UNIVERSIDADE D
COIMBRA

João Francisco Gomes Tremoço

Improving deep learning face recognition for ID and travel document applications with quality assessment

VOLUME 1

Thesis submitted to the University of Coimbra in fulfilment of the requirements of the Master's Degree in Physics Engineering under the scientific supervision of Ph.D. Nuno Gonçalves, MsC Iurii Medvedev and presented to the Physics Department of the Faculty of Sciences and Technology of the University of Coimbra.

October 2021

University of Coimbra

Integrated Master in Physics Engineering

Improving deep learning face recognition for ID and travel document applications with quality assessment

João Francisco Gomes Tremoço

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra in fulfilment of the requirements for the Master's Degree in Physics Engineering

Supervisors:

Ph.D. Nuno Gonçalves

MsC Iurii Medvedev

INSTITUTO DE SISTEMAS E ROBÓTICA
UNIVERSIDADE DE COIMBRA

Coimbra, 2021

Acknowledgements

First of all, I would like to thank Prof. Dr. Nuno Gonçalves for the opportunity to develop this project and to share the work in the parallel published paper. Thank you for the guidance and help throughout this project.

To Iurii, thank you so much for all your patience and suggestions during the developed work. Your advice, guidance and availability to advise me whenever a question arose or a frustrating problem appeared truly helped me, and I'm incredibly grateful for that.

I also would like to thank Instituto de Sistemas e Robótica - Coimbra for providing with the resources required to develop this project.

To all my friends, thank you for all the moments we've had so far and the ones we will have in the future. These five years were probably the best in my life and made me grow so much and become the person I am today. Thank you for all the magnificent memories.

Thank you, Rita, for all the support, advice, patience and for constantly pushing me to do better and to do what I like. Thank you for loving and supporting me through this so important phase of my life; you make me a better person every day.

Last but definitely not least, thank you to all my family, my grandparents, cousins. Especially, thank you to my mom, dad and sister for your love, guidance, affection every single day. For the sacrifices you made for me and for supporting me during this journey, I will always be so grateful.

Resumo

Os métodos atuais de reconhecimento facial são baseados em redes neuronais que requerem grandes quantidades de dados para serem eficazes. Os grandes conjuntos de dados disponíveis publicamente são, em sua maioria, coleções de imagens de celebridades sem restrições. Estes conjuntos de dados não são otimizados para aplicações relacionadas com a segurança de documentos. Além disso, devido a questões de privacidade, os conjuntos de dados de imagens faciais úteis para o uso em situações que envolvem documentos de identificação são pequenos e difícil acesso. Este cenário não é favorável e há espaço para otimização. Neste trabalho, uma nova abordagem de reconhecimento facial com foco na mitigação deste problema é proposta. Foi elaborada uma estratégia para incluir a qualidade das amostras numa função de perda de margem angular, a fim de otimizar o processo de treino para o cenário de documentos identificação e viagem. Isto foi conseguido alterando o parâmetro de margem na função de perda ArcFace, para um valor adaptativo que depende da qualidade de cada amostra. A margem adaptativa foi formulada de forma a aumentar com o aumento da qualidade da amostra e, como tal, aumentar o valor da perda. Para caracterizar a qualidade da amostra, cinco diferentes métricas de qualidade relacionadas com padrões da ICAO para imagens em documentos de viagem foram usadas: *Blur*, BRISQUE, FaceQNet, qualidade de iluminação facial e qualidade da pose. Três *benchmarks* específicos foram criados para testar o desempenho do método desenvolvido em diferentes cenários: sem restrições, com restrições e com restrições estritas. Com os *benchmarks* criados, o método desenvolvido foi testado e comparado com as funções de perda ArcFace e Softmax. As experiências realizadas mostraram que o método de margem adaptativa desenvolvido é superior à função de perda de margem angular (ArcFace) para o cenário de documentos de identificação. Mais especificamente, o modelo baseado na qualidade da iluminação facial provou ter o melhor desempenho nos cenários com restrições e com restrições estritas de acordo com as métricas FNMR @ FMR. Os resultados também indicam uma superioridade do método no reconhecimento facial sem restrições. modelo baseado no *blur* apresenta os melhores resultados nestas condições. Também foram testados modelos com combinações de métricas de qualidade. Estes não provaram ser superiores aos modelos que só utilisaram uma métrica, no entanto, foi obtido um resultado mais regular entre cenários.

Abstract

Current face recognition methods are based on deep neural networks that require large amounts of data to be effective. The large datasets publicly available are mostly collections of wild celebrity face images. These datasets are not optimised for document security-related applications. Moreover, due to privacy concerns, ID-compliant face image datasets are small and hardly accessible. This scenario is not favourable, and there is room for optimisation. In this work, a novel face recognition approach focused on the mitigation of this problem is proposed. A strategy was devised to include sample quality in an angular margin loss function in order to optimise the training process for the scenario of ID and Travel documents. This was achieved by changing the margin parameter in ArcFace to an adaptive value dependant on each sample's quality. The adaptive margin was formulated in such a way to increase with the increase in sample quality and as such, increase the loss value. To characterise sample quality, five different quality metrics closely related to ICAO standards were used: Blur, BRISQUE, FaceQNet, Face Illumination Quality and Pose Quality. Three specific benchmarks were designed to test the method's performance across different scenarios: Unconstrained, constrained and strictly constrained. With the designed benchmarks, the developed method was tested and compared with the ArcFace and Softmax losses. Experiments made show that the adaptive margin method developed is superior to the standard angular margin loss function (ArcFace) for the ID-compliant scenario. More specifically, the face illumination quality based model proved to better perform in the constrained and strictly scenarios according to FNMR@FMR metrics. The results also indicate a superiority of the method in unconstrained face recognition, namely the blur score model shows the best results. Models with combinations of scores were also tested. They did not prove to be superior to the single score models, however a more regular result across benchmarks was achieved.

List of Acronyms

CNN convolutional neural network

DCNN deep convolutional neural network

DET Detection Error Tradeoff

EER Equal Error Rate

FIQA Face Image Quality Assessment

FMR False Match Rate

FNIR False Negative Identification Rate

FNMR False Non-Match Rate

FPIR False Positive Identification Rate

FPR False Positive Rate

FR Face Recognition

HOG histogram oriented gradient

ICAO International Civil Aviation Organization

ID identity documents

ISR Instituto de Sistemas e Robótica

LFW Labelled Faces in the Wild

LMCL large margin cosine loss

ML machine learning

PFE Probabilistic Face Embeddings

ROC Receiver Operating Characteristic

TPR True Positive Rate

YTF YouTube faces

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter, the context and motivation of the thesis will be explained as well as the outline for the rest of the document.

1.1 Contextualisation

No two people are the same. People have a set of behavioural and biological characteristics that are unique and specific to themselves. For this reason, we as a species can distinguish a friend from a foe.

From a strict recognition view, the biological features are easier to capture and store as information than behavioural ones. Thus, to devise a generic recognition system, biological features like fingerprints, height, weight, face attributes, iris size and colour or ear shape are preferred over behavioural ones like walking, speaking mannerisms, signature or writing style, laugh, etc. Indeed this is the case for most recognition systems. For example, identity cards usually contain a face photo, some fingerprint information and height.

For recognition purposes, one of the most researched, important and useful biological human features is the face. The face contains an abundance of features that are discriminative and rich enough to distinguish one's identity. Unlike fingerprints, iris, voice or other biometric factors, the facial features can be easily extracted from unconstrained scenarios and in a non-intrusive way. This results in the face being one of the most appropriate and useful biometric data types for various applications: authentication, home security, border control, surveillance and others [1].

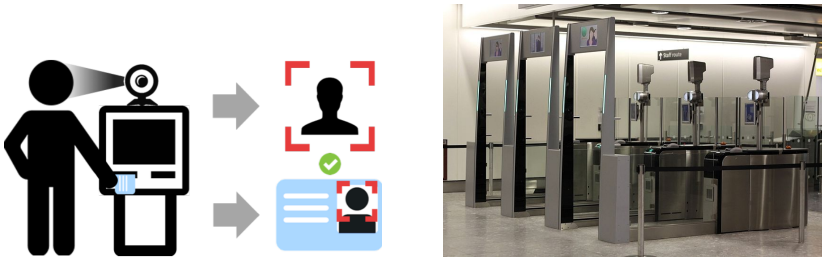
From a high-level point of view, an automatic face recognition system usually takes a two-dimensional image of a face and extracts a group of face features for representation. Those features can be used for identity verification or identification (see Section 2.4).

Automatic face recognition systems were made possible by advances in the computer vision and

machine learning (ML) fields. With deep neural networks and large collections of images, it is possible to process images (and subsequently face images) into a low dimensional feature space where further processing and recognition tasks can be made [2]. These systems convincingly outperform humans in many benchmarks.

1.2 Motivation

In some border control systems, namely in airports, face recognition is already used to verify a person's identity. In these systems, a photo is taken of the individual and then compared with his/her identity documents (ID) or travel document and access is allowed (or not). This use case proves that state-of-the-art face recognition systems can help facilitate and automate tasks that humans would otherwise perform. Although this can accelerate some tasks, it can also increase the robustness of said tasks since these systems are proven to outperform humans.



(a) (b)

Figure 1.1: Examples of ID-Face matching systems: a) Schematic of an ID - Face matching system. Image from [3]; b) Heathrow Terminal ePassport gates.

The idea of using face recognition systems to help automate and facilitate tasks like ID or passport creation is already being developed within the scope of the FACING project, a partnership including the Imprensa Nacional-Casa da Moeda (INCM) and a computer vision team from the Instituto de Sistemas e Robótica (ISR) within the University of Coimbra. This project aims to develop a mobile system for smartphones, or in the form of a web app, whose purpose is to allow the user to do a safe self-enrolment. As an example, this could be used to renovate a passport or ID from home. Apart from document-related applications, this system could also be used in banking and other services requiring safe access. The FACING project can be grouped into 3 different tasks:

- Verify if a submitted image complies with the international standards and requirements (set by International Civil Aviation Organization (ICAO)). In other words, verify if an image is ICAO standard compliant. The app also intends to help guide the user in the process of obtaining an ICAO compliant image.

- Liveness detection is the task of verifying if a human in a video is, in fact, a human or a presentation/spoofing attack is being made to the system. This type of attack has the goal of interfering with a system into thinking another individual is present. For example, this can be done by showing a picture or a screen with another individual photo to fool the system. High-level presentation attacks can also be made using custom 3-D masks and makeup, but those are harder to detect.

- Face Recognition (FR) module that has the goal of verifying if the app user is whom he/she claims to be. It is a system built from scratch to be light enough to run on mobile devices and based on deep learning techniques. Although the project's goal is not to invent a new state-of-the-art recognition system, as current solutions already have acceptable levels of performance, it is always desirable to try to improve current methods. This topic is the context where this thesis is inserted.

All ID and travel documents contain a photo of the face of the holder. This photo also follows a vast number of recommendations and requirements set by the ICAO [4]. Some examples of these requirements are that the image must be sharp enough, there must be no face occlusion, the illumination must be frontal on the face, the subject looking directly at the camera, among many others (see Fig 1.2). This is a much different scenario than unconstrained face recognition, where all variations are expected to be found.

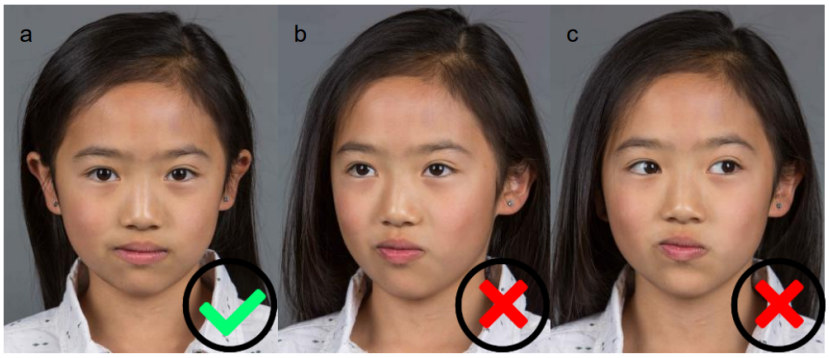
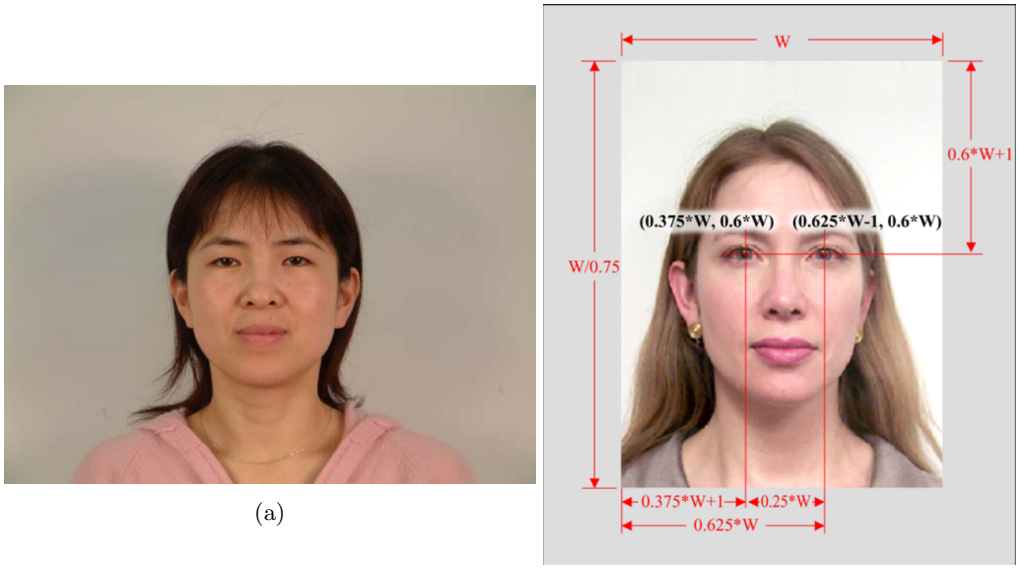


Figure 1.2: Example of pose and eye alignment ICAO requirement. a) is a compliant portrait, in b) the head not aligned toward the camera, and in c) eyes not aligned toward the camera. Image from [4].

One important requirement is that the face should be frontal and, according to the international standard ISO/IEC 19794-5 [5], two types of frontal faces can be defined:

- Full frontal face images, which should have enough resolution to be examined by humans and to be consistent for automatic face recognition purposes. This type of frontal image should contain the neck and shoulders as well as the hair of the subject.

- Token frontal face images, which follow specific geometric restrictions for image size and for the position of the eyes in the image according to image height and width. This type of portraits has fixed aspect ratio characteristics and the storage size is reduced when compared to full frontal.



(a) (b)

Figure 1.3: Examples of the two types of frontal images: a) Full frontal; b) Token frontal. Examples from [5].

Both types of images can be used for machine readable documents. Images in these types of documents can be used as-is or encoded for more secure recognition purposes. For example, Medvedev et al. [6] developed an efficient method for the protection of ID and travel documents, done by encoding the image in machine-readable code with augmentations based on the facial biometric template.

Although current developments in deep networks allowed for impressive results in FR applications, these networks require large amounts of data to be effective. For unconstrained FR, there are large public collections of labelled face images (usually from celebrities [7–9]). This is not the case for ICAO compliant images. Due to legislation concerning the use of personal data, the collection and distribution of personal data are strongly constrained. For instance, the European General Data Protection Regulation considers biometric data such as facial images as sensitive personal data, which results in several restrictions for its use. This means the large collections of ICAO compliant images are private, and the small amount of publicly available collections that exist are of insufficient size for deep learning purposes. This discrepancy between the type of data available for training (unconstrained also called "wild" data) versus the data of the final application (ID/travel documents photos) means there is room for optimisation.

As such, taking into account the current scenario of face recognition for document security, the

availability, size and type of data that can be used for these applications, this thesis project aims to contribute to developments in face recognition optimisation towards document security, with the use of deep learning techniques.

1.3 Objectives

As mentioned in the above section, the general goal of this thesis is to try to improve current deep learning-based methods performance on the task of FR, with a focus on the use case of the FACING project, which is ICAO compliant images. To achieve said goal, specific objectives were defined:

- Review current face recognition state-of-the-art literature and analyse the common datasets and deep networks used.
- Define the deep network and datasets used to develop FR models.
- Search relevant face image meta-information that could be included in the training process.
- Investigate and improve deep learning face recognition strategies for document security applications. Consider the design of a custom loss function that uses sample-specific information.
- Study protocols to evaluate the methods developed and evaluate the performance of the developed approaches.

1.4 Contributions

From the work developed in this thesis, a paper was published in the BIOSIG 2021 conference. The paper is named "QualFace: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment" and it is included in the Appendix.

The contributions of this work are listed bellow:

- Compilation of metrics to evaluate Machine Readable Travel Documents portrait quality.
- Formulation of a margin-based loss function that includes sample quality in such a way to increase inter-class separation and intra-class compactness of face embeddings for the document security scenario.
- Creation of two benchmark protocols to measure a model's performance on document security applications.

1.5 Outline of the Dissertation

The remainder of this document is structured as follows: In chapter 2, the theoretical background of the work developed is explained. Chapter 3 explores the state-of-the-art in face recognition, on which the work developed depends. Chapter 4 presents an overview of the methodology used to develop and test the models used. The results obtained are presented and discussed in chapter 5. Finally, in chapter 6, a conclusion for the work developed is made.

Chapter 2

Theoretical Background

In this section, essential aspects and concepts regarding FR systems will be presented. Then, an overview of the pipeline of a FR system will be reviewed, the main concepts behind a convolutional neural network (CNN), and how its training is done. Finally, a brief overview of recognition scenarios will be presented.

2.1 Overview

Current state-of-the-art FR technologies are based on deep learning methods [10]. The scope of this thesis is related to deep learning methods; hence only those will be detailed in this introduction. The common pipeline for a generic deep learning FR is represented in figure 2.1 [10].

The face detection module is used to localise the face or faces present in an image or video. The most common face detectors' output is a bounding box for the face in conjunction with the positions of landmark points of said face, i.e. eye centres, mouth corners, nose tip, etc.

Next in the pipeline, the face alignment module aligns the face's landmarks to some predefined coordinates or orientation and crops the image to the desired dimensions.

Finally, the FR module receives the aligned images. The image is used as input in the system's CNN and transformed into a relatively low dimensional set of features (feature vector) used to describe the said image. This feature vector is then used for the task of face matching/recognition.

2.2 Convolutional neural networks

As mentioned above, FR currently achieves the best results through the use of deep learning methods. CNN are the go-to method used for a plethora of computer vision applications, including FR.

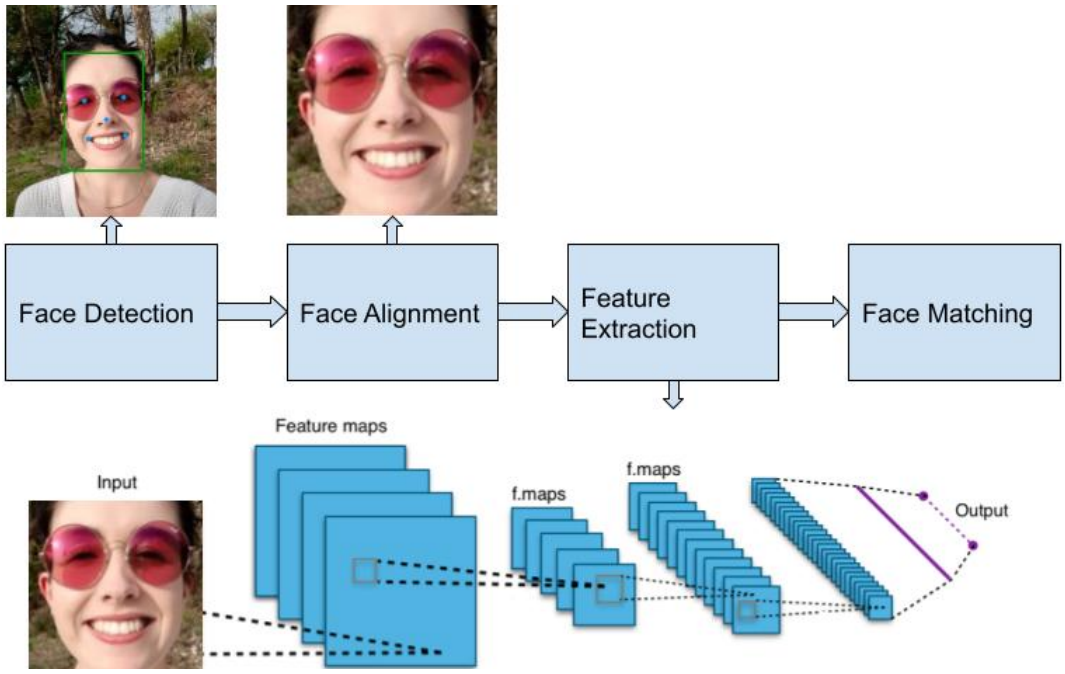
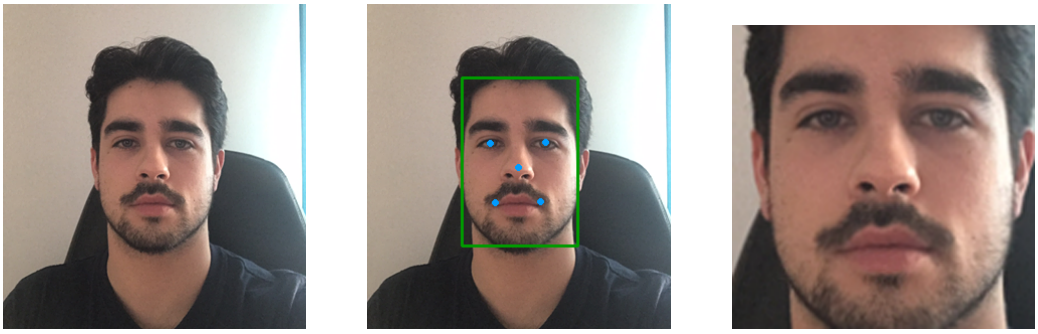


Figure 2.1: Diagram of face recognition pipeline.



(a) (b) (c)

Figure 2.2: Detected and aligned images: a) Default; b) detected; c) Aligned.

2.2.1 Artificial Neural Networks and Flat Layers

Artificial neural networks, or simply called neural networks, are a machine learning method inspired by the neurons and their connections (synapses) in animal brains. In a neural network's structure exist nodes, also called neurons, that store real number values. Groups of nodes constitute layers (named flat layers), and a group of layers and the connections between them form a neural network. (see Figure 2.3). As seen in the figure, there are three categories of layers in a neural network, the input layer, the hidden layers and the output layer. The input layer receives external inputs, the output layer contains the networks' final output, and between them exists any quantity of hidden layers.

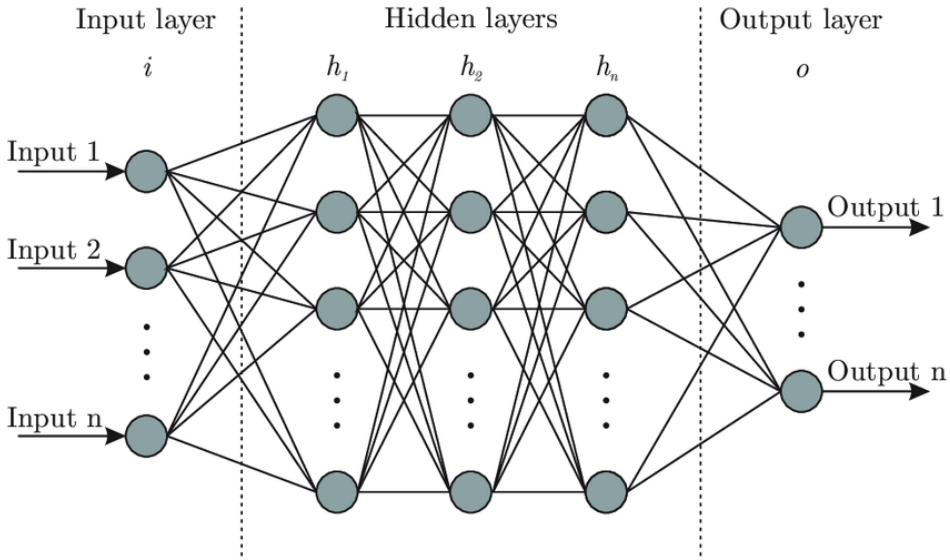


Figure 2.3: Scheme of a simple neural network architecture.

In a conventional neural network, any node of a given layer is connected to all nodes in the previous layer. Furthermore, these connections are weighted, meaning the influence of a node on another can be different for each connected pair of nodes. So, when the neurons in the input layer receive a signal, the signal is passed onto the first hidden layer, then the second, iteratively through all hidden layers until reaching the output layer.

Focusing on the connection between two arbitrary layers l and $l+1$, let x_i^l be the signal on the i'th node of the l'th layer. Also, since the weights can vary for each neuron pair, let $w_{i,j}$ be the connection weight between the i'th and j'th nodes on layers.

With this, the value of a given node in layer $l+1$ can be given by the following equation:

$$x_j^{l+1} = \sigma\left(\sum_i (w_{i,j} x_i^l) + b_i^l\right) \tag{2.1}$$

From the above equation, the signal of a node is equal to a linear combination of the signals of all the previous layer's nodes, plus a bias b_i^l and "wrapped" by a non-linear function σ. This formulation is for the case of fully connected layers, however, there are other possible ways of connecting layers.

2.2.2 Convolutional Layers

A CNN, also called ConvNet, is a type of artificial neural network that is extremely useful for processing images. As the name suggests, a ConvNet utilises convolutional layers for at least one of its layers. Convolutional layers can capture spatial features that would not be possible to capture by

flattening the image and using a flat layer.

2.2.3 Convolution operation

A convolutional layer uses a three-dimensional kernel with predefined dimensions NxMxC (hyper-parameters) where N, M correspond to height and width, respectively. C is the number of channels of the kernel, also called depth, which should be equal to the number of channels of the input volume of the layer (e.g. an RGB image has 3 channels). Another important hyper-parameter when defining a convolution is the stride. The stride is the length of each step taken by the kernel during the convolution operation.

To perform the convolution, the kernel slides across the input volume with steps that are the size of the stride previously defined. Then, at each location, element-wise multiplication between the elements of the kernel and elements of the volume the kernel overlaps is made, and the sum of these values is the output (named feature map) at the current location. This operation is visually shown in figure 2.4.

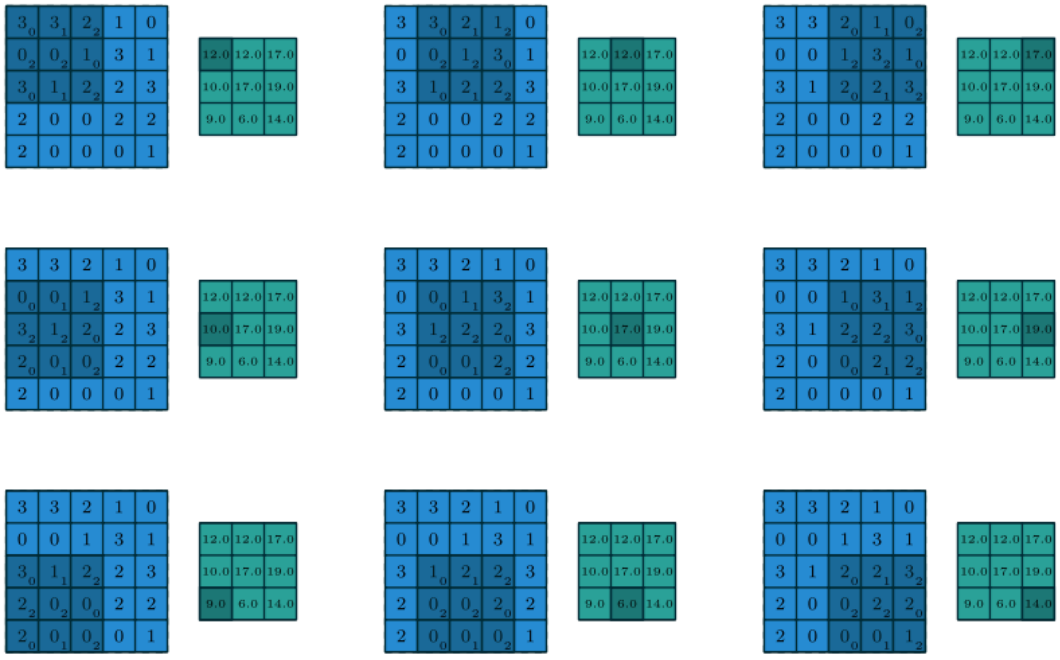


Figure 2.4: Two dimensional convolutional between 5x5x1 input map and 3x3x1 kernel with stride 1. Image taken from [11].

The width and height of output depend on the input/kernel dimensions and also the stride. From figure 2.4, it is possible to understand that unless the kernel size is 1x1, the output will always be smaller than the input. This is not desirable because if many successive convolutions are applied, which is the case for most deep CNNs, the final output map will be too small to convey any relevant

information. As such, padding is used to limit this effect. Zero padding, the most common form of padding, is the addition of zeros on the image's borders. This allows the convolution to retain the input's height and width in the case of Same Padding or even allow for increased dimensionality in Valid Padding.

Since square kernels and images are common, let O, I, K be the dimension of the side of the output, input and kernel, respectively. P is the size of padding used, and S is the stride. The output dimension is then obtained the following calculation:

$$O = \frac{I + 2P - K}{S} + 1 \tag{2.2}$$

To extract several features from the same input, it is common to stack the results of several convolutions with different kernels resulting in output with depth equal to the number of kernels used, represented by C.

2.2.4 Pooling Layers

Pooling is another common operation in convolutional neural networks and is usually done after a convolution operation. A pooling layer is used to reduce the dimensions of the feature maps [12]. This is useful since it helps reduce computational effort. For example, in a max-pooling operation, a kernel slides across the input just like in convolution. Still, the value used of the output is the maximum value in the overlapped area. Because of this, max-pooling layers help extract dominant features and remove undesired feature noise. There exist other types of pooling layers, for example, average pooling.

2.2.5 ReLU

The ReLU function is a type of activation function and stands for Rectified Linear Unit. In the context of neural networks, activation functions serve the purpose of introducing non-linearities that allow the neural network to approximate its output to non-linear functions. This is required for the vast majority of applications since most of the phenomena studied are non-linear. The ReLU function is defined as follows [13]:

$$f(x) = \max(0, x) \tag{2.3}$$

This formulation introduces some type of non-linearity required, but it is also efficient in terms of computation, which is advantageous in large architectures. This function is usually applied on the feature maps after convolution layers and also after fully connected layers.

2.3 Training

In the previous section, it was mentioned the existence of weights connecting layers and kernel weights. These parameters are what define the output of a ConvNet for any given input. How these parameters are defined is crucial to get the desired outputs instead of random numbers. To understand how the parameters are updated firstly is important to define what a loss function is.

2.3.1 Loss functions

A Loss function compares the outputs of a model with the desired outputs and returns a real number. This number simply referred to as "loss", represents some representation of the error estimate of the model. This means the larger the loss value, the farther the model is from its pretended output. There are several commonly used loss functions for simpler problems. A simple example of a loss function is the quadratic loss, commonly used when applying least-squares approaches in regression problems.

$$L_{quadratic} = A(y - x)^2 \tag{2.4}$$

where A is a set constant, y is the pretended output, and x is the model's output. For FR applications, more complex loss functions are used, which will be explained in detail in chapter 3.

2.3.2 Regularisation

When training a network, one important aspect to have in mind is the possibility of overfitting. Overfitting is the phenomenon where a model learns too well the statistical noise and details of the training dataset, which leads to poor prediction performance when presented with new unseen data. To tackle this effect, some techniques may be used. These techniques are known as regularisation techniques.

Dropout

Dropout is a technique that consists in removing ("dropping") a random set of nodes and their respective connections of a given layer each iteration with a predefined probability. As a result, nodes can learn to adapt to fix mistakes from other nodes, which may lead to complex co-adaptations [14]. This, in turn, causes overfitting since the co-adaptations are specific to the training data and thus not able to generalise to unseen data. Dropping some nodes each iteration removes this problem and leads to better generalisation and results.

Weight Decay

Weight decay is another technique that consists in penalising the growth of the weights in a neural network. This is helpful to prevent a model from becoming overly complex and overfitting to the training data. Weight decay works by introducing an L2 regularisation of the weights into the loss function:

$$L' = L + \lambda \sum_i w_i^2 \tag{2.5}$$

where w_i represents the network's weights or layer and L represents the loss function without the weight decay. The value λ is the weight decay parameter that influences how much the L2 norm of the weights impacts the value of the loss.

2.3.3 Network training

Training a neural network is a process where its trainable parameters are updated with the direction of minimising the value of the loss function. In the case of a CNN, the trainable parameters are the weights and biases in the connections and the kernels' weights.

To start the training process, first, the network's trainable parameters must be initialised with some commonly random value (random initialisation). Still, they can also be imported from other previously trained networks. Then, for each training iteration, an optimisation algorithm is used to minimise the loss function, and, in combination with the backpropagation algorithm, all the weights are updated.

The first step to update the weights in a neural network is to calculate the gradient of the loss function with respect to all the weights and biases of the network for a given input. This is done by the backpropagation algorithm that computes the loss function's gradient with respect to the network adjustable parameters using the chain rule. This is accomplished by iterating from the last layer up to the first to avoid unnecessary calculations.

After calculating the gradient of the loss function with respect to each trainable parameter of the network, the parameters are updated according to the optimising algorithm used. One of the most common optimising algorithms is the family of gradient descent methods [15] (gradient descent, stochastic gradient descent, mini-batch gradient descent). These are used to optimise a function to minimise it in the case of the loss function.

Mini-Batch gradient descent uses a mini-batch (a small group of samples) to calculate the average gradient to update the parameters. This option has smoother convergence than stochastic gradient descent, which only uses one sample per step. On the other hand, it has much less computational

effort than traditional gradient descent, which uses the entire dataset to calculate the gradient. As such, mini-batch gradient descent is the most commonly used optimiser among the three mentioned. The following formulation can describe this algorithm [15]:

$$w'_j = w_j - \eta \frac{1}{N} \sum_{1}^{N} \frac{\partial L_i}{\partial w_j} \tag{2.6}$$

where w_j represents any trainable parameter in the network and w'_j is the updated parameter. L_i is the loss value for the sample i, and N is the batch size. Finally, η is the learning rate. Bigger learning rates lead to faster convergence but may "jump" over some loss function local minimums and not converge properly if too high.

2.4 Recognition Scenarios

FR systems function in a reference-based manner. It is assumed that the system has access to a set of correct reference images (gallery). Depending on the system, this reference can be captured in different ways. For example, the user could have previously submitted it, captured it via security camera, extracted it from an ID, etc. Face recognition can be categorised into face identification and face verification.

The face verification scenario involves two images that are compared to determine if they belong to the same subject (1-1 authentication). This scenario is found in automated border control systems, e.g. in airports where a live captured image is compared to an ID or passport photo. To evaluate the performance of systems in this scenario, the commonly used metrics are the following:

- False Match Rate (FMR), which is the fraction of non-matching identities wrongly matched as the same identity.
- False Non-Match Rate (FNMR), which is the fraction of matching identities wrongly considered a non-match.

The FNMR can be interpreted as a measure of how convenient a system is, while the FMR measures the level of security. These values can be plotted to show the Detection Error Tradeoff (DET) curve, exemplified in Fig 2.5. The verification accuracy is usually reported as the FNMR for a defined FMR threshold.

Other metrics are also used, but these are usually obtained from the two above mentioned. For example, the True Accept Rate (also named True Positive Rate (TPR)): $TAR = 1 - FNMR$. This value is useful for plotting the Receiver Operating Characteristic (ROC) curve, a commonly used graphical way to show the models' performance. On the ROC curve the x-axis represents the False

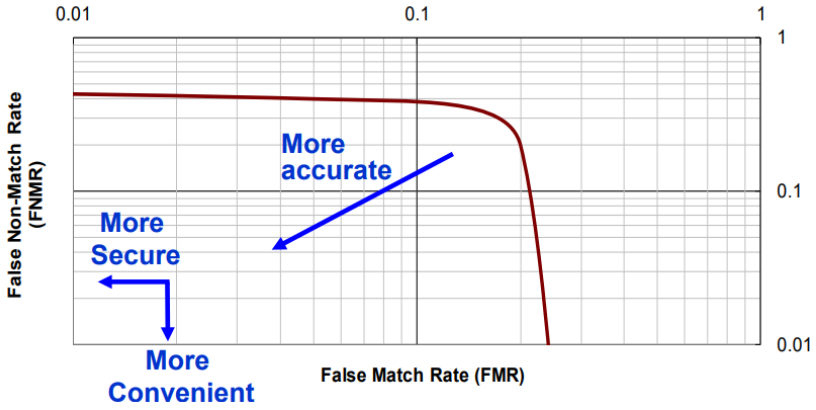


Figure 2.5: Example DET curve. Image from [16].

Positive Rate (FPR) (which in this application is the same as the FMR) and the y-axis represents the TPR. Another relevant metric is the Equal Error Rate (EER) which can be extracted from the ROC or DET curves. The EER is the point of the operation curve where $FPR + TPR = 1$ or, in other terms, the point where $FNMR = FMR$. Generally, the lower the EER the better the performance of the recognition system.

Face Identification uses a new image and the entirety of the gallery set (1-N authentication). Face identification is classified as closed-set if the probe's identity is contained in the identities present in the gallery set. In this case, the system tries to find the closest identity. To assess the performance of a face recognition system in this scenario, a common metric to use is the identification rate at rank r. For a given authentication attempt, the identification rate at rank r is the probability that in an identification attempt of an enrolled user, the user is at the top r members of the list of matched identities.

On the other hand, if new identities can be presented to the system, the task is classified as open-set identification. Open-set tasks are the scenario of most real-world applications of authentication systems. In open-set tasks, the metrics mention for the closed-set scenario can be used and in addition there are two other useful metrics to describe system performance:

- False Positive Identification Rate (FPIR), which is the fraction of non-enrolled identities that had a successful identification attempt.
- False Negative Identification Rate (FNIR), which is the fraction of enrolled identities for which the identification attempt was unsuccessful.

Similarly to 1:1 Face Verification, the performance of a system in a open-set scenario can be plotted, in this case, as the FNIR for a FPIR value at a given rank r.

Chapter 3

State of the art

In this chapter, the state-of-the-art regarding face recognition systems will be described. Firstly, some important datasets used in the field to compare system performances are presented. Then, the current advances in deep neural network face recognition are explained. The third section presents the current best-performing algorithms regarding face detection and alignment. Afterwards, a review of top-performing face recognition systems in civil identification and machine-readable travel documents is demonstrated. Finally, a short mention of recent advancements in face image quality assessment is made.

3.1 Datasets

Datasets/databases are a critical factor in developing deep learning models for FR. All state-of-the-art FR methods are heavily data-driven. A model's performance can have a significant bias depending on the quality, quantity and distribution of the data used to train it. Datasets are also used in benchmarks for system validation.

3.1.1 Training datasets

For effective deep learning-based FR, it is essential to have a sufficiently large training dataset. As Zhou et al. [17] show, large amounts of data in deep FR improve the model's performance.

At the beginning of deep learning-based FR, most state-of-the-art methods used large private datasets, so they could not be reproduced. The first publicly available dataset that helped solve this issue is CASIA-Webface [18]. Since this is a relatively small dataset with 500K images of 10K celebrities, more large-scale datasets were made available such as VGGFace [7], MS-Celeb-1M [9] among others. A collection of some widely used training datasets are presented bellow in Table 3.1.

Table 3.1: Common datasets for facial recognition training and benchmarking.

Datasets	Number of photos	Number of identities	Publish date
CASIA-WebFace [18]	494,414	10,575	2014
VGGFace [7]	2.6M	2,622	2015
VGGFace 2 [8]	3.31M	9,131	2017
IMDB-Face [19]	1.7M	59K	2018
MegaFace [20, 21]	4.7M	672,057	2016
MS-Celeb-1M [9] (challenge 1)	10M 3.8M (clean)	100,000 85K (clean)	2016
MS-Celeb-1M [9] (challenge 2)	1.5 M (base) 1K (novel)	20K (base) 1K (novel)	2016
MS-Celeb-1M [22] (challenge 3)	4M (MSv1c) 2.8M (Asian-Celeb)	80K (MSv1c) 100K (Asian-Celeb)	2018
Labelled Faces in the Wild (LFW) [23]	13,233	5,749	2008
YouTube faces (YTF) [24]	3,425(videos)	1,595	2011
IJB-A [25]	5,712(images)/2,085(videos)	500	2015
IJB-B [26]	21,798(images)/7,011(videos)	1,845	2017
IJB-C [27]	31,334(images)/11,779(videos)	3,531	2018

Datasets have two important aspects. First, depth, which measures the number of images per identity, provides the desired level of intra-class variations like pose, lighting, occlusion or even ageing. For example, VGGFace [7] has, on average, 1000 images per identity. Second, breadth, which measures the number of identities in the dataset, provides the desired level of inter-class variations and coverage for the different appearances of a sufficiently large number of people. A good example to show this is the MegaFace [20, 21] challenge with over 670K identities.

3.1.2 Evaluation datasets

Evaluation benchmarks aim to provide information regarding the viability and performance of a model. With an evaluation dataset and a benchmark protocol, it is possible to extract the performance metrics mentioned in section 2.4 in order to compare different models' performance.

The Labelled Faces in the Wild (LFW) [23] dataset is the most widely used benchmark for unconstrained FR applications. The images, obtained from the internet, are cropped using an automatic algorithm based on Haar Cascade called the Viola-Jones detector [29]. The protocols used in this dataset are intended for facial verification (pair matching). There are two variants, one where the pairs are provided, named restricted protocol, and an unrestricted protocol where the pairs can be chosen. The ROC curve is the method used to visualise the results. It is also worth mentioning the YouTube faces (YTF) [24] dataset where the images are taken from YouTube videos. The individuals

Figure 3.1: Sample images from benchmark datasets: Labelled Faces in the Wild (LFW) (top), video frames from IJB-A (middle) and MegaFace (bottom). Image from [28].

in this dataset are also present in LFW, and the protocols are the same. Finally, another dataset group worth mentioning is the IJB group of datasets [25–27]. These are manually aligned thus contain more variations in pose, occlusion and illumination that the automatic detection in LFW is not able to detect.

3.1.3 Face detection and alignment

As discussed earlier, some face datasets use some automatic face detection and alignment. This is also true for most FR systems. The scope of this work is applications on ID and travel documents. In these applications, most of the images processed are frontal face images (see section 3.6). As such, a deep study of face detection and alignment techniques is not the main target. Still, there are many published methods regarding face detection and alignment like the aforementioned Viola-Jones detector [29], histogram oriented gradient (HOG) techniques [30], Multi-Task Cascaded Convolutional Networks [31], among others. There are also deep learning-based techniques for face detection on which a survey is done by Rajan et al. [32], and Xin and Tan [33] present a survey on face alignment techniques.

3.2 Deep Learning Face Recognition

CNN are a type of Artificial Neural Network that are commonly used in the field of computer vision, usually applied in object detection, segmentation or recognition. CNN can learn high-level features from images, and those features can then be classified by the fully connected layers of a conventional neural network. Now, CNN are at the core of all state-of-the-art face identification and verification systems.

The initial evolution of deep convolutional neural network (DCNN) architectures was made by increasing the depth and number of units per level. This led to the ever-increasing complexity and computational resources required. Several breakthroughs in DCNN architectures like GoogleNet [34], AlexNet [35] and VGGnet [36] emerged, and these were more powerful and less resource-intensive than before. Further developments in image recognition and CNN were made by He et al. [37], who proposed to apply residual connections to the network and introduced the ResNet architecture. These connections result in faster training and better performances and, for these reasons, is commonly used in deep learning computer vision tasks.

In 2014, one of the first face recognition systems based on deep learning, DeepFace [2], was published by Facebook. DeepFace, at the time of its publishing, was the best performing algorithm on LFW [38] and the first to achieve near-human performance. (DeepFace: 97.35% compared with Human: 97.53% on the LFW benchmark unconstrained). The authors used a custom CNN architecture with a Softmax Loss function (see section 3.4) trained on a multi-class face classification scheme.

Current face recognition research and investigation utilise these well known and well-performing architectures as backbones for their DCNN. The recent investigation focus has shifted towards improving the loss functions used in these deep learning pipelines. Most of the recent breakthroughs and improvements are now related to enhancing the discriminative power of the features obtained from a given deep network. That is, it increases inter-class dispersion while maintaining inter-class compactness. This is done by altering or creating a new loss function with that goal in mind. That can be accomplished in two ways, using metric-learning or classification-based approaches. These concepts and related key articles are presented in the following two sections.

3.3 Metric Learning Loss Functions

Metric Learning methods consist of optimising the feature embeddings, in this case, to enhance their discriminative power.

One example of a metric learning loss function is the triplet loss introduced in face recognition applications by Schroff et al. [39] in FaceNet. As the name suggests, the triplet loss requires three

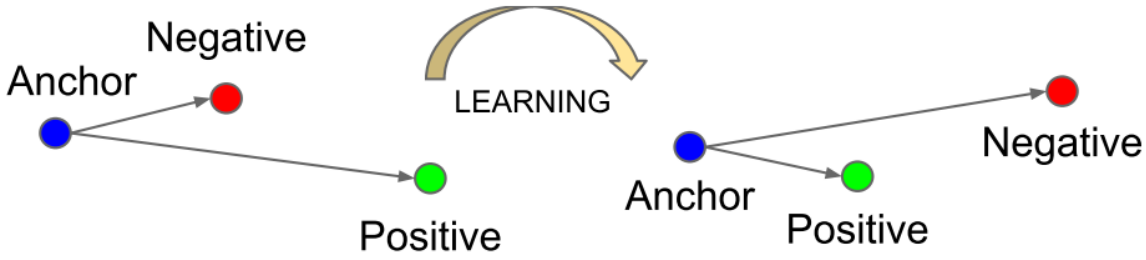


Figure 3.2: Visual representation of minimising positive pair distance while maximising the distance between negative pair. Image from [39].

data points per step known as the anchor x_i^a, the positive x_i^p and the negative x_i^n samples. The anchor and positive samples belong to the same identity while the negative sample is taken from the disjoint identity. The goal is to maximise the distance between the negative sample and the anchor while minimising the distance between the positive sample and the anchor, as illustrated in Figure 3.2.

Let $\mathbf{f}(x) \in \mathbb{R}^d$ be the d dimensional embedding or d dimensional set of deep features of an image. A constraint is made so that $\left\|f(x)\right\|_2 = 1$. Since the positive sample must be made closer to the anchor than the negative sample:

$$\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 + \alpha = \left\|f(x_i^a) - f(x_i^n)\right\|_2^2 \tag{3.1}$$

Where the triplet belongs to the set of all possible triplets $\mathbb{T}$ with cardinality N and α is the margin enforced between the distances of positive and negative pairs. Thus, the final loss function can be denoted as:

$$L_{triplet} = \sum_i^N \left(\left\|f(x_i^a) - f(x_i^p)\right\|_2^2 + \alpha - \left\|f(x_i^a) - f(x_i^n)\right\|_2^2\right) \tag{3.2}$$

The problem with the triplet loss method is that not all N triplets are useful, and using them all results in slower and worst convergence. In addition, using triplet loss requires some time-consuming data mining, even though there are methods presented to facilitate the process.

Another metric-learning based loss function worth mentioning is used in DocFace [3]. Since this article is related to document security applications, the loss function is explained in detail in section 3.6.

3.4 Classification Loss Functions

Classification based loss functions are used in problems that consist in classifying a given data point as one of the already existing classes (in this case, identities).

Most recent advances in face recognition [40–42] result from loss functions based on the softmax classification loss. The softmax classification loss' goal is to maximise the posterior probabilities of the true class. This is achieved by using the values of the output layer of the network (a layer with dimension C), $\mathbf{f} \in \mathbb{R}^C$, as an input to the softmax function, where the $i-th$ component of the resulting vector is defined as:

$$\sigma(f)_i = \frac{e^{f_i}}{\sum_{j=1}^{C} e^{f_j}} \tag{3.3}$$

Here, C is the number of classes of the classification problem. The softmax function serves as a normalisation function for the output of a network since the sum of the C components of $\sigma(\mathbf{f})$ is 1. These components can be interpreted as the probability of a given data point belonging to a certain class. This is usually implemented using a softmax activation on the network's last layer, called the softmax layer. The softmax classification loss is simply this softmax activation combined with the cross-entropy loss resulting in:

$$L_{softmax} = \frac{1}{N} \sum_{i} -\log(\frac{e^{f_{y_i}}}{\sum_{j}^{C} e^{f_{y_j}}}) \tag{3.4}$$

Here, N is the number of training samples, i refers to the $i-th$ sample, y_i is the index of the class of the $i-th$ sample and $\mathbf{x_i} \in \mathbb{R}^d$ is the output of the last fully connected layer, also called a deep feature. The value d is the dimension of the last fully connected layer.

The value f_{y_i} is the $y_i - th$ component of $\mathbf{f}$. With the weight matrix of the connections between the final two layers defined as $\mathbf{W} \in \mathbb{R}^{d \times C}$. An expression for $\mathbf{f}$ is obtained:

$$\mathbf{f} = \mathbf{W} \cdot \mathbf{x_i} + \mathbf{B} \tag{3.5}$$

where B, the bias vector of the last fully connected layer, is usually considered a vector of zeroes for simplicity. Denoting the $j-th$ columns of W as W_j, each component of $\mathbf{f}$ can be obtained by the following expression:

$$f_j = \left\|W_j\right\| \|x_i\| \cos\theta \tag{3.6}$$

The softmax classification loss function, although widely used up to this point, is not discriminative enough and results in lower performances for large intra-class variations (e.g. age gaps or pose variations) [42]. Therefore, the high separation between different classes and small intra-class separation became the main goal of modern investigation.

One common method used by some articles is to normalise the weight and feature vectors using l_2 normalisation and then re-scale the feature vectors to s: $\left\|W_j\right\| = 1$ and $\|x_i\| = s$ (however, some

approaches follow this strategy in simplified form [40]). This implies the dot product depends strictly on the cosine of the angle between these two vectors. Note that the new features are distributed across a s-radius hypersphere.

Using this method, the softmax function can be reformulated into the following:

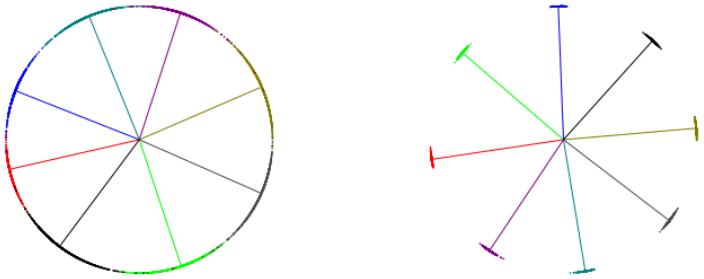
$$L_{Angular} = \frac{1}{N}\sum_{i} -\log\left(\frac{e^{s\cos(\theta_{y_i},i)}}{e^{s\cos(\theta y_i,i)} + \sum_{j\neq y_i} e^{s\cos(\theta_j,i)}}\right) \tag{3.7}$$

Various reported approaches deal with similar formulation modifying such loss function in different ways.

For instance, SphereFace [40] then introduces an angular margin by multiplying θ_{y_i} and the integer $m \geq 2$. The result is a monotonically decreasing angular function $\Phi(\theta_{y_i}, i)$ that replaces $\cos(\theta_{y_i}, i)$ and is equal to the cosine function for $\theta \in [0, \frac{\pi}{m}]$. This loss function, A-Softmax loss, led to higher discriminative features in the hypersphere and resulted in the best performance on LFW and YTF up to that point.

CosFace [41] improves on this by adding the margin directly into the cosine of the true class angle directly instead. This loss function is named large margin cosine loss (LMCL). The LMCL improves on A-Softmax and learns high-discriminative facial features, resulting in better performances than A-Softmax (although for the LFW with the same training data and CNN architecture achieved slightly worst results).

ArcFace [42] follows a similar approach but adds the margin directly to the angle θ_{y_i}. This results in a constant linear angular margin in the angular space. Through extensive testing it is found that the loss function introduced in ArcFace slightly outperforms the two prior mentioned approaches. A visualisation of the difference between softmax and ArcFace loss is presented in figure 3.3.



(a) Softmax (b) ArcFace

Figure 3.3: 2D hypersphere embedding of an 8 class problem using the Softmax versus ArcFace loss. It can be seen the increased inter-class separation and intra-class compactness of ArcFace compared to the Softmax loss. Image taken from [42].

3.5 Sample Specific Loss Functions

While previously mentioned approaches resulted in excellent performances in the evaluation datasets discussed, these same benchmarks are basically "saturated" since most modern methods achieve near-perfect performance on them. Even though this is the case, it does not mean the methods used are perfect though. Evaluation datasets like LFW present a relatively low variation in head poses [43], age [44], ethnicity [45], among other aspects. This is also true for training datasets. Even the huge ones present these kinds of biases. This lack of variation is not ideal and leads to poorer performance in more challenging datasets closer to real-world applications such as IJB-S [46], Cross-Age LFW [44], Cross Pose LFW [47], among others. Real-world scenarios offer a much broader diversity of head poses, age, ethnicity etc. Recent works aim at closing this performance gap.

Shi et al. [48] introduce some novel methods to try to tackle more challenging datasets. Firstly, the authors introduce three different data augmentation techniques: blurring, occlusion and head pose changes. These 3 augmentations were chosen since they are related to low resolution, high occlusion and lack of head pose variations which are common challenges in unconstrained/wild applications. Following this, the authors introduce a confidence-aware identification loss. This softmax based loss function introduces a sample-specific confidence parameter, s_i, where higher quality samples show higher confidence values. Usually, the class prototype w_i is simply the class centre. In other words, it does not depend on the quality of the samples. However, taking into consideration the sample confidence "pushes" the prototype towards the higher quality samples (as shown in figure 3.4). This makes higher quality samples that convey more information have a higher impact in training which is highly desirable.

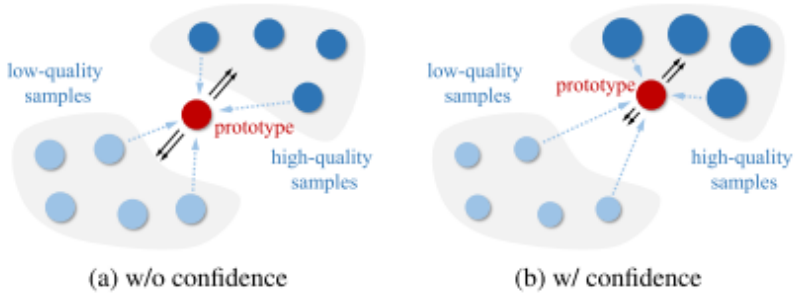


Figure 3.4: Visual representation of confidence-aware embedding learning. The learned prototype shifts towards higher quality samples. Image from [48].

Another relevant method introduced in this article is the decorrelation of the sub-embeddings. This method results in a more compact feature size and a higher representation power.

In CurricularFace [49], Huang et al. introduced curricular learning to create a loss function named Adaptive Curricular Learning Loss. Curricular learning is a method of training machine learning

methods where easier samples are introduced earlier in training while harder samples are introduced later. Using the ArcFace loss function as a base, the authors incorporated this concept in their work by including an adaptive parameter that controls the loss impact for harder samples. During training, this value is increased, and the impact of harder samples is increased as well. This parameter needs to be carefully formulated because it can greatly impact the convergence of the network. Hard samples were defined according to the performance of the network on each specific sample during training.

Sun et al. [50] observed that margin-based methods like ArcFace only enforced margin between classes, meaning no consideration was made regarding the intra-class discrepancy. For datasets with considerable class size differences (imbalanced datasets), larger classes will occupy more volume in the hypersphere resulting in bias towards these classes. This is not considered in margin-based approaches. The authors use the intra-class and inter-class similarity of a sample to measure the difficulty of the sample. The loss function takes two hyper-parameters, t_1 and t_2, the intra-class and inter-class similarity thresholds, respectively. Based on these hyper-parameters, margins are imposed in a sample-specific manner for positive and negative logits.

Zeng et al. [51] noted that hard samples could be categorised as hard positive and hard negative samples (intra and inter-class, respectively). The authors also verify that for large datasets, a hard positive sample will also generally act as a hard negative for another class.

Figure 3.5: Examples of hard positives between the first and second row and hard negatives between the second and third row. Image from [51].

To address this concern, the authors proposed the Negative-Positive Cooperation Loss (NPCFace). In NPCFace, the non-ground truth logit is altered to remove the stability-hard emphasis conflict other formulations like CurricularFace presented. This is made using a mask $M_{i,j} \in 1, 0$ that indicates if a sample i is hard or not to the $j-th$ class:

$$f_j^{M_{i,j}} = (1 - M_{i,j}) \cdot s \cos\theta_{i,j} + M_{i,j} \cdot s(t \cos\theta_{i,j} + \alpha) \tag{3.8}$$

The mask mentioned can be related with miss-classification, but other methods are also mentioned, for example, off-the-shelf options. Regarding ground-truth logit, the authors alter the ArcFace positive logit by including a cooperative margin m_i. If the sample is not considered a hard negative for any class, the margin is simply the constant $m_0 > 0$. Otherwise the margin is formulated as follows:

$$m_i = m_0 + \frac{\sum_{j \neq y_i} M_{i,j} \cos \theta_{i,j}}{\sum_{j \neq y_i} M_{i,j}} m_1 \tag{3.9}$$

where $m_1 > 0$ controls the impact of the hard negative samples. The cooperative margin is related to the average hard negative logits for each sample. If the sample is hard, the margin will increase from the baseline value m_0, increasing the loss value.

MagFace [52], unlike the works mentioned prior, utilises the magnitude of the feature vector as additional information to include in the loss function. The authors propose that the magnitude a_i can be used to indicate the quality of the sample i. They then use the ArcFace loss formulation, but instead of using a static margin m, they use an adaptive margin whose value changes with respect to the sample quality: $m(a_i)$. This results in higher margins for higher quality samples which, as will be explained further, will also be the approach in this thesis. The authors also add a term to the loss function to regularise the feature magnitudes $g(a_i)$ that rewards samples with larger magnitudes. The final MagFace Loss is defined as follows:

$$L_{MagFace} = \frac{1}{N} \sum_i -\log\left(\frac{e^{s \cos(\theta_{y_i} + m(a_i))}}{e^{s \cos(\theta_{y_i} + m(a_i)} + \sum_{j \neq y_i} e^{s \cos \theta_j}}\right) + \lambda_g g(a_i) \tag{3.10}$$

where λ_g is a coefficient that represents the trade-off between the regularisation and classification losses.

3.6 Applications on ID and travel documents

Security applications for ID pose peculiar face recognition challenges. That is why specific techniques and methods were developed for this particular subject.

Shi and Jain [3] presented a method for face matching based on ID photos entitled DocFace. This type of face matching relies on two types of images:

- The ID photo, either scanned or digital, which is obtained from the identification document. This photo usually is frontal, well lit and the subject presents a neutral expression. However, one problem that might arise with such photos is the lack of quality due to image compression.
- The live "selfie" of the individual. If the subject is cooperative, this image can be taken with relatively the same pose, lighting and expression as the ID photo, although that might not be

the case, and variations might occur.

Figure 3.6: Two ID-Selfie pairs from the private ID-Selfie-A dataset [3].

The major difficulties in matching these two different types of images from different sources come from the image quality and the time span between the issuing of the ID and the time of verification.

The approach presented uses ResNet architecture as the backbone trained on a big wild dataset (MS-Celeb-1M [9]). The loss function used is the angular margin softmax loss function. Since the target domain (ID photo - selfie pair) is different from the training domain; if this network were applied as-is, the results would be poor. As such, the authors take two copies of this trained network, called sibling networks, and then fine-tune them on the smaller ID-Selfie pair dataset. In this process, one network is trained on the ID part of the dataset and the other on the corresponding selfie part. The loss function used for the sibling network training is called the max-margin pairwise loss. This loss function is inspired by the triplet loss:

$$L_{MPS} = \frac{1}{M/2} \sum_{i=1}^{M/2} \max[0, \max_{j \neq i}(\max[\cos\theta_{j,i}, \cos\theta_{i,j}]) - \cos\theta_{i,i} + m'] \tag{3.11}$$

where $M/2$ is the number of pairs on the mini-batch. Similar to other loss functions presented earlier, $\cos\theta_{i,j} = g_i^T \cdot h_j$. g_i and h_i are the L2 normalised embeddings for the ID and selfie images, respectively. The scalar m' is the margin imposed. This method achieves state-of-the-art performance in ID-Selfie matching, although this method is trained and tested done on a private datasets.

Still in this line of research, Shi and Jain improved on the results of DocFace with DocFace+ [53]. Most of the pipeline used in DocFace is carried over to DocFace+. Apart from using a bigger ID-Selfie dataset for training, the main improvement presented is a newly introduced loss function for the training of the sibling networks. The new loss functions used, called DIAM-Softmax, results in faster convergence, crucial in this scenario where the amount of data available for training is reduced. Shi and Jain show this loss function outperforms all other state-of-the-art loss functions presented on their private ID-Selfie dataset.

3.7 Face Image Quality Assessment (FIQA)

FIQA has similarities with the generic image QA. For example, a blurry image, a distorted image or an over-saturated image will be considered a low-quality image whether it is a face image or not. Although this is the case, there are certain desirable characteristics specific to face images: Frontal pose, frontal illumination, no face occlusion, among others. A document with recommendations and requirements for high-quality portrait images (for Machine Readable Travel Documents) was made by ICAO [4]. The image quality indicators used in this work were chosen in close proximity with the standards set by ICAO and are explained in Section 4.3. Apart from the quality indicators used in this work, other useful FIQA tools could be used in future work. A survey on this topic is conducted by Schlett et. al [54]. Out of the recent FIQA literature, some noteworthy articles follow that focus on removing the human perception from the quality estimation algorithm.

In SER-FIQ [55], the authors propose a quality estimation method that is based on the usage of dropout when training a network for FR. If a network is trained via dropout, it is possible to use a set of sub-networks to generate several outputs for a given input. By comparing the proximity of the outputs, a quality estimate can be made for the sample in question. The closer all the outputs are (more robust prediction), the higher sample's quality is considered to be.

In Probabilistic Face Embeddings (PFE) [56], Shi and Jain explain that poor image quality is responsible for changes in similarity scores of genuine and impostor pairs. Furthermore, these changes increase with the increase in degradation, which increases the likelihood of mismatching genuine or impostor pairs. To fix this problem, the authors introduce a new form of face embedding named a PFE. Unlike normal deterministic face embedding, with PFE, the model used outputs two different vectors to convey the uncertainty in the representation. These two vectors are the Gaussian mean and Gaussian variance of the embedding. Furthermore, to match the PFEs from different samples, a method is proposed that penalises high levels of uncertainty (Gaussian variance).

SDD-FIQA [57] is another method that does not depend on human perception for the quality estimation of a sample. To estimate the quality of a sample, the authors propose to use the inter-class and intra-class similarity scores and map them to pseudo-labels that indicate quality by using a distribution distance metric. The final step in the method is to train a new network with the quality labels to predict the quality scores.

3.8 Discussion

Deep learning-based methods applied in computer vision led to improvements in pattern recognition, image segmentation and other tasks. These works usually use some CNN. Face recognition systems

have also adopted CNN as part of their pipeline and have shown significant growth in accuracy and performance since their inception.

Due to significant improvements in network architecture, which led to the less computational effort while improving results, new research shifted its focus to fine-tuning loss functions for specific purposes while maintaining a well-known network as the backbone used. Several types of CNN architectures can be used for face recognition purposes without a clear top-performing architecture.

New research focuses on methods to improve the discriminative power of the learned face embeddings. These are commonly under the form of loss functions. Metric learning approaches have achieved good results, but the data mining and also data quantity required leads to harder convergence. Another type of approach possible is to use a classification based loss function. This has been done by somehow enhancing the softmax loss function. This type of research achieved increasing levels of inter-class variance while also increasing intra-class compactness on the resulting face embeddings. However, datasets with larger variations (for example, head pose or lighting) still challenge current methods. For this reason, most authors are experimenting with enhancing the impact on the train of hard samples in a sample-specific way.

Applications on ID and travel documents benefit from the advances previously mentioned, but they also have specific challenges. The uniqueness of the problem requires different approaches, such as the use of sibling networks. The lack of large ID or ID-Selfie public datasets is a big challenge in training deep learning models for these applications.

There have been rapid developments in FIQA, based on classification performance instead of intrinsic image properties. This new approach shows significant improvements over previous techniques and should be further explored in future work.

Chapter 4

Methods

In this chapter, the process and decisions made for the development of the experiments done are explained.

4.1 Choice of architecture

The first step taken to develop a model was the choice of the network/backbone architecture. This is done since developing a new architecture is out of the scope of this thesis. The benefit of the loss function should be transferable to different network architectures which is another reason to define the network now. The choice of architecture for the backbone was the ResNet. ResNets are efficient and commonly used in most state-of-the-art papers regarding novel loss functions. Within the ResNet "family" of networks, the ResNet-50-V2 [58] is a good balance between performance and computational effort.

4.2 Choice of train dataset

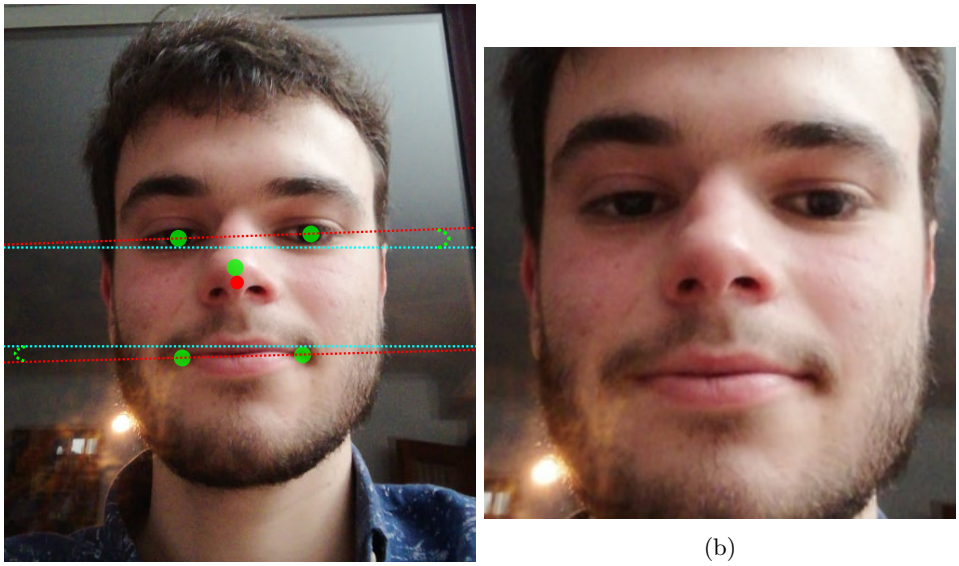
Some different datasets were considered to train the models. The three main options considered were: CASIA-Webface, VGGFace2 and MS-Celeb-1M. VGGFace2 was chosen as the training dataset since it is considerably bigger than CASIA-Webface and has much less label noise than MS-Celeb-1M. The details of these datasets are presented in table 3.1.

VGGFace2 [8] presents an unconstrained set of face images acquired from Google Images. There is an average of 362.6 face images per individual with 9131 subjects. These images have large variations in pose, lighting, age and other aspects, useful for the work developed.

4.2.1 Train dataset processing

VGGFace2 is quite large, and to compare the performance of different models; it is sufficient to train said models on a subset of this dataset. This allows for a shorter training time at the cost of a slight performance decrease. The subset chosen for training was the set of identities in VGGFace2 that had 400 or more images. So, instead of the 3.31M images and 9,131 identities of VGGFace2, the dataset used for training had 1.34M images and 2842 identities.

Before using the dataset to train a model, it is common to crop and rotate the images according to the position and orientation of the face. To do so, first, the face was detected in each image. This was done using an open-source pre-trained CNN from a project named RetinaFace [59].[1] This method (which also uses a ResNet50 architecture as a backbone) receives a face image as input and outputs a set of landmarks/key-points of the face, a bounding box for the face and a confidence score that reflects how "certain" the network is that the points return indeed belong to the face. It can happen that the face detector does not detect a face in the image, for example, if the face has really poor illumination, pose or quality. If this is the case, the image is not used. 5 important key points were then further used for face alignment: the centre of both eyes, corners of the mouth and the nose tip (as seen in Fig 2.2b)).



(a) (b)

Figure 4.1: Face image alignment process: a) Representation the 5 keypoints (green), face centre point (red) and the angles used to rotate the face (angles between red and blue lines). b) Result after alignment and cropping.

With the face detected, the final step to prepare the dataset was to align the faces according to

[1]The implementation used can be found in the following URL: `https://github.com/peteryuX/retinaface-tf2`

their landmarks. To do so, the following procedure was designed: From the previously extracted face landmarks, the face centre is calculated. This is done by averaging the X and Y coordinates of the 5 landmarks. Afterwards, the angle that the line connecting the two eye landmarks makes with the horizontal is calculated, and the same is done for the two mouth landmarks (see Fig 4.1). These two angles are averaged, and the resulting angle is used to rotate the image pivoting on the face centre. Finally, the resulting image is cropped and resized to the desired dimensions and saved.

4.3 Face image meta-information

As mentioned earlier, the main goal of this work is to improve FR performance through manipulation of the loss function. The idea of extracting extra information from the image and including it in the loss function to improve performance was explored. In total, 5 different types of meta-information were extracted from the train dataset used. These quality scores were chosen with the ICAO requirements and recommendations for portrait photographs in consideration.

4.3.1 Image blur

The first information extracted from the image was a measure of the image blur. This was done by simply taking the variance of the image after convolving it with a Laplacian filter [60]. This value gives a solid indication of the level of blurriness in an image, where lower values indicate higher blur.

4.3.2 BRISQUE

BRISQUE is a no-reference image quality assessment method, and it can be applied to any image [61]. BRISQUE stands for Blind/Reference-less Image Spatial Quality Evaluator. This method is useful to assess how distorted (or not) an image is. The output of this method is a simple value from 0-100, where lower values are best.

4.3.3 FaceQNet

FaceQNet is a face recognition specific quality assessment tool based on deep learning [62] [2]. To develop the method, the authors use a third-party framework to calculate ICAO compliance scores and use those as a baseline to train a DCNN. The trained network outputs a generic quality score ranging from 0-1, where 1 corresponds to the best face quality image.

It is shown that this quality score has a high correlation with face verification performance for commercial of-the-shelf systems. Higher quality scored samples resulted in better verification.

[2]For this work, the following FaceQNet implementation was used: `https://github.com/uam-biometrics/FaceQnet`

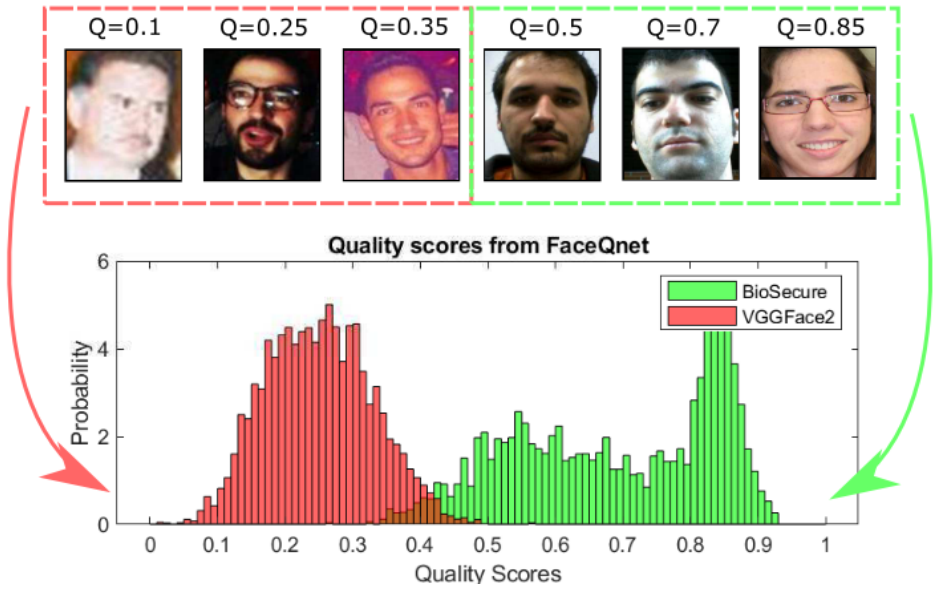


Figure 4.2: Quality score distribution from VGGFace2 and BioSecure datasets. The BioSecure dataset presents higher quality scores as its images were acquired in more controlled conditions. Image from [62].

4.3.4 Illumination Quality

One relevant variation present in face images is the quality of face illumination. According to ICAO standards and requirements for document images, portraits should have adequate and uniform lighting. To extract a measure of the quality of face lighting, a pre-trained DCNN based method was used [63] [3]. This model was trained on the Face Image Illumination Quality Database (FIIQD). Its output, similarly to FaceQNet, is a score ranging from 0-1, where the value 1 corresponds to the best possible illumination quality.

4.3.5 Pose

The pose can be characterised by the rotation in three dimensions, the yaw, pitch and roll, which are visually shown in figure 4.3

The angles of rotation for the three-axis described can be extracted for a face image in various ways. [4] These three angles can be combined to create a simple score. In this work, the average of the absolute value of the angles was used as the score for a given image. In this case, a full-frontal image would have a score of 0.

[3] For this work, the following Face Image Illumination Quality Assessment (FIIQA) implementation was used: `https://github.com/zhanglijun95/FIIQA`

[4] For this work, the following implementation was used:`https://github.com/OverEuro/deep-head-pose-lite`

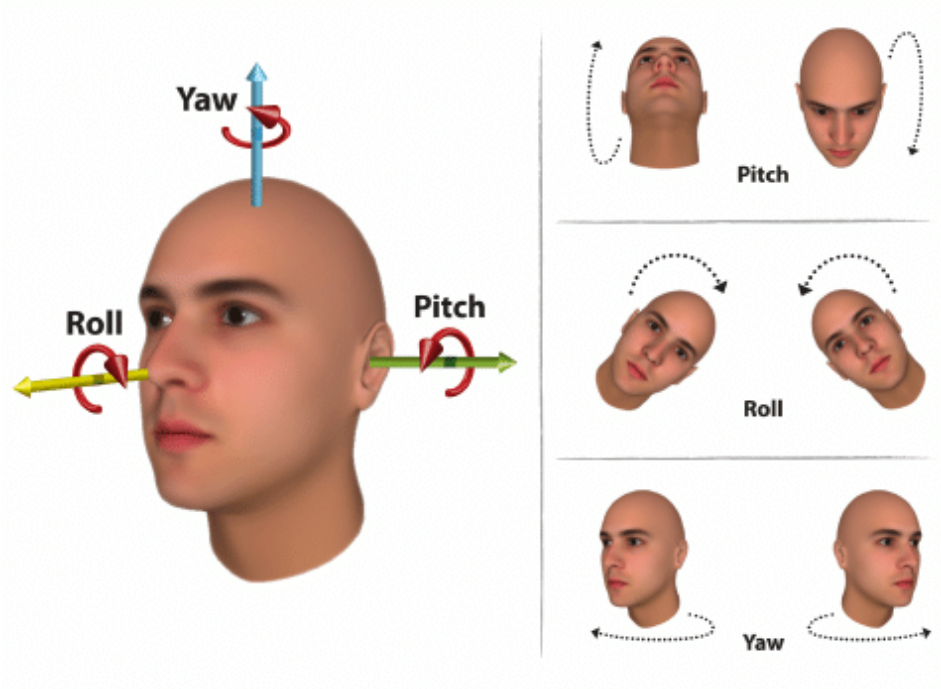


Figure 4.3: Representation of the yaw, pitch and roll axes. Image from [64].

4.4 Loss function formulation

To integrate the meta information extracted from the images in the loss function it was first defined the approach to follow. Two different directions were considered.

The first considered approach was changing the loss function to penalise harder samples more than easier samples. Harder samples can be defined in several ways, e.g. samples miss-classified by the model, samples with high loss or samples with intra/inter-class distances that are larger/smaller than desired can be used as hard samples [49–51]. These samples usually have undesirable variations, e.g. face occlusion, poor lighting, blurry photo non-frontal pose, etc. In the context of this work, a hard sample could also be defined in terms of the meta-information discussed earlier. Optimising the loss function with this approach helps the network learn harder scenarios and leads to better results in *wild* scenarios with many different possible variations.

The other approach considered has an opposite goal. Instead of increasing performance in generic scenarios by increasing the penalty on hard samples, this approach focuses on increasing the impact of higher quality samples by increasing their loss value. Higher quality samples can be defined with the meta-information previously mentioned. The desired goal of this approach is that the FR system would perform better for the use case of the FACING project that involves mostly frontal, well-lit face images, which, according to the definition used, are of higher quality. This was the approach followed.

4.4.1 Mathematical formulation

Inspired by NPCFace [51], the mathematical formulation for the loss function used was based on the ArcFace [42] loss function with an adaptive margin parameter:

$$L_1 = \frac{1}{N}\sum_i -\log\left(\frac{e^{s\cos(\theta_{y_i}+m_i)}}{e^{s\cos(\theta_{y_i}+m_i)}+\sum_{j\neq y_i} e^{s\cos\theta_j}}\right) \tag{4.1}$$

where the adaptive margin parameter m_i is defined as:

$$m_i = m_0 + \sum_j^Q w_j q_{ij} m_1 \tag{4.2}$$

and

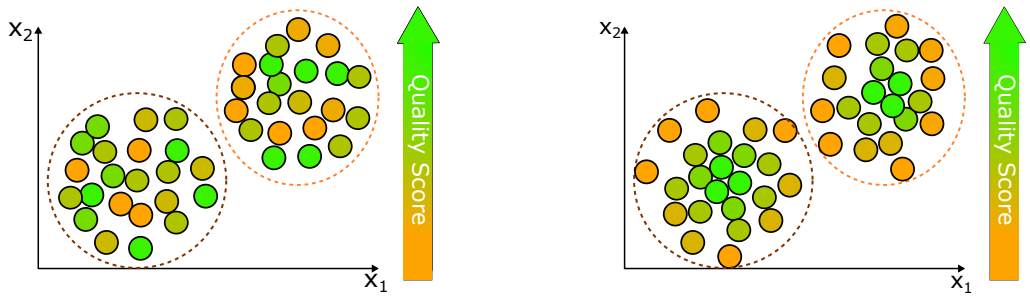
$$\sum_j^Q w_j = 1 \tag{4.3}$$

where, m_0 and m_1 are hyper-parameters that signify a baseline margin value and the amount of maximum change to that margin allowed, respectively. The value Q is the total number of quality attributes used and q_{ij} is the $j-th$ normalised quality score for sample i. Finally, $w_j \in [0,1]$ represents the importance of the $j-th$ quality score.

This formulation directly alters the ArcFace loss function intended to adapt the margin value with a linear combination of the quality scores used.

All the scores q_{ij} used should have higher values for better quality images and lower values for lower quality ones (if that's not the case, for instance, with BRISQUE, the scores should be inverted). This results in higher regularisation for higher quality images, resulting in a higher loss value, as desired.

The described adaptive margin method has a desired feature distribution with a larger concentration of higher quality samples closer to the class centre (Fig 4.4b), unlike "standard" margin methods like ArcFace, where the class centre is not affected by sample-specific attributes (Fig 4.4a).



(a) (b)

Figure 4.4: The sample distribution in two deep features; a) simple margin methods distribution; b) desired distribution with sample-specific quality information.

4.5 Benchmarking

In order to compare different models with different parameters and understand the performance of the models on real-world scenarios, benchmarks are required. Therefore, three different benchmarks were designed to test the models: a benchmark with unconstrained face images, a benchmark with images that have relaxed compliance to ICAO standards and a benchmark with images that have strict compliance to ICAO standards, named the *Wild*, *Relaxed* and *Strict* benchmarks respectively. These three scenarios were chosen to test the models under different conditions to better understand how their performance changes from wild images with high variability to relaxed and strict ICAO compliant images, which are the scope of the FACING project.

The benchmarks were designed as face verification benchmarks, a 1-1 comparison of images to check if the identities are a match or not. To design the benchmarks, firstly, datasets were chosen. A subset of VGGFace2 that was not used for training was used to generate the *Wild* benchmark.

In order to generate the *Strict* benchmark, the Face Recognition Grand Challenge (FRGC) version 2 [65] dataset was used. From this dataset, only ICAO compliant images were chosen.

To generate the relaxed ICAO compliance benchmark, the FRGC V2 dataset was also used. In this benchmark, all the images included in the *Strict* benchmark were used with the inclusion of other images. The added images chosen were allowed the following deviations from strict ICAO compliance: small variations from the frontal pose, the existence of facial expressions, non-frontal face lighting, some degree of blurriness and non-uniform background.[5] Example face images from the three benchmarks mentioned above are represented in figure 4.5.

After selecting the datasets, a protocol was generated for each benchmark. To generate the protocol, first, a certain ratio of positive (same identity) to negative (different identities) pairs were chosen. For the three benchmarks, a ratio of 1/2 was used, meaning that in the protocol for each positive comparison existed two negative comparisons. The pairs of images were chosen randomly through the use of random number generation. The total number of comparisons was chosen to be around 330.000, which resulted in approximately 110.000 positive pairs and 220.000 negative pairs.

To evaluate a model's performance, the trained model is used to extract the feature embeddings of all images that comprise the benchmark. Afterwards, the cosine similarity scores are calculated for all the pairs in the protocol. Then, for a given threshold of the similarity score, all values below said threshold are considered different identities, and all values above it are considered the same identity. This is compared with the real pairwise label (same or different identity), and the FMR and FNMR are obtained. Finally, by sweeping the threshold across the entire cosine similarity range, pairs of the FMR and FNMR values are extracted, which are the metrics used to evaluate the models' performance.

[5]The choice of images from the FRGC V2 dataset was done in the context of other works in the Computer Vision investigation group and used as-is.

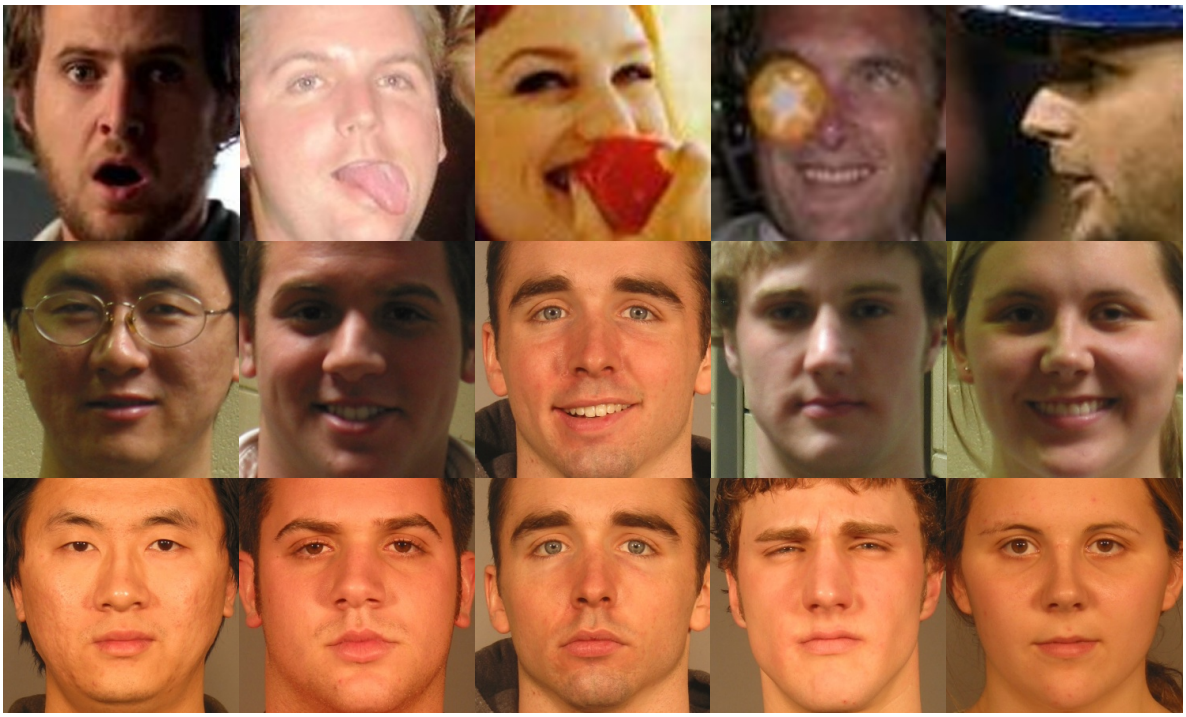


Figure 4.5: Example images from the three benchmarks: First row - *Wild* Benchmark; Second row - *Relaxed* Benchmark; Third row - *Strict* Benchmark.

4.6 Tools/Technical implementation

The code for defining and training all the models used was developed in python 3.7.9 with the TensorFlow 2.5 and Keras libraries. The training was performed on a NVIDIA RTX 3090 GPU. The ResNet50V2 model was imported from Keras. Then, an additional layer for the feature embedding (512 nodes fully connected) was added. Finally, the ArcFace and adaptive margin losses were added as a custom Keras layer. For the adaptive margin loss, an extra input was added for the score/scores of each image which was then used to calculate the margin, while for ArcFace, the margin was a constant value. The logits calculated are then used as input for a softmax function, and afterwards categorical cross-entropy is applied to get the final loss value.

Chapter 5

Results

5.1 Data analysis

In the work developed, four datasets were used: One for training and three for benchmarking. The main properties of these datasets are presented in table 5.1.

Table 5.1: Details of the datasets used for training and benchmarking.

Dataset	Train	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
Number of images	1338468	31117	35410	11732
Number of identities	2842	147	568	565
Average of images per identity	470.9±57.7	211.7±48.4	62.3±48.6	20.7±16.7

To better understand the datasets used in model training and benchmarking, the meta-information mentioned in section 4.3 was extracted, after cropping, alignment and normalisation of the datasets as described in section 4.2.1. For visualisation and comparison purposes, after extracting this information, the scores were normalised to a 0-1 range with respect to the minimum and maximum values found in the training dataset (VGGFace2). Then, the scores for which lower values signified higher quality face images (BRISQUE and Pose) were inverted. The distributions of these normalised quality scores are represented in figure 5.1 and the mean and standard deviation for each distribution are also represented in table 5.2.

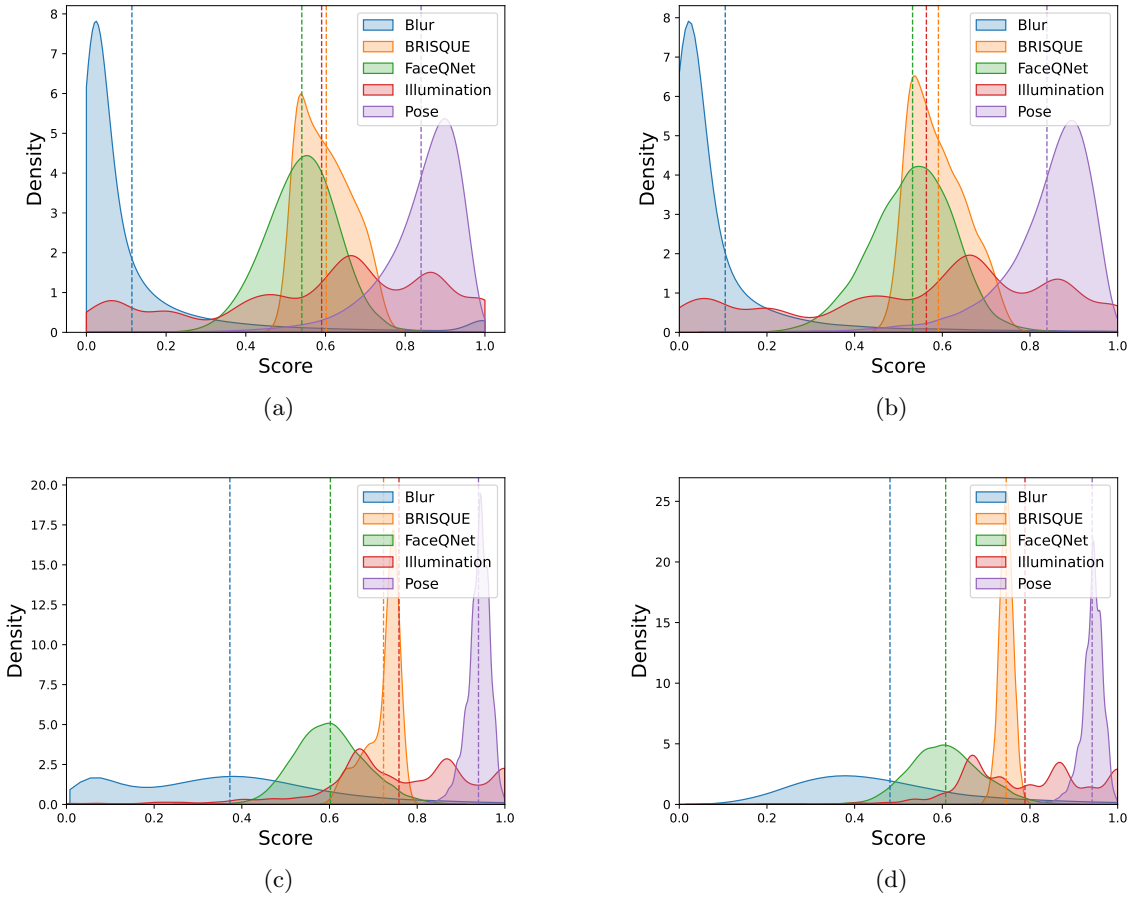


Figure 5.1: Normalised score distributions (vertical lines represent the mean of each distribution): a) VGGFace2 train; b) VGGFace2 *Wild* benchmark; c) FRGC V2 *Relaxed* benchmark; d) FRGC V2 *Strict* benchmark.

Table 5.2: Mean and standard deviation of normalised scores.

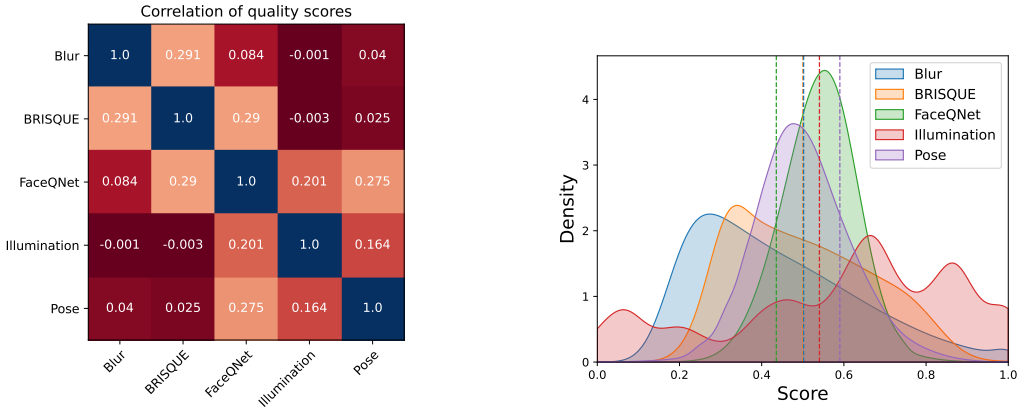
Datasets		Train Dataset	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
Blur	Mean	0.0016	0.1325	0.3727	0.4806
	Standard Deviation	0.0050	0.5914	0.2651	0.2424
BRISQUE	Mean	0.6018	0.5910	0.7232	0.7457
	Standard Deviation	0.0662	0.0648	0.0397	0.0162
FaceQNet	Mean	0.5405	0.5321	0.6019	0.6076
	Standard Deviation	0.0913	0.0954	0.0811	0.0835
Illumination	Mean	0.5902	0.5636	0.7584	0.7885
	Standard Deviation	0.2840	0.2883	0.1830	0.1538
Pose	Mean	0.8398	0.8386	0.9398	0.9418
	Standard Deviation	0.1054	0.1081	0.0263	0.0254

By analysing these distributions, some observations were made:

- The score distributions in the train dataset and *Wild* benchmark are almost identical as expected, since they are subsets of the same dataset.

- Both FRGC V2 benchmarks have higher face image quality than the training dataset, for all scores tested.
- The *Strict* benchmark has a small but noticeable superiority (more shifted to the right) in terms of score distribution when compared to the *Relaxed* benchmark.

The correlation between pairs of quality scores was also analysed and is plotted in Fig 5.2a. This value can go from -1 for complete inverse correlation to 1 for complete correlation. The 0 value corresponds to no correlation. A small degree of correlation between some pairs of scores can be seen from the figure, as expected. The small correlation between blur and BRISQUE can be explained since the BRISQUE scores also include information regarding image blur. Apart from this, the correlation between FaceQNet and the other scores can be justified since FaceQNet is a generic face quality indicator based on ICAO compliance, and as such, implicitly learns information regarding pose, illumination and blurriness/intrinsic image quality.



(a) (b)

Figure 5.2: a) Correlation of the 5 scores extracted from the datasets before any transformation or normalisation. b) Distribution of the 5 normalised and transformed scores used for training.

5.2 Model training and results

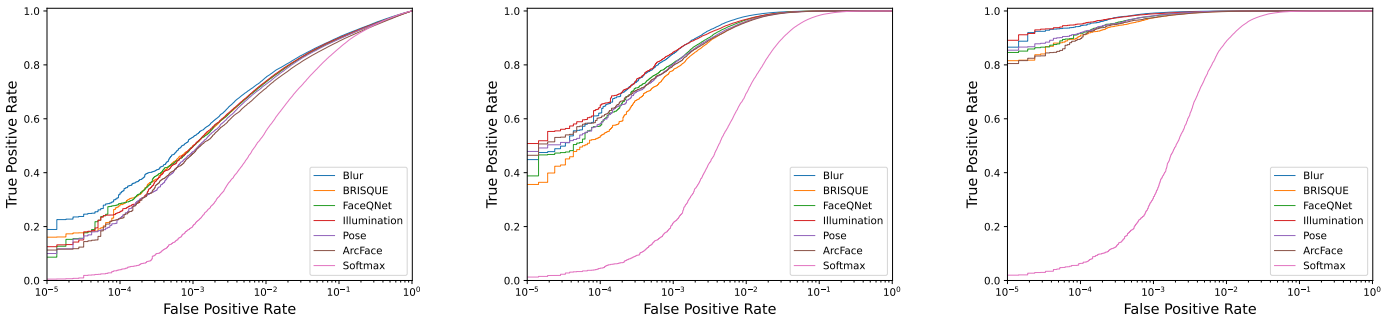
For all the following experiments (unless explicitly mentioned), the same training setup was used. As mentioned previously, the backbone used was the ResNet50-V2 with a fully connected 512 node layer. These models were initialised with the imagenet pre-trained weights. The batch size used was 24 images. The learning rate decay was done with a cosine annealing scheduler [66], from $5e-3$ to $1e-5$. The optimiser used was Mini-Batch Gradient Descent and with momentum parameter of 0.5 and weight decay of 0.0005.

After cropping the images to 299×299 pixels, aligning and normalising the dataset was used for training. The five quality scores mentioned in section 4.3 were extracted, transformed and normalised to then be used as additional inputs in training. Each score was transformed independently in such a way to reduce the skewness of the data and centre it with respect to its mean (transformed scores represented in Fig 5.2b).

Single Score Model Training

To test the value of the information given by each of the five quality scores, two different models were trained per score. These two models were trained using the loss function formulation in equation 4.1, with $m_0 = 0.4,\ m_1 = 0.1$ and $m_0 = 0.4,\ m_1 = 0.2$. These values for m_0 and m_1 were chosen for a specific reason: in ArcFace, the authors found that the margin value $m = 0.5$ produced the best results. By setting $m_0 = 0.4,\ m_1 = 0.1$ the max margin will be $max(m_i) = 0.5$ since the scores were normalised to the $[0, 1]$ interval. By setting $m_0 = 0.4,\ m_1 = 0.2$ and taking into account that the scores were transformed to have an average value close to 0.5, this combination of parameters leads to an average margin value close to ~ 0.5 and $max(m_i) = 0.6$.

The FNMR@FMR metrics are represented in tables 5.3 and 5.4 for the 3 benchmarks, respectively. The ROC curves for the $m_1 = 0.1,\ m_1 = 0.2$ models are also represented in figures 5.3 and 5.4, respectively.



(a) (b) (c)

Figure 5.3: ROC curves for the single score $m_1 = 0.1$ adaptive margin trained models plus ArcFace and Softmax. a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.

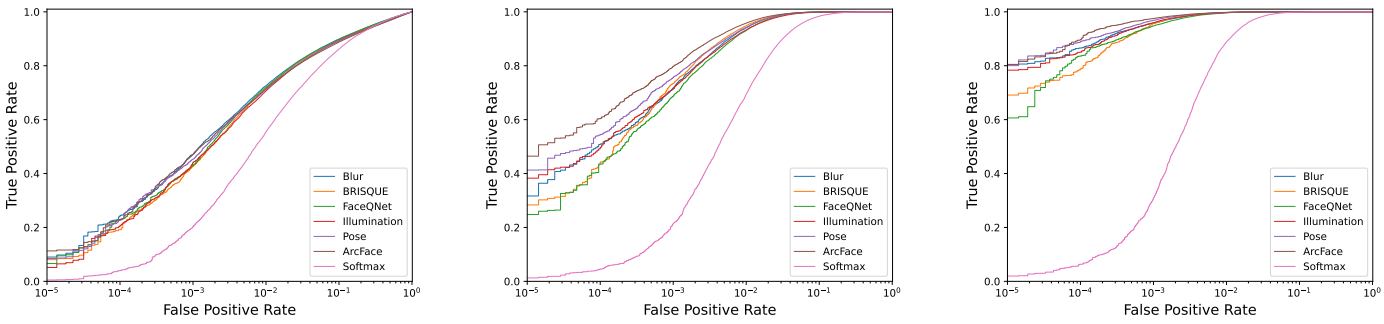
From the ROC curves in figure 5.3 it is possible to conclude most of the adaptive margin models have better operation curves than either the ArcFace or Softmax models. Namely, the blur score model is the most effective in the *Wild* benchmark scenario while the illumination model is generally the best performer in the *Relaxed* and *Strict* benchmarks. (The AUC scores of all models trained are represented in the Appendix).

Table 5.3: FNMR@FMR thresholds for ArcFace, Softmax and the adaptive margin models (underlined) with $m_0 = 0.4,\ m_1 = 0.1$.

Method	*Wild* Benchmark		*Relaxed* Benchmark			*Strict* Benchmark		
	1e-2	1e-3	1e-3	1e-4	1e-5	1e-3	1e-4	1e-5
Softmax	0.44502	0.79633	0.78645	0.95409	0.98727	0.69017	0.93655	0.98027
ArcFace	0.28680	0.52938	0.20263	0.39509	0.53552	0.02486	0.10205	0.19507
Blur	**0.24600**	**0.46806**	0.15828	0.37839	0.55140	**0.00793**	0.05453	0.13429
BRISQUE	0.26185	0.49934	0.21957	0.46290	0.64373	0.02556	0.08950	0.18444
FaceQNet	0.26383	0.50290	0.19293	0.42316	0.61194	0.01874	0.08284	0.15398
Illumination	0.26037	0.50076	**0.15251**	**0.34849**	**0.49179**	0.01066	**0.04835**	**0.10878**
Pose	0.27177	0.52186	0.19694	0.41607	0.52128	0.01805	0.08011	0.14550

From table 5.3 the conclusions taken from the ROC curves can indeed be confirmed. The blur and illumination models are the best performers at all the thresholds tested.

Analysing the ROC curves from the models with $m_1 = 0.2$ (see Fig 5.4) the results are generally worse than with $m_1 = 0.1$. Still, apart from the *Relaxed* benchmark, it is possible to generally see some situations where the adaptive margin method is superior, namely with the blur and pose scores, with the first being better in the unconstrained scenario and the latter outperforming in the higher quality scenarios of *Relaxed* and *Strict* benchmarks.



(a) (b) (c)

Figure 5.4: ROC curves for the single score $m_1 = 0.2$ adaptive margin trained models plus ArcFace and Softmax. a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.

The metrics from the table 5.4 indeed confirm the conclusions made from the ROC curves. Although this margin value shows some good results, for all the thresholds tested in all benchmarks the models with $m_1 = 0.1$ outperformed. Another downside for the $m_1 = 0.2$ formulation comes from the fact that these models had some problems with convergence with the same setup as the previous models.

In the single score experiments the validity of the proposed method is proven. In the same training conditions, the adaptive margin methods have enhanced verification performance in strictly ICAO compliant and relaxed ICAO compliant face images compared to ArcFace. This fact is evident in the

Table 5.4: FNMR@FMR thresholds for ArcFace, Softmax and the adaptive margin models (underlined) with $m_0 = 0.4,\ m_1 = 0.2$.

Method	*Wild* Benchmark		*Relaxed* Benchmark			*Strict* Benchmark		
	1e-2	1e-3	1e-3	1e-4	1e-5	1e-3	1e-4	1e-5
Softmax	0.44502	0.79633	0.78645	0.95409	0.98727	0.69017	0.93655	0.98027
ArcFace	0.28680	0.52938	**0.20263**	**0.39509**	**0.53552**	0.02486	**0.10205**	0.19507
Blur	**0.27460**	0.52942	0.28090	0.49019	0.68323	0.04183	0.13423	0.19618
BRISQUE	0.28253	0.57363	0.26548	0.55709	0.71651	0.04329	0.21139	0.30880
FaceQNet	0.27945	0.56911	0.31194	0.56485	0.75202	0.05226	0.16409	0.39373
Illumination	0.29351	0.56006	0.28666	0.50236	0.61717	0.04046	0.14946	0.21644
Pose	0.28781	**0.51604**	0.24307	0.45505	0.58721	**0.01146**	0.11058	**0.11783**

models with $m_1 = 0.1$. Taking into consideration wild benchmarks, the proposed approach also outperforms ArcFace in these training conditions. It is possible to conclude that, in these conditions, the adaptive margin method allows to regularise the training process in a deeper manner. In other words, the models are not just adapting to qualitative samples but are generally learning more qualitative and discriminative face features.

Combined Scores Model Training

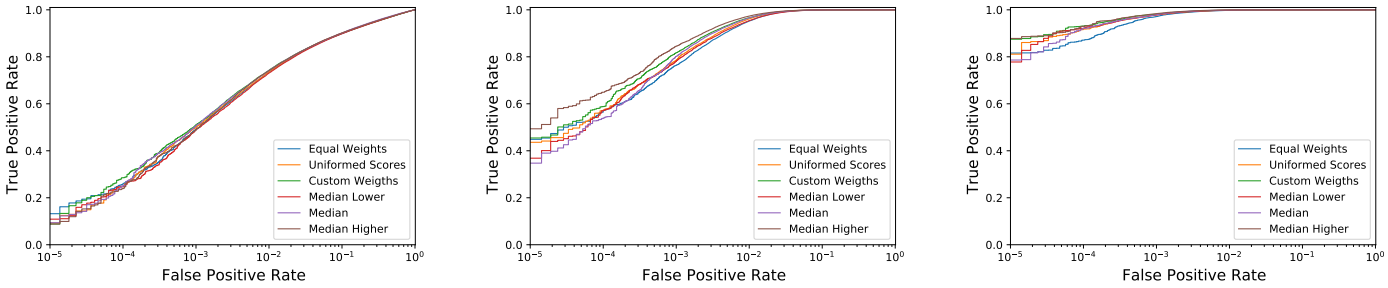
After verifying the validity of the loss function formulation and the usefulness of the information conveyed in the extracted scores, tests were made using the combination of all scores.

To test the utilisation of all scores simultaneously, several forms of mixing the scores were tried. Specifically, the simple mean, a weighted mean, and three different types of median value. Three median combinations of the scores were formulated as follows:

A model that averaged the three lower scores, named *Median Lower*, one where the three centre scores were averaged - the *Median* model and the *Median Higher* model which averaged the three highest scores. This sorting and averaging of scores was done on for each sample individually. Experiments with uniforming score distributions in the range $[0, 1]$ before averaging for equalising their impact were also made. All the models tested in these conditions have $m_1 = 0.1$ since its superiority was proven. The ROC curves for all the models mentioned above are represented in figure 5.5 and the FNMR@FMR metrics are presented in table 5.5.

Table 5.5: FNMR@FMR thresholds for the combined score models in the three benchmarks.

Models	*Wild* Benchmark		*Relaxed* Benchmark			*Strict* Benchmark		
	1e-2	1e-3	1e-3	1e-4	1e-5	1e-3	1e-4	1e-5
Equal Weights	0.26495	0.50912	0.23612	0.42770	0.55127	0.02869	0.12879	0.18398
Uniformed Scores	0.26393	0.50244	0.21462	0.42619	0.56392	0.02171	0.08221	0.18881
Custom Weights	**0.25735**	**0.48964**	0.18288	0.41150	0.54557	0.01834	**0.06875**	0.12521
Median Lower	0.26914	0.51016	0.21888	0.42815	0.63156	0.02195	0.07807	0.22204
Median	0.25829	0.49400	0.19761	0.46021	0.65266	0.02087	0.07853	0.21371
Median Higher	0.25877	0.51080	**0.15494**	**0.35069**	**0.50644**	**0.01629**	0.07027	**0.12184**



(a) (b) (c)

Figure 5.5: ROC curves for the combined score adaptive margin trained models. a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.

In the combined scores experiments, the *Custom Weights* and *Median Higher* were the best performers in the *Wild* and *Relaxed*/*Strict* benchmarks, respectively. Various observations were made from the results: Uniforming the scores distributions did result in slightly better results than the normal averaging of scores but not a significant difference.

Weighting the scores before averaging proved to be a good improvement from the *Equal Weights* model. The scores were weighted in accordance to their corresponding individual model performance (in the experiments, the made best performing *Custom Weights* model had the following weighting: Blur - 0.3, BRISQUE - 0.1, FaceQnet - 0.15, Illumination - 0.3, Pose - 0.15).

Regarding the models with median averaging, the *Median Higher* model had the most promising results. This means that the sampling strategy used in the median models should be biased towards better scores.

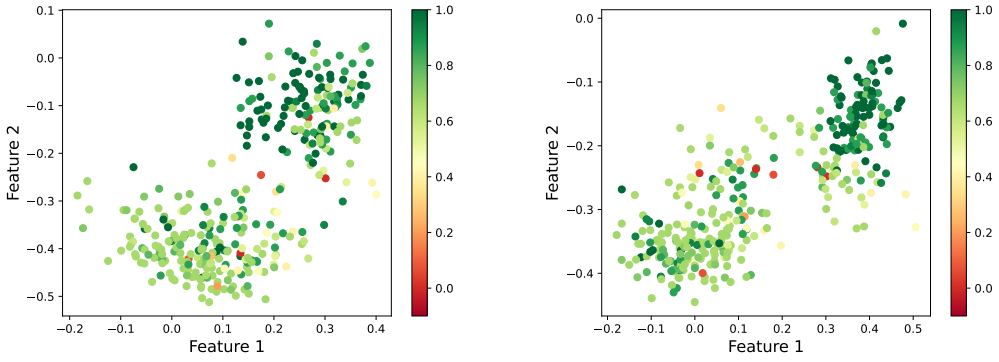
Although the use of combined scores did not outperform the single score models in any particular benchmark, it allowed some combined score models to achieve more uniform results across the three benchmarks. This result could be useful for applications in an unspecified scenario where more universal face representation is key. For example, the *Custom Weights model* outperforms the blur model in the $FMR = 1e-5$ threshold in both document security related benchmarks, combining the strengths of the blur and illumination score.

5.2.1 Other Experiments

Other types of experiments were made to better understand the behaviour and characteristics of the developed method.

Feature distribution

The goal of this experiment was to test if the feature distribution of the developed method pulled higher quality samples towards the class centre while pushing low-quality samples away, like hypothesised in figure 4.4.



(a) (b)

Figure 5.6: Feature distribution of 2 FRGC V2 identities (04430 and 02463), the score represented is the illumination quality score: a) ArcFace model; b) Adaptive margin illumination model.

To do so, the 512 dimensional feature distribution of all the images from the FRGC V2 dataset was extracted. In order to better visualise the distributions, through the use of Principal Component Analysis, the embeddings were reduced to 2 dimensions. The distribution plots with 2 identities are represented in figure 5.6.

From the feature distributions, it is possible to make two conclusions. Firstly, the spatial separation between identities common in margin-based methods is visible in both ArcFace and adaptive margin models, as expected. Secondly, ArcFace does not take into account the quality of the samples: some low-quality samples are close to the class centre, and there is some dispersion in high-quality samples. However, in the adaptive margin case, the high-quality samples are more compacted together and closer to the class centre while lower quality samples are further away, as hypothesised.

This is an interesting result since it shows the process of how the verification performance increases for high quality samples: higher quality samples of different identities are further apart than in the ArcFace scenario (since they are spread in a more compact manner around the class centre), which leads to more discriminative power for the target ID and travel document scenario.

Additional experiments

For better understanding of the developed method, two more formulations of the adaptive margin were tested. First, a model with margin parameter $m_1 = 0.15$. Another model was designed to test

the use of non-linear adaptive margin. The loss function remains the same as 4.1 but unlike equation 4.2, the m_i parameter was designed as follows:

$$m_i = m_0 + \sum_{j}^{Q} w_j(q_{ij}m_1 + q_{ij}^2 m_2) \tag{5.1}$$

where m_2 is an additional hyper-parameter. The goal of this formulation is to increase the impact of the good scores. The testing of these formulations was made on models trained with blur score. In table 5.6 the FNMR@FMR metrics are represented for the models mentioned above:

Table 5.6: FNMR@FMR thresholds for different blur models in the three benchmarks ($m_0 = 0.4$).

Method	*Wild* Benchmark		*Relaxed* Benchmark			*Strict* Benchmark		
	1e-2	1e-3	1e-3	1e-4	1e-5	1e-3	1e-4	1e-5
$m_1 = 0.1$	**0.24600**	**0.46806**	**0.15828**	**0.37839**	0.55140	**0.00793**	**0.05453**	**0.13429**
$m_1 = 0.15$	0.26941	0.51831	0.19995	0.40163	**0.49769**	0.01820	0.07151	0.14045
$m_1 = 0.2$	0.27460	0.52942	0.28090	0.49019	0.68323	0.04183	0.13423	0.19618
$m_1 = 0.1, m_2 = 0.1$	0.25800	0.48868	0.22497	0.43605	0.57471	0.02445	0.09612	0.20571

From these experiments, it is possible to conclude that increasing the impact of the score in the margin does not translate in better results, instead increasing levels of margin result in poorer performances as well as poorer convergence during training.

Inverted Quality Scores

Another experiment was made to better understand the impact of the quality scores. All 5 scores used in this work, were inverted (subtracted 1 and multiplied by -1) and then adaptive margin models with $m_0 = 0.4, m_1 = 0.1$ were trained with this scores. Intuitively, the expectation is that these models might perform better than the standard models on the *Wild* benchmark due to harder samples having higher weight in the training process and under-perform in *Relaxed* and *Strict* scenarios. The results from this experiment are represented in table 5.7.

Table 5.7: FNMR@FMR thresholds for the inverted score models ($m_0 = 0.4, m_1 = 0.1$) in the three benchmarks. The bold numbers highlight the conditions where the inverted models outperformed the base models.

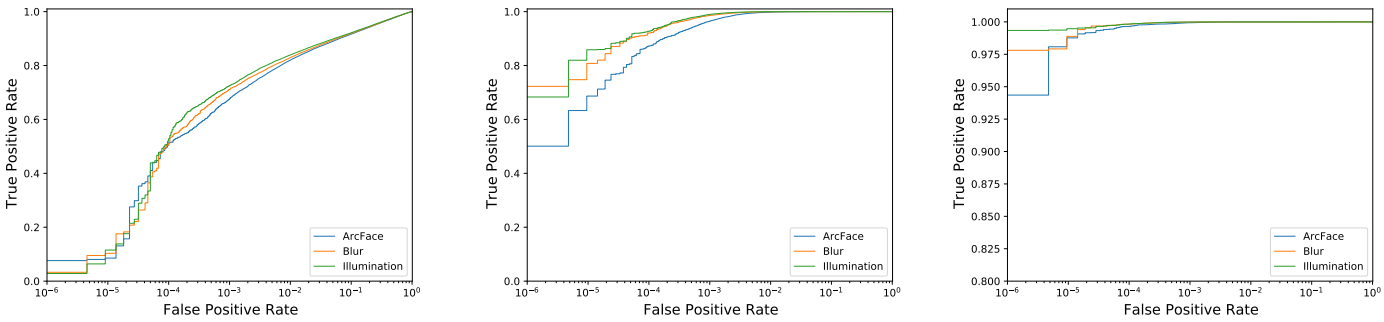
Method	*Wild* Benchmark		*Relaxed* Benchmark			*Strict* Benchmark		
	1e-2	1e-3	1e-3	1e-4	1e-5	1e-3	1e-4	1e-5
Blur	0.25676	0.49993	0.21335	0.45042	0.65084	0.02279	0.08321	0.20641
BRISQUE	**0.25419**	**0.49398**	0.19377	0.41990	0.58251	**0.01800**	**0.07416**	0.21906
FaceQNet	0.26650	0.50846	0.21402	0.43829	0.63282	**0.01762**	0.08537	0.17598
Illumination	**0.25565**	**0.47977**	0.16520	0.37467	0.55442	0.01351	0.06568	0.14062
Pose	**0.25818**	**0.50916**	0.20509	0.42735	0.57781	0.02310	0.09207	0.19440

Comparing these results with table 5.3 it is possible to conclude that although the performance

difference between the normal and the inverted models is not equal for all scores, generally a trend can be recognised. For the pose, illumination and BRISQUE scores, focusing on poorer quality images during training increases the performance in unconstrained conditions. It is also interesting to see that for the inverted blur score case the results were poorer in all benchmarks, this can be understood as blurrier images not conveying enough facial features and thus hindering the training process. Although the BRISQUE and FaceQNet have some better results in the *Strict* Benchmark, this is not maintained for all thresholds and is not verified for the lower threshold of FMR= $1e-5$. As such, it is possible to understand that generally, inverting the scores results in better performance for the wild scenario and lower performance for the stricter conditions.

Longer Refined Training

Finally, after proving the validity of the adaptive margin method, a longer, more refined training setup was designed to show that the increase in performance from the method presented is transferable and can achieve much better results. The training setup designed used the full VGGFace 2 train set compromised of 8.631 identities and 2.474.216 images used for training purposes, and 618.554 images for validation purposes. Instead of 6 epochs, 20 epochs were used, and the learning rate was started at $1e-1$ and decreased until $1e-5$. The same batch size and network were used (only the last layer was replaced with an 8631-dimensional layer). The momentum parameter in Mini-Batch Gradient Descent was set to 0.9.



(a) (b) (c)

Figure 5.7: ROC curves for the refined training models: a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.

With the more refined training setup, the results are clearly superior to the previous results trained on the "cropped" VGGFace2 dataset; ArcFace and adaptive margin models outperform all previous models at all thresholds. This is an expected result given the importance of the dataset in the training of deep neural networks.

Table 5.8: FNMR@FMR thresholds for ArcFace and the adaptive margin blur and illumination models with $m_0 = 0.4,\ m_1 = 0.1$ for the longer and refined training conditions

Method	*Wild* Benchmark			*Relaxed* Benchmark				*Strict* Benchmark			
	1e-2	1e-3	1e-4	1e-3	1e-4	1e-5	1e-6	1e-3	1e-4	1e-5	1e-6
ArcFace	0.17989	0.32435	**0.47122**	0.03569	0.12723	0.31330	0.49924	0.00065	0.00350	0.01232	0.05647
Blur	0.17018	0.28878	0.48877	0.01418	0.07902	0.19204	**0.27731**	0.00017	0.00172	0.01117	0.02192
Illumination	**0.16061**	**0.27483**	0.48542	**0.01033**	**0.07237**	**0.14148**	0.31670	**0.00010**	**0.00170**	**0.00526**	**0.00658**

Comparing the ArcFace model with the adaptive margin models, the performance difference across the *Relaxed* and *Strict* benchmarks at all thresholds is evident. In these benchmarks, the illumination and blur adaptive margin models are considerably better than ArcFace, and the performance enhancement of the adaptive margin methods is even more noticeable than for the first experiments. This result can indicate that over larger datasets with more identities and more pronounced image variations, the effects of sample-specific techniques (and subsequently the method developed) can become more noticeable. Namely, the increase in identities represented in the 512-dimensional feature space make it more challenging to separate classes in the embedding hyper-sphere, which in turn makes the effect of the adaptive margin feature distribution more noticeable. When comparing the blur and illumination models, the relative performance trend seen in the first experiment is repeated. For the *Strict* and *Relaxed* benchmarks, the results are still favourable for the illumination model for most thresholds except for the *Relaxed* scenario at $FMR = 1e-6$ case.

By analysing the *Wild* benchmark results, the performance difference between the three models is more negligible. ArcFace is the best model for the lower FMR thresholds ($FMR \leq 1e-4$ while the illumination model achieves better results for the higher FMR thresholds. This statement can be confirmed in table 5.8 and from figure 5.7a, and is a different result from earlier experiments. Before commenting on these results, it is relevant to mention that lower FMR thresholds present a more challenging task where performance on harder samples is crucial for good results.

So, the results on the *Wild* benchmark suggest that the adaptive margin models, in more complete training conditions, do indeed learn more discriminative features (showed by their performance in higher FMR thresholds when compared to ArcFace. However, for lower FMR thresholds (more challenging conditions), ArcFace proves to be superior. So, it is possible to conclude that although the adaptive margin method helps to improve the performance on unconstrained face recognition by learning more discriminative face features from higher quality samples when performance on bad quality samples matters most, ArcFace still proves to be superior.

Chapter 6

Conclusion

6.1 General Conclusion

In this work, a method for optimising face recognition for document security applications based on deep neural networks was developed. A novel loss function solution was created by modifying the angular margin loss and introducing sample quality to alter the margin adaptively in a sample-specific way. Furthermore, five different quality estimators were used (Blur, BRISQUE, FaceQNet, Illumination Quality, Pose). Following the loss formulation, benchmarks were designed for the specific purpose of document security and models were trained on each individual score. Other experiments, including the combination of scores, feature distribution visualisation and inverting scores, were made.

In the extensive experiments with hyper-parameters and quality metrics, the adaptive margin method generally proved to be superior in the ICAO compliance based benchmarks for almost all thresholds and scores utilised. Training with the illumination quality score proved the best in both document security benchmarks, significantly outperforming the standard margin method. This result proves that the developed method significantly enhances face verification performance compared with regular margin-based methods like ArcFace, for the ICAO compliant face image case.

Another interesting result obtained is that the method developed also outperformed ArcFace in the unconstrained scenario benchmark. This result is more evident in the Blur score case. This result allows the conclusion that the developed method improves learning, generally allowing the network to learn better, more discriminative face embeddings.

Training with all scores combined, although not outperforming the best single models, showed that combining the scores resulted in more uniform results across different scenarios. Namely, using weighted averaging of scores and averaging the best 3 scores per image resulted in the best results for combined score models.

In the feature distribution visualisation, it was possible to see how the adaptive margin results in changing the sample distribution in such a way that approximates higher quality samples towards the class centre and pushes bad quality samples away. Additional experiments suggested that increasing the score impact by increasing the m_1 hyper-parameter or adding additional squared score dependency resulted in poorer performance and worst model convergence.

The experiments with inverting the quality scores showed that for most scores used, training a model while emphasising worst quality scores generally improves verification performance in the wild scenario.

Finally, the results for the longer refined training showed that the performance benefits of the adaptive margin model are transferable and enhanced in larger datasets. The illumination score model convincingly outperformed ArcFace in both document security benchmarks. The results on the wild scenario solidified the claim that the developed method increases the quality of the learned face features to make them more discriminative, however showed , for lower FMR threshold where performance on bad samples is important, ArcFace is still superior.

In conclusion, the method proposed in this work improves face recognition performance compared to regular marginal loss functions for the document security use case without significant performance loss in the universal face recognition scenario.

6.2 Future Work

Although the advantage of the developed method was demonstrated, there are further experiments and ideas to pursue. The future work will focus on the study of different image quality metrics, for example, SER-FIQ, PFE's or SDD-FIQA that were mentioned in the state-of-the-art. It will also include the study of other possible score transformation techniques as well as other possible score combinations. The validity of the method on other margin-based loss functions, for example, on Equalised Margin Loss, will also be tested. Finally, including the concept of feature vector magnitude as a quality indicator (presented in MagFace) will be explored.

Bibliography

[1] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. Face recognition systems: A survey. *Sensors (Switzerland)*, 20(2), 2020.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[3] Yichun Shi and Anil K. Jain. DocFace: Matching ID document photos to selfies. *arXiv*, 2018.

[4] Portrait quality - reference facial images for mrtd. `https://www.icao.int/Security/FAL/TRIP/Documents/TR%20-%20Portrait%20Quality%20v1.0.pdf`. Version: 1.0 Date – 2018-04, Accessed: 2021-04-04.

[5] Understanding the face image format standards. `https://www.nist.gov/system/files/documents/2021/02/25/ansi-nist_2007_griffin-face-std-m1.pdf`. Accessed: 2021-09-20.

[6] Iurii Medvedev, Farhad Shadmand, Leandro Cruz, and Nuno Gonçalves. Towards facial biometrics for id document validation in mobile devices. *Applied Sciences*, 11(13), 2021.

[7] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. pages 41.1–41.12, 01 2015.

[8] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. *CoRR*, abs/1710.08092, 2017.

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9907 LNCS:87–102, 2016.

[10] Mei Wang and Weihong Deng. Deep face recognition: A survey. *CoRR*, abs/1804.06655, 2018.

[11] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. pages 1–31, 2016.

[12] Hossein Gholamalinezhad and Hossein Khosravi. Pooling methods in deep neural networks, a review. *CoRR*, abs/2009.07485, 2020.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[15] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

[16] Biometric authentication. `https://www.nist.gov/system/files/applyingmeasurementscienceworkshopjan12_13_2016.pdf`. Version: 1.0 Date – 2015-01, Accessed: 2021-04-15.

[17] Erjin Zhou, Zhimin Cao, and Qi Yin. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *CoRR*, abs/1501.04690, 2015.

[18] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014.

[19] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. *CoRR*, abs/1807.11649, 2018.

[20] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *CoRR*, abs/1512.00596, 2015.

[21] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. *CoRR*, abs/1705.00393, 2017.

[22] Trillion pairs challenge. `http://trillionpairs.deepglint.com/overview`. Accessed: 2021-01-24.

[23] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[24] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011.

[25] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015.

[26] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017.

[27] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, Feb 2018.

[28] Samadhi P K Wickrama Arachchilage. *Deep-learned faces : a survey.* EURASIP Journal on Image and Video Processing, 2020.

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[30] Yousri Ouerhani, Ayman Alfalou, and Christian Brosseau. Road mark recognition using HOG-SVM and correlation. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 10395 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 103950Q, August 2017.

[31] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *CoRR*, abs/1604.02878, 2016.

[32] Rajeev Ranjan, Swami Sankaranarayanan, Ankan Bansal, Navaneeth Bodla, Jun Cheng Chen, Vishal M. Patel, Carlos D. Castillo, and Rama Chellappa. Deep learning for understanding faces: Machines may be just as good, or better, than humans. *IEEE Signal Processing Magazine*, 35(1):66–83, January 2018.

[33] Xin Jin and Xiaoyang Tan. Face alignment in-the-wild: A survey. *CoRR*, abs/1608.04188, 2016.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 2016.

[38] Gary B Huang, Marwan Mattar, Tamara Berg, Eric Learned-miller Labeled, Real-life Images, and Erik Learned-miller. Labeled Faces in the Wild : A Database for Studying Face Recognition in Unconstrained Environments. *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:815–823, 2015.

[40] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:6738–6746, 2017.

[41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. *arXiv*, 2018.

[42] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *arXiv*, 2018.

[43] Eric López-López, Xosé M. Pardo, Carlos V. Regueiro, Roberto Iglesias, and Fernando E. Casado. Dataset bias exposed in face verification. *IET Biometrics*, 8(4):249–258, 2019.

[44] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *arXiv*, 2017.

[45] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network. *arXiv*, pages 692–702, 2018.

[46] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. Ijb–s: Iarpa janus surveillance video benchmark. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2018.

[47] T. Zheng and W. Deng. Cross-pose lfw : A database for studying cross-pose face recognition in unconstrained environments. 2018.

[48] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6816–6825, 2020.

[49] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5900–5909, 2020.

[50] Jingna Sun, Wenming Yang, Jing Hao Xue, and Qingmin Liao. An Equalized Margin Loss for Face Recognition. *IEEE Transactions on Multimedia*, 22(11):2833–2843, 2020.

[51] Dan Zeng, Hailin Shi, Hang Du, Jun Wang, Zhen Lei, and Tao Mei. NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition. *arXiv*, pages 1–11, 2020.

[52] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, June 2021.

[53] Yichun Shi and Anil K. Jain. DocFace+: ID document to selfie* matching. *arXiv*, 1(1):56–67, 2018.

[54] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face image quality assessment: A literature survey. 09 2020.

[55] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. 2020.

[56] Yichun Shi, Anil K. Jain, and N. Kalka. Probabilistic face embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910, 2019.

[57] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, June 2021.

[58] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks, 2016.

[59] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *CoRR*, abs/1905.00641, 2019.

[60] J. L. Pech-Pacheco, G. Cristöbal, J. Chamorro-Martínez, and J. Fernândez-Valdivia. Diatom autofocusing in brightfield microscopy: A comparative study. *Proceedings - International Conference on Pattern Recognition*, 15(3):314–317, 2000.

[61] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[62] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. *arXiv*, 2019.

[63] Lijun Zhang, Lin Zhang, and Lida Li. Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10636 LNCS:583–593, 2017.

[64] Euclides Arcoverde, Rafael Duarte, Rafael Barreto, Joao Magalhaes, Carlos Bastos, Tsang Ing Ren, and George Cavalcanti. Enhanced real-time head pose estimation system for mobile device. *Integrated Computer Aided Engineering*, 21:281–293, 04 2014.

[65] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, Jin Chang, K. Hoffman, J. Marques, Jaesik Min, and W. Worek. Overview of the face recognition grand challenge. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 947–954 vol. 1, 2005.

[66] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016.

Appendix

Table 6.1: AUC scores for the adaptive margin single score models of the $m_0 = 0.4, m_1 = 0.1$ (underlined) and $m_0 = 0.4, m_1 = 0.2$ models, for the 3 benchmarks.

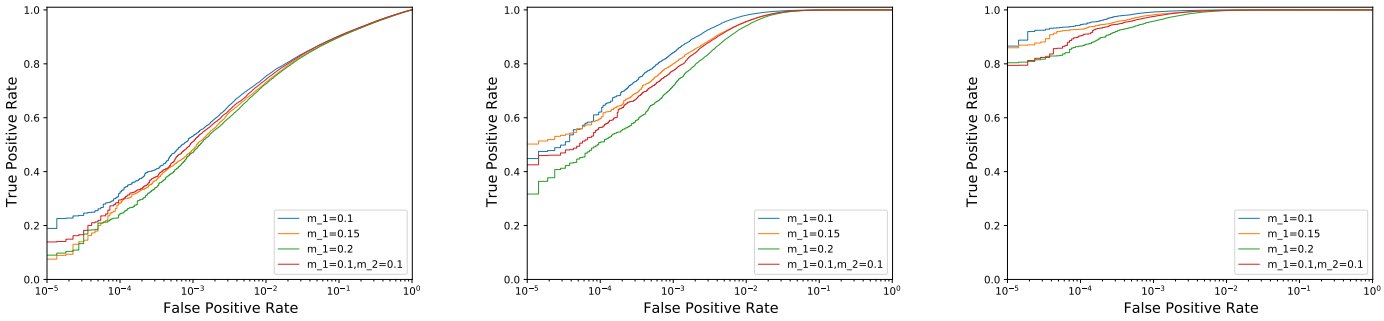
Method	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
Blur	**0.959089**	**0.999007**	**0.999957**
BRISQUE	0.954925	0.998387	0.999878
FaceQNet	0.956515	0.998600	0.999910
Illumination	0.956797	0.998717	0.999936
Pose	0.954926	0.998385	0.999917
Blur	0.956314	0.997744	0.999813
BRISQUE	0.953276	0.997853	0.999772
FaceQNet	0.956252	0.997410	0.999743
Illumination	0.952535	0.997372	0.999792
Pose	0.951155	0.997860	0.999838

Table 6.2: AUC scores for the adaptive margin combined score models for the 3 benchmarks.

Method	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
Equal Weights	0.956074	0.998130	0.999850
Uniformed Scores	0.957280	0.998444	0.999897
Custom Weights	0.956706	0.998599	0.999907
Median Lower	0.955681	0.998203	0.999891
Median	0.958679	0.998518	0.999905
Median Higher	0.957604	0.998849	0.999930

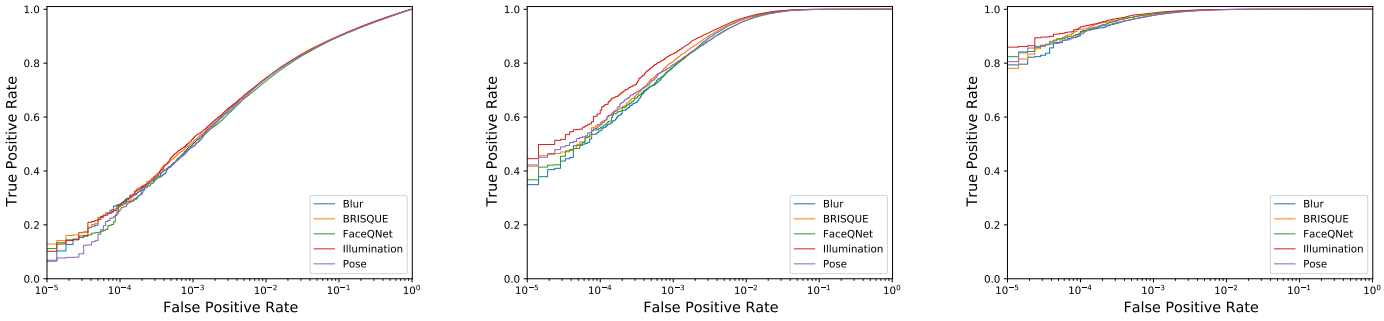
Table 6.3: AUC scores for the "small experiments" models, for the 3 benchmarks

Method	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
0.1	0.959089	0.999007	0.999957
0.15	0.957346	0.998266	0.999882
0.2	0.956314	0.997744	0.999813
0.1, 0.1	0.959106	0.998262	0.999891



(a) (b) (c)

Figure 6.1: ROC curves for the "small experiments" models: a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.



(a) (b) (c)

Figure 6.2: ROC curves for the adaptive margin inverted score models: a) *Wild* benchmark; b) *Relaxed* benchmark; c) *Strict* Benchmark.

Table 6.4: AUC scores for the adaptive margin inverted score models, for the 3 benchmarks.

Method	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
Blur	0.957053	0.998433	0.999890
BRISQUE	0.957521	0.998643	0.999909
FaceQNet	0.956470	0.998352	0.999918
Illumination	0.957856	0.998791	0.999935
Pose	0.955912	0.998278	0.999890

Table 6.5: AUC scores for the refined training models, for the 3 benchmarks.

Method	*Wild* Benchmark	*Relaxed* Benchmark	*Strict* Benchmark
ArcFace	0.964413	0.999830	0.99999704
Blur	0.965521	0.999933	0.99999893
Illumination	0.966744	0.999950	0.99999927

QualFace: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment

João Tremoço*
*Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
joao.tremoco@isr.uc.pt

Iurii Medvedev*
*Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal
iurii.medvedev@isr.uc.pt

Nuno Gonçalves*†
†INCM Lab
Imprensa Nacional - Casa da Moeda
Lisbon, Portugal
nunogon@deec.uc.pt

***Abstract*—Modern face recognition biometrics widely rely on deep neural networks that are usually trained on large collections of wild face images of celebrities. This choice of the data is related with its public availability in a situation when existing ID document compliant face image datasets (usually stored by national institutions) are hardly accessible due to continuously increasing privacy restrictions. However this may lead to a leak in performance in systems developed specifically for ID document compliant images. In this work we proposed a novel face recognition approach for mitigating that problem. To adapt deep face recognition network for document security purposes, we propose to regularise the training process with specific sample mining strategy which penalises the samples by their estimated quality, where the quality metric is proposed by our work and is related to the specific case of face images for ID documents. We perform extensive experiments and demonstrate the efficiency of proposed approach for ID document compliant face images.**

***Index Terms*—face recognition, biometric template, document security**

I. Introduction

Security border control applications widely embed biometrics recognition where the face image is one of the most popular biometric source for such applications. The standard approach to the face recognition nowadays implies learning deep face features that are combined into a biometric template. This template is further utilised for distinguishing identities with relatively simple similarity metric and may be stored in a secured database or even embedded to the document itself for performing verification in the match-on-document scenario [1].

The features of the template may be learned explicitly by contrastive methods (i.e. by the contrast between match/non-match pairs [2]) or implicitly in the multiclass (identities) classification manner [3]. The deep networks, which are used for extracting the biometric template, usually have complex architectures of stacked convolutional layers. These networks are usually trained on big collections of labelled face images of celebrities [4], [5].

This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020

Face recognition for document security applications possesses specificities. Official identification documents (i.e. biometric passports, national ID cards) adopt only the frontal face images compliant to ICAO standards [6], [7]. In comparison with unconstrained face recognition systems, which adapts to variations in illumination, pose, occlusion, facial expressions, document security solutions deal with more regular conditions especially in a situation when biometric enrolment tends to become more controlled [8].

At the same time, the collections of ICAO compliant enrolled images, which are usually stored by national institutions, are hardly available for the research and development due to privacy issues. As an example, European GDPR (General Data Protection Regulation) categorise face images as sensitive personal data which results in many constraints for their collecting and distributing [9]. Recently, following this trend, many of the face datasets (even public wild datasets of celebrities) were withdrawn and usually available only in a form of redistribution.

That is why there is a challenge for face recognition in document security when for efficient training of the face recognition algorithms one require large ICAO compliant face image datasets which remain private, and the ones that are public available are of insufficient size. In this situation the most effective approach is to follow training on available wild datasets and then apply some optional measures (like fine-tuning) for achieving better performance in the deploy scenario [10].

In this work we address the problem of this inconsistency between the training and deploy data and introduce a novel approach to mitigate this issue. We propose to emphasise the face features which are more characteristic for ID document compliant images by designing a sophisticated sample mining strategy which regularises the training process. The developed strategy penalises the samples by their quality score (estimated by several metrics). Our approach allows to learn facial biometric template which better suits the document security applications.

II. Related Work

A. Loss function

Loss function design have been in a focus of many recent investigations of deep learning face recognition. The general trend of these works was directed onto the increasing the discriminative power of learned features. Most of the current state of the art methods follow the approach of multi-class classification with use of softmax based loss functions. To increase intra-class compactness and inter-class dispersion, several marginal modifications of softmax were proposed. For instance SphereFace, CosFace and ArcFace introduced the margin (in different manner) to the feature logits in the angular domain [3], [11], [12]. These methods demonstrated clear geometric interpretation at the same time having relatively simple implementation. Although these loss functions allowed to achieve state of the art performance in several benchmarks, they do not account the hardness and variability of each sample.

B. Hard sample mining strategies

Hard sample mining strategies allowed to improve the face recognition performance in several approaches. For instance, MV-Softmax [13] treats miss-classified samples as hard samples increasing their weights in the training process. CurricularFace [14] also uses miss-classification for indicating hard samples and adapts the curricular learning strategy to the face recognition. Hard samples are emphasised increasingly over the training duration with an additional hyper-parameter. NPCFace [15] makes the important distinction between hard positive and hard negative samples and show that for large datasets hard positives will usually be hard negatives for another class as well. The form of the negative logit is defined with use of a binary mask that indicates whether a sample is hard or not. Following the ArcFace approach, the NPCFace also utilises a margin for the positive logits, which is controlled by the hardness of the sample.

These methods try to optimise their performance towards hard samples, however we propose that for the document security applications emphasising higher quality samples during training better suits the target scenario. Unlike the previous works mentioned, MagFace [16] includes the quality of the samples in the training process in a way that pulls easy (high quality) samples closer to the class centre and pushes harder (lower quality) samples away. The authors follow a formulation similar to ArcFace where the margin parameter varies for each sample with accordance to its quality. In MagFace, the quality of each sample is defined by magnitude of the feature vector. This approach shares several conceptual similarities to our approach, however we shift our attention to adapting the quality sampling to the document security images scenario.

C. Document security specific face recognition

Document security specific face recognition investigation is reported in several works. DocFace [10] present a method for matching Identification Document (ID) photos to live photos. The authors use a pair of trained sibling networks and fine-tune them on a small private ID-Selfie dataset. The method achieves better performance over general methods, however the dataset used for benchmarking is private. Several improvements on the ID-Selfie dataset and the loss function for fine-tuning were introduced in the DocFace+ [17].

D. Face Image Quality Assessment (FIQA)

FIQA inherits aspects from general image QA also considering several other attributes such as pose, illumination, face occlusion or facial expressions. A survey on this topic was done recently by Schlett et. al [18]. Blur is good baseline indicator for the quality of any image. The blur of an image can be extracted by convolving the image with a Laplacian filter and then calculating the variance of the result [19]. BRISQUE [20] is a no-reference generic image quality assessment method. Through the use of scene statistics this method is able to quantify the "naturalness" and quality of an image. Regarding face specific attributes, several works have been recently developed to extract face specific meta-information from images. The pose of a face in an image can be characterised as a rotation in three dimensions, the yaw pitch and roll. Estimating these angles is helpful to understand a datasets pose distribution. Ruiz et. al [21] use a Convolutional Neural Network (CNN) to estimate these three angles. The quality of facial illumination is also a useful indicator of the quality of a facial image. Zhang et. al [22] use a CNN, which is trained on the FIIQD dataset to score the quality of illumination. FaceQnet [23] is a face image QA CNN based method. It used a third party framework to calculate ICAO compliance scores used as ground-truth values to train the network. The authors also show high correlation between the resulting scores and face biometric verification performance for a variety of off-the-shelf biometric recognition systems.

Some recently developed methods of face image quality assessment were developed in such a way to remove human perception from the quality estimation process. SER-FIQ [24] is a quality estimation method based on the use of dropout during the training of a model. The quality of a sample is defined with respect with the robustness of its embeddings in different sub-networks. The closer the outputs are for different sub-networks, the higher the quality of the sample is. Shi and Jain introduced the concept of Probabilistic Face Embedding (PFE) [25]. This work shows that poor image quality affects the similarity scores of genuine and impostor pairs in such a way that higher degradation of an image leads to higher probability of false reject or false accept of these pairs (named Feature Ambiguity Dilemma). As such, instead of the normal deterministic face embedding, the authors propose to encode the uncertainty in the representation of the face with two different output vectors one representing the Gaussian mean and the other for the Gaussian variance. The authors also introduce a method for matching the PFEs that penalises high levels of uncertainty (variance). SDD-FIQA [26] also bases its quality classification on the recognition performance of the sample in question. This is done by mapping the inter-class and

intra-class similarity scores to quality pseudo-labels through the use of a distribution distance metric. Afterwards, these quality values are used to train a network to predict quality scores.

III. Methodology

Deep learning classification approaches usually utilise softmax loss function, which now serves as basis for most of recently developed loss functions in the field of face recognition. It is usually formulated as follows:

$$L_{softmax} = \frac{1}{N}\sum_{i} -\log(\frac{e^{f_{y_i}}}{\sum_j^C e^{f_{y_j}}}) \tag{1}$$

where C is the number of classes in the classification problem, y_i is the index of the class of the $i-th$ sample, N is the number of samples in a batch and f_{y_j} is the $y_j - th$ component of the final layer's logits $\mathbf{f}$. If l2 normalisation of the weights $\mathbf{w_j}$ and biometric feature set $\mathbf{x_i}$ is performed, then f_{y_j} can be represented as: $f_{y_j} = w_j^T x_i = \cos(\theta_j)$. The normalised features are constrained on the hyper sphere in $\mathbb{R}^d$ space (where d is the size of $\mathbf{f}$), which leads to the angular similarity metric between samples. By reformulating softmax with this normalisation and adding an angular margin parameter m to the positive logit we obtain the ArcFace loss:

$$L_{arcface} = \frac{1}{N}\sum_{i} -\log(\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}) \tag{2}$$

A. QualFace

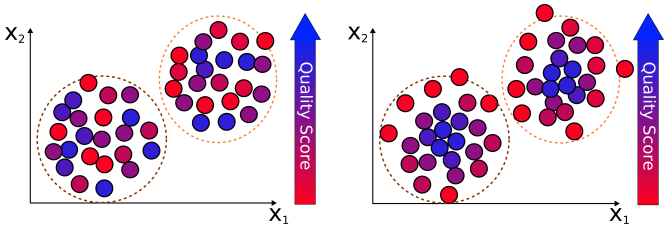
Basing on the cooperative margin presented in NPCFace [15], we introduce the concept of adaptive margin with regard to image quality. Our approach, unlike others previously mentioned, implies developing the sample mining strategy, which enhance the impact of higher quality samples instead of harder samples. In this case deep feature distribution is characterised by the concentration of the qualitative samples closer to the class feature centre (see Fig. 1). With this approach, higher impact means higher loss value for samples with better quality. This is done by increasing the margin parameter in the ArcFace loss in an adaptive way, which results in the following formulation:

$$L_q = \frac{1}{N}\sum_{i} -\log(\frac{e^{s\cos(\theta_{y_i}+m_i)}}{e^{s\cos(\theta_{y_i}+m_i)} + \sum_{j\neq y_i} e^{s\cos\theta_j}}) \tag{3}$$

where the adaptive margin parameter m_i is defined as a baseline value plus an added constant dependent on the quality of the image:

$$m_i = m_0 + \sum_{j}^{Q} w_j q_{ij} m_1 \tag{4}$$

Here, m_0 and m_1 are hyper-parameters, q_{ij} represents the normalised $j-th$ quality score value for the sample i. Q is the total number of quality attributes and w_j is the weight of each score. For travel document photos, we consider high quality samples as samples that have high ICAO standards compliance [6]. For instance, images with frontal poses, clear background, frontal face lighting, no face occlusion, no facial expressions, etc. In our work we use five different indicators of quality that are inspired by ICAO recommendations for portrait photographs: Blur [19], FaceQNet scores [23], BRISQUE scores [20], Face Illumination quality [22] and a pose score [21]. The pose scores used were calculated as the average of absolute values of the yaw, pitch and roll angles. QualFace strengthens the supervision on higher quality samples through the use of external quality indicators. The following section will show the advantages of QualFace on document security applications.



(a) (b)

Fig. 1. The spatial distribution of two high level features; a) default feature distribution; b) desired distribution in our method.

IV. Experiments and Results

We perform extensive training experiments with QualFace and several baseline loss functions and benchmark the result models in a following way.

A. Training

As a training data we used the subset of public VGGFace2_train dataset [4], selecting classes with more than 400 images per identity. The resulting dataset has a total of 1.34M images and 2842 identities. Face detection and alignment to 299×299 is performed with use of RetinaFace method [27]. Each image channel is normalised by subtracting the mean of the training dataset. The scores (FaceQNet, BRISQUE, pose score) were extracted from the aligned images. They are normalised and fed to the model as additional input.

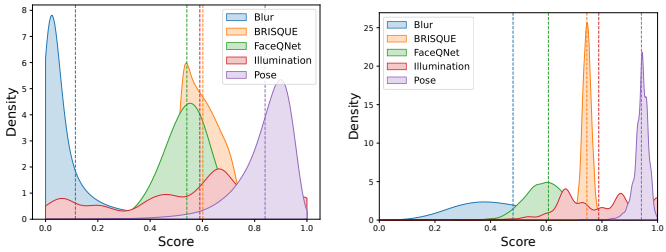
As a backbone CNN architecture we choose the ResNet50V2 [28], adding the fully connected feature layer with 512 nodes. We initialise all models with the imagenet weights before training. The training was performed on a NVIDIA RTX 3090 GPU. We limit the batch size with 24 images and decay the learning rate with cosine annealing scheduler from $5e-3$ in the beginning to $1e-5$ in the end. The model is trained with SGD optimiser for $6-th$ epochs with a momentum parameter of 0.5 and weight decay of 0.0005.

B. Benchmarking

In order to demonstrate the effect of our method, and its superiority for ID document compliant images, we designed

TABLE I
FNMR@FMR THRESHOLDS AND AUC SCORES FOR TWO BENCHMARKS.

Method		Wild			Strict			
		1e-2	1e-3	AUC	1e-3	1e-4	1e-5	AUC
Softmax		0.44502	0.79633	0.944118	0.69017	0.93655	0.98027	0,995333
ArcFace		0.28680	0.52938	0.951181	0.02486	0.10205	0.19507	0,999871
QualFace (m_0=0.4, m_1=0.1)	Blur	**0.24600**	**0.46806**	**0.959089**	**0.00793**	0.05453	0.13429	**0.999957**
	BRISQUE	0.26185	0.49934	0.954925	0.02556	0.08950	0.18444	0.999878
	FaceQNet	0.26383	0.50290	0.956515	0.01874	0.08284	0.15398	0.999910
	Illumination	0.26037	0.50076	0.956797	0.01066	**0.04835**	**0.10878**	0.999936
	Pose	0.27177	0.52186	0.954926	0.01805	0.08011	0.14550	0.999917
QualFace (m_0=0.4, m_1=0.2)	Blur	0.27460	0.52942	0.956314	0.04183	0.13423	0.19618	0.999813
	BRISQUE	0.28253	0.57363	0.953276	0.04329	0.21139	0.30880	0.999772
	FaceQNet	0.26524	0.54649	0.956252	0.03185	0.05963	0.19958	0.999944
	Illumination	0.29351	0.56006	0.952535	0.04046	0.14946	0.21644	0.999792
	Pose	0.28781	0.51604	0.951155	0.01146	0.11058	0.19958	0.999838



(a) (b)

Fig. 2. Normalised quality scores distributions across the datasets; a) VGGFace2_train dataset (identical to VGGFace2_test); b) FRGC_V2 test strict dataset.

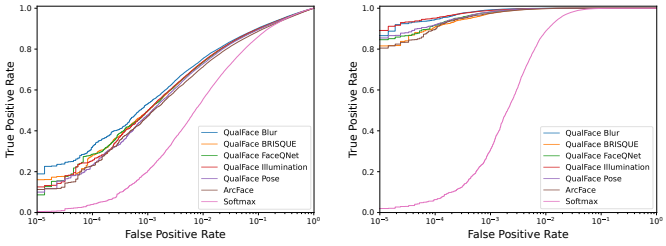
two different benchmarks datasets. The first one includes "wild" images, and the second one is comprised of images that are compliant to ICAO standards (we call it "strict"). The wild benchmark dataset was created basing on a subset of VGGFace2_test part and include identities disjoint from the training set. It contains 31k face images of 147 identities. The strict dataset was created with images from the Face Recognition Grand Challenge V2 (FRGC_V2) dataset [29]. Since its default version includes wild images, we performed its filtering in a semi-automatic way choosing only ICAO compliant images. The final strict dataset contains 11.7k images from 565 identities. For each dataset we generated the protocols for 1-1 for verification by random selecting of comparison pairs. Each protocol contains around 110K pairs for match comparison and 220K pairs for non-match comparison. [1]

To demonstrate the relative difference of distributions across two benchmark datasets we performed min-max normalisation with respect to the minimum and maximum scores values for the VGGFace2_train. One can see that the designed strict benchmark (see Fig. 2b) has better image quality with respect to the five scores presented. The wild benchmark dataset distributions, as expected, turned out to be identical to the train dataset distributions (see Fig. 2a).

[1]https://github.com/visteam-isr-uc/QualFace

C. Results Discussion

We performed intensive experiments training deep networks with QualFace and observed that the strong applied adaptation usually lead to a problem with the convergence. However, applying regular and careful adaptation, we could attain the superiority of our method. We achieved the best results in two following configurations: $m_0 = 0.4$ with $m_1 = 0.1$ and $m_1 = 0.2$. For each of those we trained five different models using a single score: Blur, BRISQUE, FaceQNet, Illumination and Pose. The Receiver Operating Characteristic (ROC) curves of the trained QualFace models (with $m_0 = 0.4$ and $m_1 = 0.1$) are represented in Fig. 3 as well as ArcFace and Softmax models for comparison.



(a) (b)

Fig. 3. ROC curves; a) Wild Benchmark; b) Strict ICAO compliance Benchmark.

From the ROC curves one can see that most of the QualFace models have better operation curves than ArcFace and Softmax. For the strict benchmark, the illumination score QualFace model exhibits the best results while for the wild benchmark the blur scores is the best performing. We estimate the performance by several metrics: False Non-Match Rate at False Match Rate (FNMR@FMR) and Area Under Curve (AUC) of ROC (see Table I).

From the results obtained, we conclude that QualFace significantly enhances the biometric verification performance in ICAO compliant face images when compared to a simple margin based loss function like ArcFace. This statement can be verified for most of the models trained in both configurations, however the models with $m_1 = 0.1$ clearly show superior results. Considering wild benchmarks, our approach performs

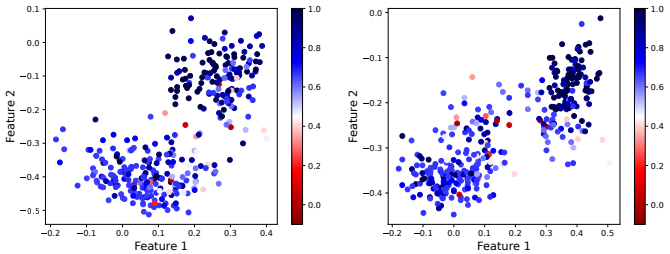
TABLE II
FNMR@FMR THRESHOLDS AND AUC SCORES FOR TWO BENCHMARKS USING ALL FIVE SCORES QUALFACE MODELS WITH m_0=0.4, m_1=0.1.

Models	Wild			Strict			
	1e-2	1e-3	AUC	1e-3	1e-4	1e-5	AUC
Equal Weights	0.26495	0.50912	0.956074	0.02869	0.12879	0.18398	0.999850
Uniformed Scores	0.26393	0.50244	0.957280	0.02171	0.08221	0.18881	0.999897
Custom Weights	**0.25735**	**0.48964**	0.956706	0.01834	**0.06875**	0.12521	0.999907
Median Lower	0.26914	0.51016	0.955681	0.02195	0.07807	0.22204	0.999891
Median	0.25829	0.49400	**0.958679**	0.02087	0.07853	0.21371	0.999905
Median Higher	0.25877	0.51080	0.957604	**0.01629**	0.07027	**0.12184**	**0.999929**

on par with the baseline models. However, most of QualFace experiment results still slightly outperform ArcFace. We conclude that our method allows to regularise the training process in deeper manner (not just adapting to qualitative samples) but generally learns better (more qualitative/discriminative) face features. From that point of view, our approach inherently shares conceptual similarities with the curriculum learning strategy.

D. Feature distribution

To better understand the QualFace impact to the learning process we analysed the real feature distribution for several particular identities in the benchmark datasets. To constrain the analysis in the 2D case we extract two principal components from the 512 dimensional embeddings with PCA (Principal Component Analysis). We represent the resulting feature distributions for two identities from the FRGC_V2 Dataset Fig. 4.



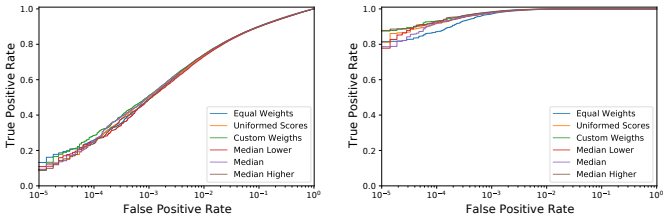
(a) (b)

Fig. 4. Features distribution of 2 different identities (*04430* and *02463*) from the FRGC_V2 dataset with Illumination scores represented in colour; a) ArcFace Model; b) Illumination Score QualFace Model with $m_0 = 0.4$ and $m_1 = 0.1$.

Basing on our results we make two observations. First, the separation between identities, which is commonly seen in margin based methods can be confirmed both in ArcFace and QualFace cases. Second, while ArcFace does not take into account image quality, QualFace pulles high quality samples towards the class centre and compacts their distribution, while the low quality samples are pushed away as theoretically hypothesised in Fig. 1b.

E. Combined scores experiments

After the experiments with sampling by a single score we intuitively investigated several scores averaging techniques. Namely, we utilised straight forward mean value, weighted mean and several median value implementations. The median implementations used three scores each. The *Median Lower* model averaged the three lower scores, the *Median* model - the three centre scores and the *Median Higher* averaged the three highest scores, for each image. We also made experiments with uniforming scores distributions in the range $[0, 1]$ before averaging for equalising their impact. The ROC curves of the combined models are represented on Fig. 5



(a) (b)

Fig. 5. Combined model ROC curves; a) Wild Benchmark; b) Strict ICAO compliance Benchmark.

In our experiments the models with *Custom Weights* and *Median Higher* weights averaging demonstrated the best performance in benchmarks. This can also be confirmed from the AUC and FNMR@FMR metrics, which are represented in table II.

We made several observations regarding the usage of combined scores. Scores uniforming indeed allowed better regularise the training and achieve better performance results. Scores weighing demonstrated its importance and the best performance was achieved when the weights were selected according to the results of single score models (Blur - 0.3, BRISQUE - 0.1, FaceQnet - 0.15, Illumination - 0.3, Pose - 0.15 in our experiments).

In the list of models with median averaging *Median Higher* case gave the most promising result, which means that the QualFace sampling strategy should be good score biased. In other words, it is better to treat a sample by its best scores rather than consider it a bad sample even if it has some lower scores.

The use of combined scores did not demonstrate the superiority in any particular benchmark. However, it allowed to achieve more regular results across the two utilised benchmarks (strict and wild) making the face representation more universal in applications with unspecified scenario. This can be verified when comparing the *Custom Weight* and *Median*

Higher model with the single score blur and illumination models.

We conclude that sampling of face images with single generic illumination and blur quality metrics allow to learn better face representation when applying the QualFace technique. Particularly, illumination quality is better suitable in application to the document security scenario, while blur score better shifts the performance towards wild face recognition scenario.

V. Conclusions

In this work we proposed a novel approach of adapting deep learning face recognition methods for document security applications. We introduced a sophisticated sample mining strategy that regularises the training process by careful emphasising the impact of samples which are better suitable for document security. The method allows to effectively train face recognition networks on big wild datasets and at the same time reduce the effect of "wildness" of these datasets. The extensive experiments with the selected marginal loss function (ArcFace) proved the superiority of adapted models against the default ones in tests with ID compliant images. The introduced strategy can also be applied to other loss functions. Our future work will focus on the study of additional image quality metrics more specific to concrete ICAO requirements. Experiments with different loss functions and finding better normalisation for the quality scores are also part of our future work plan.

VI. Acknowledgements

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics - University of Coimbra for the support of the project Facing. This work has been supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

References

[1] I. Medvedev, N. Gonçalves, and L. Cruz, "Biometric System for Mobile Validation of ID And Travel Documents," in *2020 International Conference of the BIOSIG*, 2020, pp. 1–5.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *2019 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *2018 13th IEEE International Conference on AFGR*, 2018, pp. 67–74.

[5] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," in *2016 ECCV*, vol. 9907, 10 2016, pp. 87–102.

[6] ISO/IEC JTC1 SC17 WG3. (2018) Portrait Quality - Reference Facial Images For MRTD. https://www.icao.int/Security/FAL/TRIP/Documents/TR - Portrait Quality v1.0.pdf. Version: 1.0 Date – 2018-04, Accessed: 2021-04-04.

[7] ISO/IEC JTC 1/SC 37 Biometrics. (2019) Information technology — Extensible biometric data interchange formats — Part 5: Face image data. ISO/IEC 39794-5:2019.

[8] European Commission. (2018) European Enrolment Guide for Biometric ID Documents. CEN/TC 224.

[9] European Commission. (2016) General Data Protection Regulation. Official Journal of the European Union.

[10] Y. Shi and A. K. Jain, "Docface: Matching id document photos to selfies*," in *2018 IEEE 9th International Conference on BTAS*, 2018, pp. 1–8.

[11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," in *IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[12] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large Margin Cosine Loss for Deep Face Recognition," in *2018 IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5265–5274.

[13] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12 241–12 248, 04 2020.

[14] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: Adaptive curriculum learning loss for deep face recognition," in *2020 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2020, pp. 5900–5909.

[15] D. Zeng, H. Shi, H. Du, J. Wang, Z. Lei, and T. Mei, "NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition," *CoRR*, vol. abs/2007.10172, 2020. [Online]. Available: https://arxiv.org/abs/2007.10172

[16] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14 225–14 234.

[17] Y. Shi and A. K. Jain, "Docface+: Id document to selfie matching," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 56–67, 2019.

[18] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ArXiv*, vol. abs/2009.01103, 2020.

[19] R. Bansal, G. Raj, and T. Choudhury, "Blur image detection using laplacian operator and open-cv," in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, 2016, pp. 63–67.

[20] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[21] N. Ruiz, E. Chong, and J. Rehg, "Fine-grained head pose estimation without keypoints," in *The IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 06 2018.

[22] L. Zhang, L. Zhang, and L. Li, "Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model," *Lecture Notes in Computer Science*, vol. 10636 LNCS, pp. 583–593, 2017.

[23] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality assessment for face recognition based on deep learning," *arXiv*, 2019.

[24] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Serfiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," 2020.

[25] Y. Shi, A. K. Jain, and N. Kalka, "Probabilistic face embeddings," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6901–6910, 2019.

[26] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang, "Sdd-fiqa: Unsupervised face image quality assessment with similarity distribution distance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 7670–7679.

[27] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-stage Dense Face Localisation in the Wild," *CoRR*, vol. abs/1905.00641, 2019. [Online]. Available: http://arxiv.org/abs/1905.00641

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," *CoRR*, vol. abs/1603.05027, 2016. [Online]. Available: http://arxiv.org/abs/1603.05027

[29] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition

grand challenge," in *2005 IEEE Computer Society Conference on Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 947–954.