.

[101] Yunfan Liu, Qi Li, and Zhenan Sun. Attribute-aware face aging with wavelet-based generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11877–11886, 2019.

[102] Zicheng Liu, Zhengyou Zhang, and Ying Shan. Image-based surface detail transfer. *IEEE Computer Graphics and Applications*, 24(3):30–35, 2004.

[103] Jiwen Lu, Venice Erin Liong, and Jie Zhou. Cost-sensitive local binary feature learning for facial age estimation. *IEEE Transactions on Image Processing*, 24 (12):5356–5368, 2015.

[104] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821. IEEE, 2017.

[105] Xudong Mao, Qing Li, Haoran Xie, Raymond Yiu Keung Lau, Zhen Wang, and Stephen Paul Smolley. On the effectiveness of least squares generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[106] Leonard S Mark and James T Todd. The perception of growth in three dimensions. *Attention, Perception, & Psychophysics*, 33(2):193–196, 1983.

[107] Leonard S Mark, James T Todd, and Robert E Shaw. Perception of growth: A geometric analysis of how different styles of change are distinguished. *Journal of Experimental Psychology: Human Perception and Performance*, 7(4):855, 1981.

[108] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2017.

[109] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[110] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6707–6717, 2020.

[111] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–59, 2017.

[112] Shigeru Mukaida and Hiroshi Ando. Extraction and manipulation of wrinkles and spots for facial image synthesis. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 749–754. IEEE, 2004.

[113] Tu Nguyen, Trung Le, Hung Vu, and Dinh Phung. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2670–2680, 2017.

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

[115] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 271–279, 2016.

[116] Gokhan Ozbulak, Yusuf Aytar, and Hazim Kemal Ekenel. How transferable are cnn-based features for age and gender classification? In *Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the*, pages 1–6. IEEE, 2016.

[117] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5285–5294, 2018.

[118] Evangelia Pantraki and Constantine Kotropoulos. Face aging as image-to-image translation using shared-latent space generative adversarial networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 306–310. IEEE, 2018.

[119] Unsang Park, Yiying Tong, and Anil K Jain. Age-invariant face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):947–954, 2010.

[120] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Machine Vision Conference (BMVC)*, volume 1, page 6, 2015.

[121] Niki Parmar, Prajit Ramachandran, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 68–80, 2019.

[122] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[123] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can

be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302): 157–175, 1900.

[124] Narayanan Ramanathan and Rama Chellappa. Modeling age progression in young faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 387–394. IEEE, 2006.

[125] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 17–24. IEEE, 2017.

[126] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 341–345. IEEE, 2006.

[127] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1999.

[128] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, pages 1–14, 2016.

[129] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Some like it hot-visual guidance for preference prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5553–5561, 2016.

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

[131] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A uni-

fied embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[132] M Saad Shakeel and Kin-Man Lam. Deep-feature encoding-based discriminative model for age-invariant face recognition. *Pattern Recognition*, 93:442–457, 2019.

[133] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018.

[134] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2304–2313, 2018.

[135] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan Loddon Yuille. Deep differentiable random forests for age estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[136] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, Shuicheng Yan, et al. Personalized age progression with bi-level aging dictionary learning. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):905–917, 2018.

[137] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[138] Diana Sungatullina, Jiwen Lu, Gang Wang, and Pierre Moulin. Multiview discriminative learning for age-invariant face recognition. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[139] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Ra-

binovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[140] Shahram Taheri and Önsen Toygar. On the use of dag-cnn architecture for age estimation with multi-stage features fusion. *Neurocomputing*, 2018.

[141] Shahram Taheri and Önsen Toygar. On the use of dag-cnn architecture for age estimation with multi-stage features fusion. *Neurocomputing*, 329:300–310, 2019.

[142] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[143] James T Todd, Leonard S Mark, Robert E Shaw, and John B Pittenger. The perception of human growth. *Scientific american*, 242(2):132–145, 1980.

[144] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[145] Hao Wang, Dihong Gong, Zhifeng Li, and Wei Liu. Decorrelated adversarial learning for age-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3527–3536, 2019.

[146] Haoyi Wang, Xingjie Wei, Victor Sanchez, and Chang-Tsun Li. Fusion network for face-based age estimation. In *IEEE International Conference on Image Processing (ICIP)*, pages 2675–2679. IEEE, 2018.

[147] Haoyi Wang, Victor Sanchez, and Chang-Tsun Li. Age-oriented face synthesis with conditional discriminator pool and adversarial triplet loss. *arXiv preprint arXiv:2007.00792*, 2020.

[148] Haoyi Wang, Victor Sanchez, Wanli Ouyang, and Chang-Tsun Li. Using age

information as a soft biometric trait for face image analysis. In *Deep Biometrics*, pages 1–20. Springer, 2020.

[149] Xiaolong Wang, Rui Guo, and Chandra Kambhamettu. Deeply-learned feature for age estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 534–541. IEEE, 2015.

[150] Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. *arXiv preprint arXiv:1810.07599*, 2018.

[151] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7947, 2018.

[152] Yandong Wen, Zhifeng Li, and Yu Qiao. Latent factor guided convolutional neural networks for age-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4893–4901, 2016.

[153] Yin Wu, Nadia Magnenat Thalmann, and Daniel Thalmann. A dynamic wrinkle model in facial animation and skin ageing. *The journal of visualization and computer animation*, 6(4):195–205, 1995.

[154] Yin Wu, Prem Kalra, Laurent Moccozet, and Nadia Magnenat-Thalmann. Simulating wrinkles and skin aging. *The visual computer*, 15(4):183–198, 1999.

[155] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742, 2018.

[156] Min Xia, Xu Zhang, Liguo Weng, Yiqing Xu, et al. Multi-stage feature constraints learning for age estimation. *IEEE Transactions on Information Forensics and Security*, 15:2417–2428, 2020.

[157] Chenfei Xu, Qihe Liu, and Mao Ye. Age invariant face recognition and retrieval by coupled auto-encoder networks. *Neurocomputing*, 222:62–71, 2017.

[158] Juefei Xu, Khoa Luu, Marios Savvides, Tien D Bui, and Ching Y Suen. Investigating age invariant face recognition based on periocular biometrics. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7. IEEE, 2011.

[159] Hongyu Yang, Di Huang, and Yunhong Wang. Age invariant face recognition based on texture embedded discriminative graph model. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[160] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31–39, 2018.

[161] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1092–1099, 2018.

[162] Dong Yi, Zhen Lei, and Stan Z Li. Age estimation by multi-scale convolutional network. In *Asian Conference on Computer Vision (ACCV)*, pages 144–158. Springer, 2014.

[163] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[164] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833. Springer, 2014.

[165] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 7354–7363, 2019.

[166] Yu Zhang and Dit-Yan Yeung. Multi-task warped gaussian process for personalized age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2622–2629. IEEE, 2010.

[167] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision (ECCV)*, pages 286–301, 2018.

[168] Zhaoyu Zhang, Mengyan Li, and Jun Yu. On the convergence and mode collapse of gan. In *SIGGRAPH Asia 2018 Technical Briefs*, page 21. ACM, 2018.

[169] Zhaoyu Zhang, Mengyan Li, and Jun Yu. D2pggan: Two discriminators used in progressive growing of gans. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3177–3181. IEEE, 2019.

[170] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[171] Jian Zhao, Yu Cheng, Yi Cheng, Yang Yang, Fang Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9251–9258, 2019.

[172] Shuyang Zhao, Jianwu Li, and Jiaxing Wang. Disentangled representation learning and residual gan for age-invariant face verification. *Pattern Recognition*, 100:107097, 2020.

[173] Tianyue Zheng, Weihong Deng, and Jiani Hu. Age estimation guided convolutional neural network for age-invariant face recognition. In *IEEE Conference*

*on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 12–16, 2017.

[174] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.