Received: 15 January 2022 | Revised: 25 May 2022 | Accepted: 5 August 2022

DOI: 10.1049/bme2.12095

IET Biometrics

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Towards understanding the character of quality sampling in deep learning face recognition

Iurii Medvedev[1] | **João Tremoço**[1] | **Beatriz Mano**[2] | **Luís Espírito Santo**[2] | **Nuno Gonçalves**[1,2]

[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal

[2] Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal

Correspondence

Iurii Medvedev, Institute of Systems and Robotics, Department of Electrical and Computer Engineering, University of Coimbra, Rua Silvio Lima- Polo II, 3030-290 COIMBRA - Portugal.
Email: iurii.medvedev@isr.uc.pt

Abstract

Face recognition has become one of the most important modalities of biometrics in recent years. It widely utilises deep learning computer vision tools and adopts large collections of unconstrained face images of celebrities for training. Such choice of the data is related to its public availability when existing document compliant face image collections are hardly accessible due to security and privacy issues. Such inconsistency between the training data and deploy scenario may lead to a leak in performance in biometric systems, which are developed specifically for dealing with ID document compliant images. To mitigate this problem, we propose to regularise the training of the deep face recognition network with a specific sample mining strategy, which penalises the samples by their estimated quality. In addition to several considered quality metrics in recent work, we also expand our deep learning strategy to other sophisticated quality estimation methods and perform experiments to better understand the nature of quality sampling. Namely, we seek for the penalising manner (sampling character) that better satisfies the purpose of adapting deep learning face recognition for images of ID and travel documents. Extensive experiments demonstrate the efficiency of the approach for ID document compliant face images.

1 | INTRODUCTION

With the recent development of deep learning tools, face images have become one of the most important biometric sources for various security authentication applications. The general approach of face recognition is based on transforming the acquired face image to its low-dimensional representation in some deep feature domain. This domain is learnt to be highly discriminative for identities. However, it straightforwardly inherits the properties of the training data. Then, the characteristic face representation is encoded into the biometric facial template.

The distinguishing of face images by their identities is performed by the comparison of their respective biometric templates in a number of scenarios. For instance, a template may be used for one-to-one verification, when the comparison of testing and the trusted genuine samples is performed. This situation takes place when the secured template is embedded in the ID document, following the match-on-document scenario [1, 2]. Also, it may be enroled into a secured database, or searched within the collection of enroled templates, following the 1-N identification.

The properties of the deep face features domain allow utilising computationally simple similarity metrics (e.g. Euclidean distance and scalar product) for templates comparison, which significantly simplify the process of authentication within the large database of enroled identities.

The particular strategy of learning the features of a template may be different. The most popular techniques are based on learning the contrast between match/non-match identity pairs [3] or on the multiclass (identities) classification [4–7]. The deep network, which acts as a backbone of modern face recognition systems, usually has a complex architecture of stacked convolutional layers. The training data is usually collected and labelled in an automatic or semiautomatic manner (due to large size requirement) and based on public face images of celebrities [8, 9].

The application of face recognition tools set to the document security systems possesses a number of specificities. The approved forms of identification documents (i.e. biometric passports and national ID cards) in many countries allow the enrolment of only the frontal face images, which are compliant to International Civil Aviation Organization (ICAO) standards [10, 11].

In contradistinction to unconstrained face recognition methods (e.g. embedded in surveillance systems), which are focussed to cover variations of acquisition parameters (illumination, pose, occlusion, and facial expressions), document security solutions deal with more regular conditions especially with a recent tendency to control the procedure of biometric enrolment [12]. Nowadays, this process even usually includes the face quality estimation stage, which rejects samples when their estimated quality score is below a certain threshold.

At the same time, the collections of ICAO compliant enroled images, which are usually stored by national institutions, are hardly available for research and development due to privacy issues. As an example, the European GDPR (General Data Protection Regulation) categorises face images as sensitive personal data, which results in many constraints for their collection and distribution [13]. Following this trend, many of the face datasets (even public wild datasets of celebrities) were recently withdrawn and are usually available only in the 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 requires large ICAO compliant face image datasets, which remain private, and the publicly available ones 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 [14].

In this work, we extend our recent research [15], which was presented at the BIOSIG2021 conference and continue investigation towards understanding the properties of quality sampling in deep learning face recognition. Our approach is designed to reduce the impact of image *wildness*, which is the common characteristic property of training data of face recognition, for adapting the deep network to the document security scenario.

We propose to emphasise the facial 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 learning a facial biometric template, which better suits document security applications.

Along with the above improvements, we also revisit the process of benchmarking, which is based on stressing the trained networks in different scenarios and cross-comparison of the results.

This problem becomes extremely important from the perspective of face quality estimation, which has become a separate area of face recognition with its specific metrics and benchmarks [16].

2 | RELATED WORK

2.1 | Face recognition

Most recent advancements in the field of Face Recognition are related to manipulation of the loss function. Generally, the intended outcome of this manipulation is the increase of the discriminative power of the learnt feature embeddings. With this in mind, most current approaches use a classification-based approach, transforming the problem of face recognition into a multi-class classification problem. Based on the softmax loss function, some works achieved increased inter-class dispersion and intra-class compactness by introducing margin parameters (SphereFace, CosFace, and ArcFace). However, these methods do not account for the variability, hardness, or properties of each sample, and as such, further improvements can be made.

With this in mind, some works focussed on introducing hard sample mining strategies. For example, NPCFace [17] distinguishes between hard-positive and hard-negative samples, showing that for large datasets, hard-positives for one identity usually are hard-negatives for another. The authors introduce a binary mask to identify if a sample is hard or not, which impacts the formulation of the negative logit. For the positive logit, they follow the ArcFace formulation with a margin parameter that is also influenced by the hardness of the sample.

Although hard sample mining methods improve results in wild scenarios where the system's performance on hard samples is crucial, we propose that for the document security application scenario, focussing on better quality samples is indeed better suited. Following this approach, a recent work named MagFace [18] uses the quality of samples to regularise the training process by increasing the importance of higher quality samples. A loss function formulation inspired by ArcFace is followed and the margin parameter is designed to vary with a measure of sample quality previously mentioned. The MagFace approach is conceptually similar to ours and tends to control the distribution of deep features during the training process. However, the approach is limited to utilising the magnitude of deep features as an implicit indication of sample quality.

Regarding document security-specific face recognition investigation, there exist some works namely DocFace [14]. In this work, the authors present a method for matching live portraits to Identification Document (ID) photos. This was done by using a pair of sibling networks and fine-tuning them on a private ID-Selfie dataset. DocFace achieves better performance over more general face recognition approaches; however, the dataset used in benchmarking is private, and that is why no comparisons can be made. Some improvements on the ID-Selfie dataset and the loss function were introduced in DocFace+ [19], resulting in better performance.

2.2 | Face image quality assessment

As mentioned above, introducing sample quality as additional information during the training process can lead to improved results. A survey was done on face image quality assessment (FIQA) by Schlett et al. [20]. Some quality indicators can be used related to ICAO compliance: For example, a measure of the Blur in an image, although not a sufficient indicator of face image quality, can already help classify an image as low quality if the image is too blurry. A measure of image blur can be extracted by convolving the image with a Laplacian filter and then calculating the variance of the result [21]. The BRISQUE method is an image quality assessment tool that, through the use of statistics, can quantify the 'naturalness' and quality of any image. The generic quality estimation may be done also with the use of image characteristics (like saturation).

Several recent works utilised face-specific attributes to extract quality characteristics.

For example, Zhang et al. focussed their research on face illumination [22]. The authors used a Convolutional Neural Network (CNN), which is trained on the manually labelled FIIQD dataset in order to return a measure of the quality of face illumination.

The face pose is also a promising quality metric since the face should be frontal according to ICAO requirements. Ruiz et al. [23] use a CNN to estimate the three angles that describe the face pose (yaw, pitch and roll).

The facial geometry, which is usually estimated by the detection of specific landmarks can be utilised to extract the quality [24].

FaceQnet [25] is a face image quality assessment CNN-based method. To train the CNN, the authors used a third party framework to generate ICAO compliance scores that were then used as ground-truth values. They demonstrated that there is a high correlation between the FaceQnet scores and face biometric verification performance for several off-the-shelf face recognition systems.

The above-mentioned methods are heavily influenced by the human perception of quality and are related to ICAO standards, but not directly related to a performance increase. As such, some recent methods try to detach the image quality definition from human perception.

SER-FIQ [26] is a face image quality estimation method reliant on applying dropout during the training of a network. With this method of training, the quality of a sample is defined regarding the robustness of its embeddings in different sub-networks: for a given sample, the closer the outputs are for different sub-networks, the higher the quality of the sample.

In Probabilistic Face Embedding (PFE) [27], Shi and Jain show that poor image quality impacts the similarity scores both of genuine and impostor pairs. It concluded that the degradation of an image leads to a higher probability of false reject or false acceptance of these pairs. So, unlike the common deterministic face embedding, the authors propose to encode a measure of uncertainty in the embedding with two different output vectors: one representing the Gaussian mean and the other for the Gaussian variance.

PCNet [28] introduced a scheme for learning predictive confidence to reduce the proportion of errors caused by images with bad quality. The training of PCNet is performed with the use of pairwise verification scores, which is then disentangled to single images.

Like the previously mentioned methods, SDD-FIQA [29] bases its quality assessment on the recognition performance for a given sample. 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. The quality values are then used to train a CNN to predict quality scores.

Boutros et al. [30] proposed to learn the face image quality estimation by predicting its relative classifiability. Their CR-FIQA method is trained by optimising the feature representation of a sample in angular space with respect to its class centre and the nearest negative class centre.

However, many deep learning methods are usually criticised by the weak explainability of the estimated *quality* since they usually learn it via the performance of the face recognition system. That is why, the resulting scores are hardly interpreted from the perspective of standard ICAO requirements.

Fu et al. [31] investigated a number of the above face quality metrics in a combination with general image quality measures and handcrafted quality features. In our work, we draw several similar conclusions specifically to the face image quality metrics (namely the correlation of the metrics). However, we are mainly focussed on the understanding of the impact of quality sampling on the deep network training process, rather than evaluating the improvement of face verification performance by rejecting low-quality data.

3 | METHODOLOGY

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

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

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 $f_{y_j} = w_j^T x_i = \cos\left(\theta_j\right)$. The normalised features are constrained on the hyper-sphere in $\mathbb{R}^d$ space (where d is the size of $\mathbf{f}$), which lead 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:

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

3.1 | QualFace

Basing on the cooperative margin presented in NPCFace [17], 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 enhances 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 Figure 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:

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

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

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

where m_0 and m_1 are hyper-parameters and 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. Indeed, the m_0 sets the baseline margin, when m_1 defines the variation range of the quality score. Our strategy implies a linear loss function modification; however, the variations of non-linear effect may be achieved by explicit regularisation of quality score distribution (see Section 4.7).

For travel document photos, we consider high-quality samples as samples that have high ICAO standards compliance [10], for instance, images with frontal poses, clear background, frontal face lighting, no face occlusion, no facial expressions, etc. In our work, we use a number of different indicators of quality that are inspired by ICAO recommendations for portrait photographs: Blur [21], BRISQUE scores [32], FaceQnet [25], Face Illumination quality (FIIQA) [22], a Pose score [23], SER-FIQ [26], Saturation, Eyes openness [24], CR-FIQA [30], and MagFace [18]. The pose scores are calculated as the average of absolute values of the yaw, pitch and roll angles. Saturation is computed as the distance between the image saturation value and some manually chosen optimal value. The quality score values are normalised to the [0, 1] range prior to utilising in our strategy. MagFace quality measures signify the magnitude of deep feature embedding.

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.

4 | EXPERIMENTS AND RESULTS

We have performed extensive training experiments with QualFace and the baseline loss function and have estimated the performance of the resulting models in the following way.

In our basic experiments with the QualFace approach, we simplified the original training dataset. We used the subset of public VGGFace2_train dataset [8], selecting classes with more than 400 images per identity to employ the variance of the quality metric within the class. The resulting dataset has a total of 1.34 M images and 2842 identities.

Face detection and alignment to 299 × 299 dimensions were performed with the use of RetinaFace method [33]. For training, we use the custom alignment, which is based on 5 fiducial face landmarks and rigid transform operations for face centring and vertical alignment (see Figure 2). Each image channel is normalised by subtracting the mean of the training dataset before the batch generation.

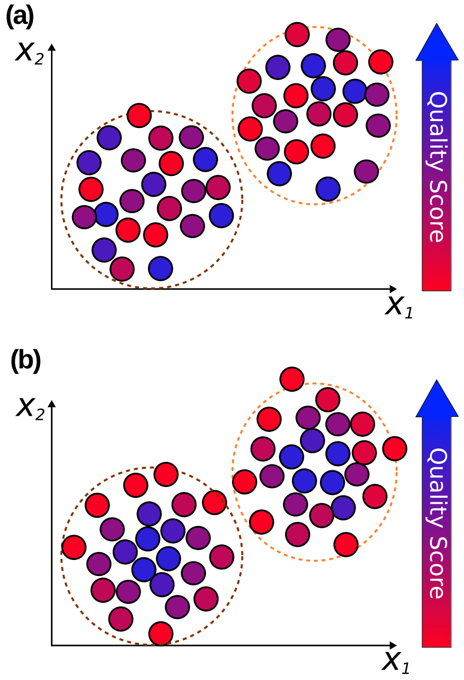

FIGURE 1 The spatial distribution of two high-level features. (a) Default feature distribution; (b) desired distribution in our method

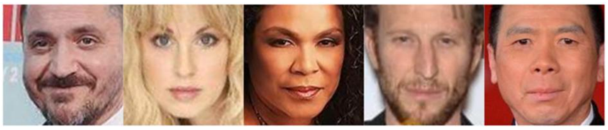

FIGURE 2 Examples of aligned training images

The same alignment is used to extract Blur, BRISQUE, Saturation, and FIIQA quality scores. FaceQnet, SER-FIQ, CR-FIQA, MagFace, and Pose metrics follow the alignment, which are defined in their public implementations. Eyes Openness score does not require a specific alignment for extraction. The selected quality scores are normalised and fed to the model as additional input for the QualFace loss function.

For the backbone CNN, we choose the ResNet50V2 [34] architecture and added a fully connected feature layer with 512 nodes. The default training settings are the following. We initialised all models with the ImageNet weights before training. The batch size used was 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 six epochs with a momentum parameter of 0.5 and weight decay of 0.0005.

4.1 | Benchmarking

The performance of face recognition in existence of quality estimation is usually estimated in a quality aware manner by rejecting images of insufficient quality from the benchmark [35]. In this case, different images are rejected for each quality metric, and the performance is indeed compared on slightly different one-to-one protocols.

We follow another approach by using several fixed one-to-one benchmark protocols for different scenarios. This is also better from the perspective of employing many different quality metrics in our work and several experiment of their joint usage. Such approach allows to better assess the effect of QualFace and its superiority for ID compliant images. The first benchmark includes "*wild*" or non-constrained images, while the second one is composed of images compliant to ICAO standards (named the "*Strict*" benchmark). The *wild* benchmark dataset was created using the test subset of VGGFace2. It contains 166k face images of 500 identities. The *Strict* dataset was created with images belonging to the Face Recognition Grand Challenge V2 (FRGC_V2) dataset [36]. Since it includes non-compliant images, we filtered the dataset in a semi-automatic way, choosing only ICAO compliant images. The final *Strict* benchmark contains 11.7 k images from 565 identities. For both benchmark datasets, the protocols for one to one for verification were generated by randomly selecting image pairs. Each protocol contains around 110 K pairs for match comparison and 220K pairs for non-match comparison (https://github.com/visteam-isr-uc/QualFace). We also adopted the set of LFW benchmarks to our work (LFW [37], CALFW [38], CPLFW [39], and XQLFW [40]), which can describe the performance under the variation of different parameters and give a better understanding of the investigated effect.

We estimate the performance by the comparison of the following metrics between the trained models for the benchmarks: False Non-Match Rate at False Match Rate (FNMR@FMR) and Area Under Curve (AUC) of ROC. With our strategy, we generally expect to achieve the boost of the performance in the *Strict* benchmark, sacrificing the performance in the *Wild* benchmark.

To understand the relative quality metric distributions across the two benchmark datasets, min-max normalisation (with respect to the minimum and maximum score values for the VGGFace2_train) was performed. It can be observed that the *Strict* benchmark (see Figure 3b) has better image quality for the used quality metrics. Also, the *Wild* benchmark dataset distributions, as expected, are identical to the train dataset distributions (see Figure 3a).

4.2 | Single-score experiments

We performed extensive experiments with QualFace and observed that the strong adaptation (high m_1 parameter values) with our method usually leads to a convergence problem. However, applying a more careful adaptation, the superiority of QualFace could be attained. The best results came from the following configurations: $m_0 = 0.4$ with $m_1 = 0.05$, $m_1 = 0.1$ and $m_1 = 0.2$. For each configuration, we trained a set of models using a single score: Blur, BRISQUE, FaceQnet, FIIQA, Pose, SER-FIQ, Saturation, Eyes openness, CR-FIQA, and MagFace. The Receiver Operating Characteristic (ROC) curves for these models (with $m_0 = 0.4$ and $m_1 = 0.1$), the baseline ArcFace (with margin $m = 0.5$, which corresponds to the mean range of m_i for this configuration), and MagFace (with default parameters) are represented in Figure 4.

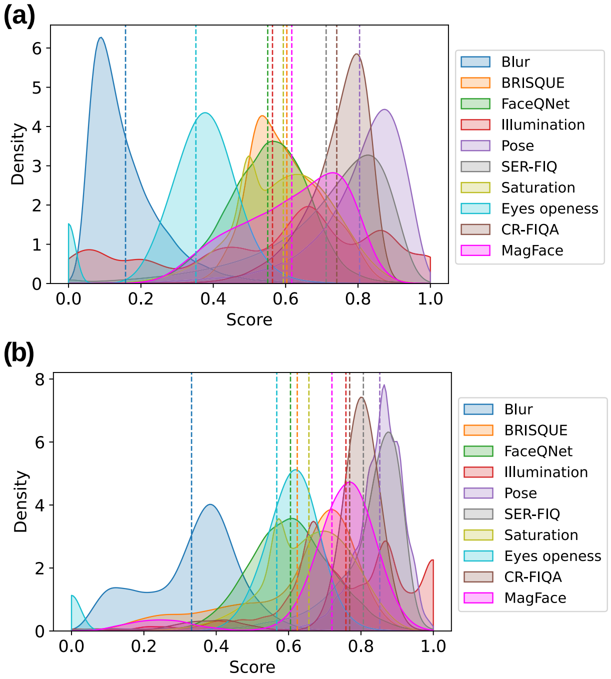

FIGURE 3 Normalised quality score distributions across the datasets. (a) VGGFace2_train dataset (identical to VGGFace2_test); (b) Face Recognition Grand Challenge (FRGC)_V2 test strict dataset

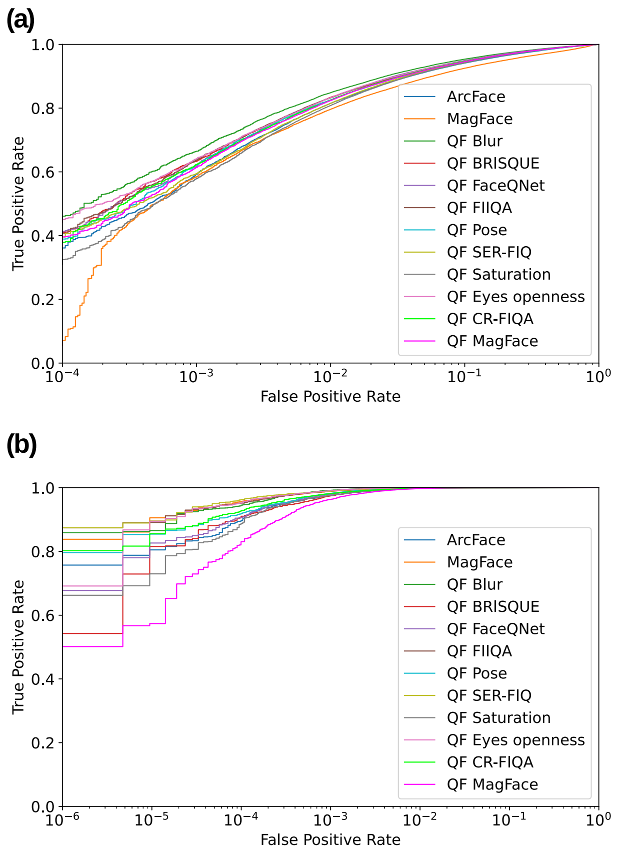

FIGURE 4 Receiver operating characteristic (ROC) curves. (a) *Wild* Benchmark; (b) *Strict* international civil aviation organisation (ICAO) compliance Benchmark

The presented results (see Tables 1 and 2, Figure 4) demonstrate that in most of the configurations, the usage of the QualFace sampling allows to outperform the baseline loss function performance in the target *Strict* scenario. Several configurations of QualFace also outperform the MagFace. Namely, QualFace can significantly facilitate the accurate face verification of ICAO compliant face images. This is verified for most of the models in all configurations; however, the configuration with $m_1 = 0.1$ demonstrates the most regular overall result. However, the usage of the strategy is score-specific. Various quality metrics require specific hyper-parameters settings.

Considering the *Wild* benchmark, our approach performs at least on par with ArcFace in most of the cases usually slightly outperforming the baseline. We conclude that our method also allows to regularise the training process in more general manner: not just adapting to qualitative samples but generally learning better (more qualitative/discriminative) facial features.

We also observe that the most regular result across configurations for considered scenarios is given by the generic metrics, like Blur, Face Illumination, Pose, when the deep metrics, which are based on the face recognition performance, does not demonstrate comparable results and tend to be sensitive to the hyper-parameters settings. For instance, it may be verified for SER-FIQ and MagFace, which give irregular results and outperform the baseline only when the weak adaptation is applied (see Table 1).

4.3 | Feature distribution

To understand how QualFace impacts the learning process, we analysed the real feature distribution for several identities in the benchmark datasets.

To constrain the analysis in the 2D case for visualisation purposes, we extract the two principal components from the 512 dimensional embeddings with the PCA (Principal Component Analysis) method.

We perform this analysis on the wild benchmark dataset (due to larger variability of the quality scores within the class) as shown in Figure 5.

The real effect on the distribution indeed is not as straightforward as desired (see Figure 1); however, observations about the QualFace effect can be made. First, our method retains the spatial separation between identities, which is a typical property of margin-based loss functions. Second, as expected, QualFace slightly pull high-quality samples towards the class centre compacting their distribution, while pushing lower quality samples away. ArcFace, which does not have image quality supervision, has a scattered quality distribution (see Figure 5).

4.4 | Combined scores experiments

After experimenting with sampling by a single score, we investigated several strategies of scores combination. First, we analysed the correlation between the used quality metrics to estimate the redundancy while using scores in a combination. The resulting correlation matrix is represented in Figure 6. The correlation value range is $[-1, 1]$, which varies from complete inverse correlation to complete correlation, while 0 means no correlation presence.

Some small level of correlation for several score pairs is expected and observed. The correlation between BRISQUE and Blur exists since BRISQUE scores also include information regarding image blur. Saturation and FIIQA are correlated since they both deal with the illumination properties. FaceQnet, SER-FIQ, CR-FIQA, and MagFace are obtained from the perspective of face recognition task, which leads to their high coupling. At the same time, those metrics are correlated with several other generic ones. For the FaceQnet, this is expected, since it utilises a face quality indicator of ICAO compliance as the ground-truth label during the training. For the SER-FIQ, this dependency is implicit since it is intrinsically focussed to capture high-level face image factors, which are relevant for the face recognition system.

To perform experiments with combined scores sampling, we have selected five quality metrics, which represent a sample from different perspectives and have a low level of correlation: Blur, BRISQUE, FaceQnet, FIIQA, and Pose.

TABLE 1 FNMR@FMR and area under curve (AUC) values for the various QualFace configurations, Baseline ArcFace, and MagFace for *Wild* and *Strict* benchmarks

Method		*Wild* benchmark			*Strict* benchmark			
		1e-2	1e-3	AUC	1e-3	1e-4	1e-5	AUC
ArcFace (baseline)		0.18483	0.40730	0.975532	0.02486	0.10205	0.19507	0.999871
Mag face		0.20376	0.41549	0.965708	0.01099	0.04362	0.09449	0.999944
QualFace ($m_0 = 0.4$, $m_1 = 0.05$)	Blur	**0.15703**	0.35077	0.979291	0.01049	0.04444	0.14623	0.999943
	BRISQUE	0.16631	0.35381	0.978996	0.02656	0.09491	0.18212	0.999876
	FaceQnet	0.15814	0.34545	**0.979759**	0.02562	0.09896	0.26051	0.999866
	FIIQA	0.15730	**0.34262**	0.979431	0.01390	0.08449	0.23264	0.999907
	Pose	0.16701	0.34667	0.978542	0.01380	0.06806	0.10515	0.999927
	SER-FIQ	0.16128	0.36441	0.978776	0.01558	0.08406	0.16226	0.999902
	Saturation	0.16698	0.35694	0.978076	0.01990	0.08222	0.19853	0.999902
	Eyes openness	0.16457	0.35045	0.979593	**0.00724**	**0.03282**	**0.07619**	**0.999959**
	CR-FIQA	0.15545	0.34798	0.980271	0.01189	0.05439	0.10434	0.999935
	MagFace	0.17850	0.39177	0.977441	0.02247	0.08262	0.15150	0.999868
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	**0.15162**	**0.33566**	**0.980600**	**0.00793**	0.05453	0.13429	**0.999957**
	BRISQUE	0.16603	0.36687	0.978071	0.02556	0.08950	0.18444	0.999878
	FaceQnet	0.16773	0.38094	0.978969	0.01874	0.08284	0.15398	0.999910
	FIIQA	0.16782	0.36317	0.978176	0.01066	**0.04835**	**0.10878**	0.999936
	Pose	0.17413	0.37836	0.976835	0.01805	0.08011	0.14550	0.999917
	SER-FIQ	0.18528	0.40587	0.976898	0.04708	0.14825	0.33822	0.999734
	Saturation	0.19192	0.42160	0.975996	0.02426	0.11208	0.27028	0.999869
	Eyes openness	0.16698	0.35918	0.978328	0.00961	0.04908	0.11502	0.999958
	CR-FIQA	0.17518	0.37720	0.977633	0.01856	0.07150	0.14453	0.999897
	MagFace	0.17621	0.38658	0.977498	0.03651	0.16957	0.42629	0.999802
QualFace ($m_0 = 0.4$, $m_1 = 0.2$)	Blur	**0.17476**	**0.39103**	**0.978118**	0.04183	0.13423	**0.19618**	0.999813
	BRISQUE	0.18796	0.42866	0.975520	0.04329	0.21139	0.30880	0.999772
	FaceQnet	0.18501	0.41202	0.976702	**0.03185**	**0.05963**	0.19958	**0.999944**
	FIIQA	0.19250	0.43931	0.975917	0.04046	0.14946	0.21644	0.999792
	Pose	0.18806	0.40735	0.975961	0.01146	0.11058	0.19958	0.999838
	SER-FIQ	0.19519	0.42910	0.975961	0.03531	0.12467	0.22893	0.999778
	Saturation	0.18629	0.40323	0.975961	0.02962	0.13005	0.27291	0.999838
	Eyes openness	0.18090	0.39833	0.977220	0.04032	0.14621	0.32314	0.999786
	CR-FIQA	0.18715	0.40388	0.975943	0.04169	0.15682	0.29013	0.999734
	MagFace	0.18708	0.41124	0.977581	0.02297	0.08210	0.20454	0.999872

Note: Bold numbers indicate the best performance per configuration.

Namely, at this stage, we seek for a better strategy of combining several uncorrelated scores to understand how to treat face images in multi-score sampling.

We performed experiments with the mean and median combined score values. The median implementations use three scores each: The *Median Lower* sampling averaged the three lower scores, the *Median* model – the three 'centre' scores and the *Median Higher* averaged the three highest scores, per sample. We also made experiments with transforming the score distributions to the uniform distribution before averaging to equalise their impact. The ROC curves of the combined models are represented in Figure 7.

The model with *Median Higher* averaging demonstrated the best performance in the benchmarks for all the combined models. This can be confirmed from the AUC and FNMR@FMR metrics, which are represented in Tables 3 and 4.

TABLE 2 FNMR@FMR and area under curve (AUC) values for the various QualFace configurations, Baseline ArcFace and MagFace for LFW, CALFW, CPLFW, and XQLFW benchmarks

		LFW			CALFW			CPLFW			XQLFW		
Method		**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**
ArcFace		0.0480	0.1180	0.9872	0.5336	0.6916	0.9000	0.7153	0.9083	0.8570	0.6356	0.8263	0.8793
MagFace		0.0493	0.1480	0.9866	0.4770	0.7486	0.9103	0.7967	0.9993	0.7783	0.9183	0.9939	0.7540
QualFace ($m_0 = 0.4$, $m_1 = 0.05$)	Blur	0.0400	0.0899	0.9895	0.4340	0.8000	0.9163	0.6300	0.8009	0.8736	0.5389	0.7340	0.9021
	BRISQUE	0.0426	0.0946	0.9887	0.5263	0.8056	0.9128	0.6420	0.8520	0.8626	0.6013	0.7850	0.8826
	FaceQnet	0.0353	0.1103	0.9891	0.4433	0.7043	0.9163	0.6547	0.8373	0.8683	0.5937	0.7903	0.8834
	FIIQA	0.0330	0.0887	0.9890	0.5033	0.7326	0.9155	0.6083	0.864	0.8714	0.5543	0.7163	0.8931
	Pose	0.0470	0.1187	0.9901	0.4786	0.6833	0.9162	0.6390	0.8156	0.8693	0.6063	0.7680	0.8919
	SER-FIQ	0.0400	0.1213	0.9889	0.5063	0.7806	0.9189	0.6553	0.8156	0.8708	0.5796	0.7283	0.8992
	Saturation	0.0380	0.1160	0.9887	0.5290	0.6970	0.9117	0.6737	0.8660	0.8704	0.5860	0.7827	0.8899
	Eyes openness	0.0453	0.1159	0.9889	0.5033	0.7053	0.9124	0.6616	0.8036	0.8604	0.5970	0.7603	0.8891
	CR-FIQA	0.0430	0.1113	0.9897	0.5013	0.8487	0.9153	0.6157	0.8210	0.8696	0.5870	0.7083	0.8934
	MagFace	0.0483	0.1103	0.9891	0.4990	0.8003	0.9127	0.6830	0.8550	0.8618	0.6360	0.8370	0.8689
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	0.0370	0.0959	0.9882	0.4573	0.7937	0.9162	0.6337	0.8250	0.8750	0.5740	0.7570	0.8977
	BRISQUE	0.0450	0.1296	0.9880	0.5223	0.7250	0.9123	0.6880	0.8326	0.8582	0.6060	0.8080	0.8903
	FaceQnet	0.0416	0.0703	0.9886	0.5296	0.7067	0.9133	0.6616	0.8753	0.8681	0.6063	0.7783	0.8816
	FIIQA	0.0416	0.0840	0.9896	0.4883	0.6983	0.9151	0.6250	0.7667	0.8633	0.5747	0.7737	0.8820
	Pose	0.0396	0.0953	0.9894	0.5403	0.7383	0.9088	0.6760	0.8727	0.8618	0.6330	0.8410	0.8837
	SER-FIQ	0.0440	0.1496	0.9881	0.5730	0.7536	0.9018	0.671	0.838	0.8598	0.5907	0.7633	0.8874
	Saturation	0.0430	0.1197	0.9897	0.5713	0.7610	0.9019	0.7030	0.8360	0.8542	0.6387	0.8667	0.8774
	Eyes openness	0.0396	0.1129	0.9894	0.5230	0.7470	0.9158	0.6550	0.8367	0.8684	0.6127	0.8230	0.8860
	CR-FIQA	0.0470	0.1077	0.9891	0.4987	0.7637	0.9128	0.6273	0.8437	0.8632	0.6217	0.8080	0.8833
	MagFace	0.0440	0.1079	0.9899	0.4923	0.7253	0.9153	0.6370	0.8029	0.8667	0.5827	0.8150	0.8896
QualFace ($m_0 = 0.4$, $m_1 = 0.2$)	Blur	0.0410	0.1083	0.9890	0.5416	0.7497	0.9089	0.7216	0.8636	0.8585	0.6177	0.7860	0.8875
	BRUSQUE	0.0450	0.1090	0.9890	0.5717	0.7717	0.9019	0.6867	0.8903	0.8592	0.6037	0.7903	0.8795
	FaceQnet	0.0373	0.1079	0.9896	0.4983	0.7396	0.9080	0.6550	0.8357	0.8566	0.6357	0.8280	0.8800
	FIIQA	0.0520	0.1943	0.9891	0.5633	0.7993	0.9036	0.6797	0.8570	0.8499	0.6353	0.7747	0.8707
	Pose	0.0503	0.1009	0.9880	0.5360	0.7220	0.9063	0.7233	0.9017	0.8595	0.6087	0.8023	0.8876
	SER-FIQ	0.0457	0.1186	0.9882	0.5356	0.8063	0.8976	0.7220	0.8700	0.8466	0.5940	0.7717	0.8818
	Saturation	0.0430	0.1426	0.9886	0.5726	0.8310	0.9046	0.7050	0.8830	0.8524	0.6230	0.7963	0.8772
	Eyes openness	0.0410	0.1220	0.9878	0.5870	0.8167	0.9028	0.6850	0.9023	0.8621	0.6603	0.8247	0.8760
	CR-FIQA	0.0426	0.1473	0.9894	0.5246	0.7840	0.9034	0.6719	0.8443	0.8560	0.6547	0.8197	0.8781
	MagFace	0.0406	0.0923	0.9893	0.5546	0.7050	0.9012	0.6793	0.8940	0.8508	0.6403	0.7700	0.8806

We have made several observations regarding the usage of combined scores. Score uniforming indeed allowed better regularising the training process and achieving better results.

From the median averaging models, the *Median Higher* case has the best results for the document compliant images, while preserving similar performance for the wild images. Namely, it means that the QualFace sampling strategy should be good score biased. In other words, it is better to use a sample's best scores rather than consider it a bad sample even if it has few low scores.

Using combined scores did not demonstrate to be a superior choice in any particular benchmark when compared to the singular score models. However, it allowed achieving a more regular result across the different scenarios, which can be useful in applications with unspecified scenarios. By comparing the *Median Higher* model with the single-score

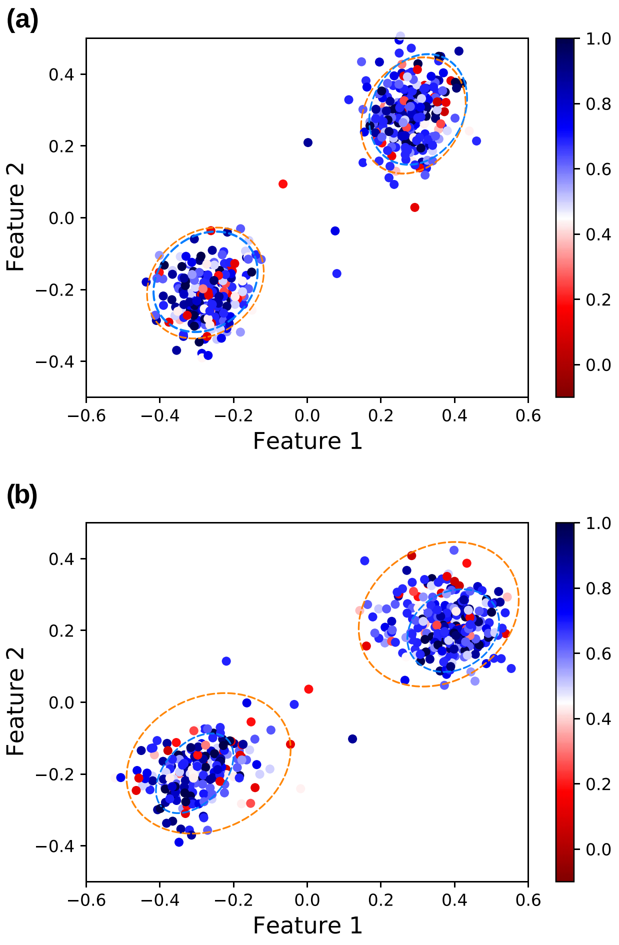

FIGURE 5 Features distribution of two different identities (*n004078* and *n002475*) from the VGGFace2 benchmark dataset with Face Illumination scores represented in colour. (a) ArcFace Model; (b) Face Illumination Score QualFace Model with $m_0 = 0.4$ and $m_1 = 0.1$. Orange-dotted circle – the approximate area of bad-quality samples (<0.2). Blue-dotted circle – the approximate area of good-quality samples (>0.8)

Blur and Face Illumination models, the previous statement can be verified.

4.5 | Inverted sampling scenario

In order to better understand the nature of our sampling strategy, we inverted previously used scores (e.g. subtracted 1 and multiplied by − 1) and afterwards QualFace single-score models with $m_0 = 0.4$, $m_1 = 0.1$ were trained using these inverted scores. The training settings were set similarly to ones in Section 4.2. This strategy is indeed more common for approaches directed onto unconstrained face recognition where the impact of hard samples is emphasised.

Intuitively, it is expected that training with the inverted scores might lead to better performance than the standard models on the *Wild* benchmark. This is due to harder samples having higher weight in the training process. However, it is

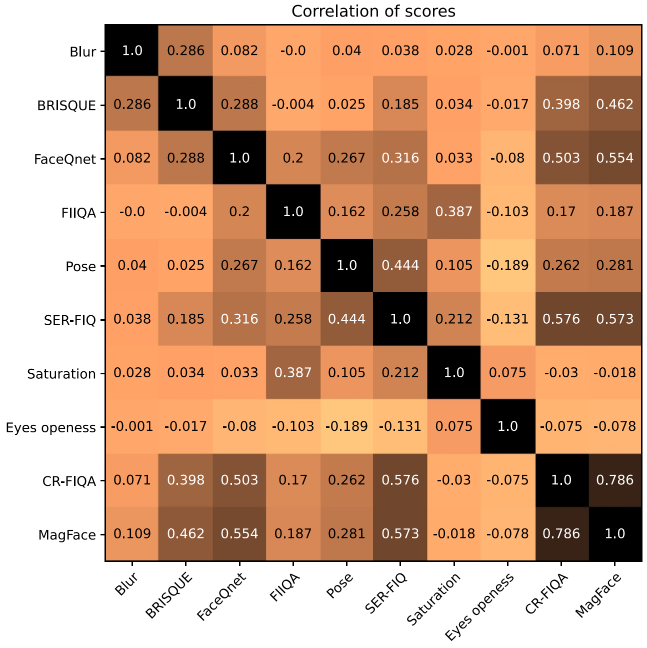

FIGURE 6 Correlation matrix for the quality scores extracted from the VGGFace2 dataset

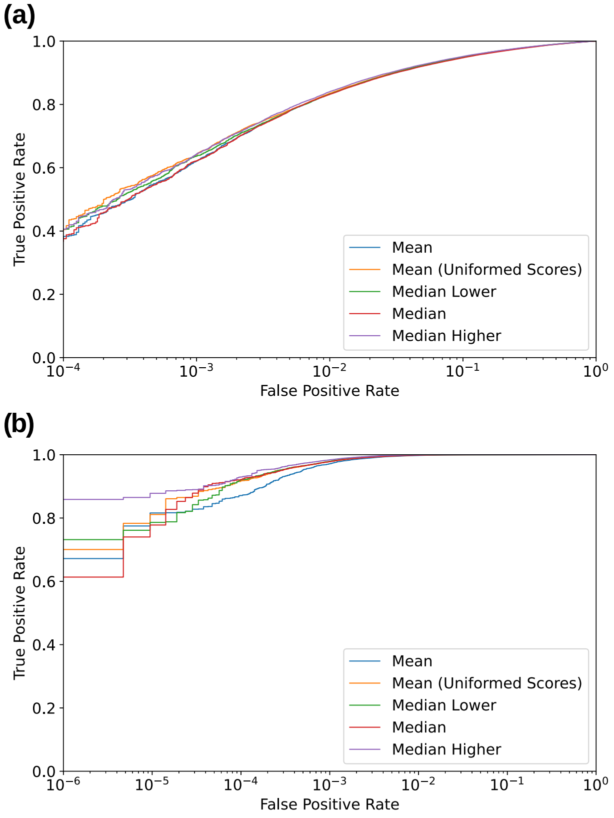

FIGURE 7 Combined model receiver operating characteristic (ROC) curves. (a) *Wild* Benchmark; (b) *Strict* international civil aviation organisation (ICAO) compliance Benchmark

TABLE 3 FNMR@FMR and area under curve (AUC) scores for two benchmarks using five scores QualFace models with $m_0 = 0.4$, $m_1 = 0.1$. *Wild* and *Strict* benchmarks

	Wild benchmark			*Strict* benchmark			
Models	**1e-2**	**1e-3**	**AUC**	**1e-3**	**1e-4**	**1e-5**	**AUC**
Mean	0.16705	0.37935	0.978644	0.02869	0.12879	0.18398	0.999850
Mean (uniformed scores)	0.16480	0.35785	0.979140	0.02171	0.08221	0.18881	0.999897
Median lower	0.16830	0.37824	0.978203	0.02195	0.07807	0.22204	0.999891
Median	0.16729	0.36361	0.979351	0.02087	0.07853	0.21371	0.999905
Median higher	**0.15986**	**0.35767**	**0.979873**	**0.01629**	**0.07027**	**0.12184**	**0.999929**

Note: Bold numbers indicate the best performance.

TABLE 4 FNMR@FMR and area under curve (AUC) scores for two benchmarks using five scores QualFace models with $m_0 = 0.4$, $m_1 = 0.1$. LWF, CALFW, CPLFW, and XQLFW benchmarks

	LFW			**CALFW**			**CPLFW**			**XQLFW**		
Model	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**
Mean	0.0413	0.1039	0.9882	0.5476	0.7703	0.9150	0.6619	0.8403	0.8665	0.5650	0.832	0.8907
Mean (U.S.)	0.0380	0.0939	0.9888	0.5363	0.8130	0.9108	0.6656	0.8160	0.8645	0.5986	0.8296	0.8890
Median lower	0.0436	0.0916	0.9896	0.5356	0.7476	0.9122	0.6779	0.8360	0.8652	0.5810	0.8053	0.8826
Median	0.0426	0.0926	0.9901	0.5176	0.7416	0.9099	0.6430	0.7667	0.8631	0.5913	0.7760	0.8875
Median higher	0.0443	0.0999	0.9886	0.4736	0.6816	0.9148	0.6240	0.7743	0.8682	0.5870	0.7633	0.8921

expected to underperform in the *Strict* scenarios since the models are not as well optimised. The results from this experiment are represented in Tables 5 and 6.

We analyse the performance difference between the normal (see Tables 1 and 2), the inverted (see Tables 5 and 6) score cases and the baseline model. Indeed, results vary across quality scores; however, several observations can be made. First, most of the inverted score models (except Blur and CR-FIQA) perform better in unconstrained (wild) conditions, which is the expected result of those experiments.

The inverted generic metrics (Blur, BRISQUE, FIIQA, Pose, Saturation, and Eyes Openness) lose their performance benefits in the *Strict* Benchmark in comparison to the normal case. Some metrics even under-perform the baseline model (especially for the lowest tested threshold of FMR $= 1e - 5$).

We also observe the improvements of the performance in the *Strict* Benchmark for the quality scores, which are trained on the face recognition performance (FaceQnet, SER-FIQ, CR-FIQA, and MagFace). Recall that these scores demonstrate high cross correlations. Due to the nature of those quality metrics, the training of a deep network by default implicitly tends to generate the desired (see Figure 1) feature distribution. That is why the additional penalising of the low-quality samples consequently also penalises the feature distribution of the overall class with the high-quality samples, improving the performance in the *Strict* scenario.

4.6 | Extended training

To understand the impact of larger scale datasets with the higher number of classes and more variable quality sampling, we have performed a set of experiments on a full VGGFace2 dataset with the QualFace configuration $m_0 = 0.4$, $m_1 = 0.1$. In these experiments, we increased the number of epochs to 20 and changed the learning rate settings, which here start at $1e - 1$ and decrease 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 SGD was set to 0.9.

We trained a set of networks with several quality metrics, which demonstrated the best performance for the selected configuration in Section 4.2: Blur, FIIQA, FaceQnet, and Eyes openness.

From the obtained results (see Tables 7 and 8), we observe that the performance of all models is clearly superior to the ones trained on the "cropped" version of the VGGFace2 dataset. The ArcFace and QualFace models outperform all previous models at all thresholds. This is an expected result with the increased breadth of the training data.

Also, by comparing the ArcFace model with the QualFace in the *Strict* benchmark, the performance gap across all thresholds is clear. In the strict scenario, the QualFace models are significantly better than baseline and this effect is more noticeable than from the results in Section 4.2. It is then possible to conclude that over larger datasets face variations and identities, the effects of sample-specific methods are enhanced. With the increase in the number of identities to be represented in the 512-dimensional feature space, the task of separating these identities in the hyper-sphere representation is more challenging. As such, the effect of the QualFace feature distribution benefits is much more noticeable. For the *Wild* benchmark, the same trend is verified. The QualFace models also increase performance in the non-restriction scenario achieving stronger performance than ArcFace, strengthening

TABLE 5 FNMR@FMR for the models with $m_0 = 0.4$, $m_1 = 0.1$, trained with inverted scores. *Wild* and *Strict* benchmarks

		Wild benchmark			*Strict* benchmark			
Method		**1e-2**	**1e-3**	**AUC**	**1e-3**	**1e-4**	**1e-5**	**AUC**
ArcFace (baseline)		0.18483	0.40730	0.975532	0.02486	0.10205	0.19507	0.999871
MagFace		0.20376	0.41549	0.965708	0.01099	0.04362	0.09449	0.99994
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	0.16666	0.36240	0.978831	0.02279	0.08321	0.20641	0.999890
	BRISQUE	0.16067	0.35909	0.979462	0.01800	0.07416	0.21906	0.999909
	FaceQnet	0.16849	0.37806	0.978761	0.01762	0.08537	0.17598	0.999918
	FIIQA	0.15888	0.35184	0.979424	0.01351	0.06568	0.14062	0.999935
	Pose	0.16789	0.36928	0.978299	0.02310	0.09207	0.19440	0.999890
	SER-FIQ	0.15981	0.35117	0.979586	0.00757	0.03103	0.07653	0.999959
	Saturation	0.16635	0.36030	0.978698	0.01517	0.05905	0.23058	0.999902
	Eyes openness	0.16365	0.35296	0.979565	0.01911	0.08163	0.19597	0.999915
	CR-FIQA	0.17097	0.38032	0.978731	0.02135	0.06527	0.11984	0.999876
	MagFace	0.16878	0.34644	0.978946	0.00975	0.04872	0.12234	0.999942

TABLE 6 FNMR@FMR for the models with $m_0 = 0.4$, $m_1 = 0.1$, trained with inverted scores. LFW, CALFW, CPLFW, and XQLFW benchmarks

		LFW			CALFW			CPLFW			XQLFW		
Method		**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**
ArcFace (Baseline)		0.0480	0.1180	0.9872	0.5336	0.6916	0.9000	0.7153	0.9083	0.8570	0.6356	0.8263	0.8793
MagFace		0.0493	0.1480	0.9866	0.4770	0.7486	0.9103	0.7967	0.9993	0.7783	0.9183	0.9939	0.7540
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	0.0393	0.1370	0.9887	0.5416	0.7613	0.9117	0.6603	0.8346	0.8625	0.5703	0.7853	0.8824
	BRISQUE	0.0390	0.1257	0.9901	0.5180	0.7707	0.9145	0.6597	0.8373	0.8695	0.5980	0.8410	0.8779
	FaceQnet	0.0380	0.0807	0.9888	0.5423	0.7480	0.9165	0.6759	0.8383	0.8622	0.6073	0.7837	0.8849
	FIIQA	0.0373	0.1653	0.9896	0.4737	0.7463	0.9133	0.6503	0.7897	0.8713	0.6017	0.7997	0.8945
	Pose	0.0400	0.1273	0.9894	0.5143	0.7880	0.9100	0.627	0.8586	0.8711	0.6096	0.7810	0.8849
	SER-FIQ	0.0373	0.0889	0.9884	0.4897	0.8167	0.9135	0.6747	0.8073	0.8605	0.5583	0.7390	0.9005
	Saturation	0.0416	0.0829	0.9887	0.5166	0.7563	0.9100	0.6877	0.8363	0.8620	0.5840	0.7863	0.8823
	Eyes openness	0.0353	0.0986	0.9887	0.4813	0.6903	0.9213	0.5936	0.8276	0.8692	0.6436	0.7853	0.8794
	CR-FIQA	0.0363	0.0643	0.9905	0.4970	0.7620	0.9147	0.6619	0.8540	0.8662	0.6010	0.8610	0.8883
	MagFace	0.0380	0.1193	0.9893	0.4983	0.7480	0.9154	0.6697	0.8380	0.8696	0.5647	0.8283	0.8890

the claim that the QualFace method produces more discriminative features.

4.7 | Sampling character

Our methodology introduces sample-specific penalisation in a linear manner when the various nonlinear cases also pose research interest. Instead of introducing the non-linearity into the methodology, we investigate its effect by modifying the distribution of the quality score itself, which gives a better visual representation of the non-linear impact. With this interpretation, we indeed just want to find the form of the score distribution, which is better for practical usage.

Scores are first transformed to default Gaussian distribution with the quantile transform $\{q_i'\} = Q(\{q_i\})$. To introduce the non-linearity, we apply a sigmoid function to the modified scores, which is multiplied by the control coefficient α:

$$q_i'' = 1/\left(1 + exp\left(-\alpha q_i'\right)\right) \tag{5}$$

With such techniques by varying α, we obtain several characteristic distribution patterns (see Figure 8), in case of $\alpha = 0$, which equalise each sample score to 0.5 that leads to the result margin $(m_0 + m_1/2)$ across all the samples (the generic ArcFace case).

Such a strategy indeed destroys the probability properties of originally extracted sampling data. However, we neglect this

TABLE 7 FNMR@FMR for ArcFace and the adaptive margin models with $m_0 = 0.4$, $m_1 = 0.1$ for the longer and refined training conditions

		***Wild* benchmark**				***Strict* benchmark**				
Method		**1e-2**	**1e-3**	**1e-4**	**AUC**	**1e-3**	**1e-4**	**1e-5**	**1e-6**	**AUC**
ArcFace		0.09741	0.20330	0.35757	0.984907	0.00065	0.00350	0.01232	0.05647	0.99999704
MagFace		0.11562	0.23680	0.38473	0.983958	0.00047	0.00440	0.01427	0.01657	0.99999722
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	0.09025	0.18353	0.33886	0.985778	0.00017	0.00172	0.01117	0.02192	0.99999893
	FIIQA	**0.08482**	**0.16534**	**0.28218**	**0.986231**	**0.00010**	**0.00170**	**0.00526**	**0.00658**	**0.99999927**
	FaceQnet	0.09404	0.19503	0.32994	0.985260	0.00028	0.00216	0.00575	0.04570	0.99999852
	Eyes openness	0.09617	0.19376	0.30802	0.985011	0.00055	0.00423	0.01000	0.01756	0.99999752

Note: Bold numbers indicate the best performance.

TABLE 8 FNMR@FMR for ArcFace and the adaptive margin models with $m_0 = 0.4$, $m_1 = 0.1$ for the longer and refined training conditions. LFW, CALFW, CPLFW, and XQLFW benchmarks

		LFW			**CALFW**			**CPLFW**			**XQLFW**		
Method		**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**
ArcFace		0.027	0.042	0.9901	0.2756	0.4686	0.9513	0.4569	0.6406	0.9162	0.4443	0.6446	0.9231
MagFace		0.0290	0.0513	0.9901	0.2813	0.5433	0.9455	0.4916	0.7190	0.9003	0.6623	0.8413	0.8804
QualFace ($m_0 = 0.4$, $m_1 = 0.1$)	Blur	0.0246	0.0296	0.9900	0.2353	0.4536	0.9529	0.4356	0.7106	0.9181	0.4513	0.6713	0.9228
	FIIQA	0.0240	0.0353	0.9894	0.236	0.3996	0.95353	0.4190	0.5987	0.9209	0.4360	0.5642	0.9281
	FaceQnet	0.0256	0.0440	0.9892	0.2763	0.4033	0.9493	0.4296	0.6933	0.9128	0.4609	0.6270	0.9206
	Eyes op	0.0253	0.0293	0.9898	0.2493	0.5060	0.94990	0.4356	0.6560	0.9145	0.4780	0.6977	0.9225

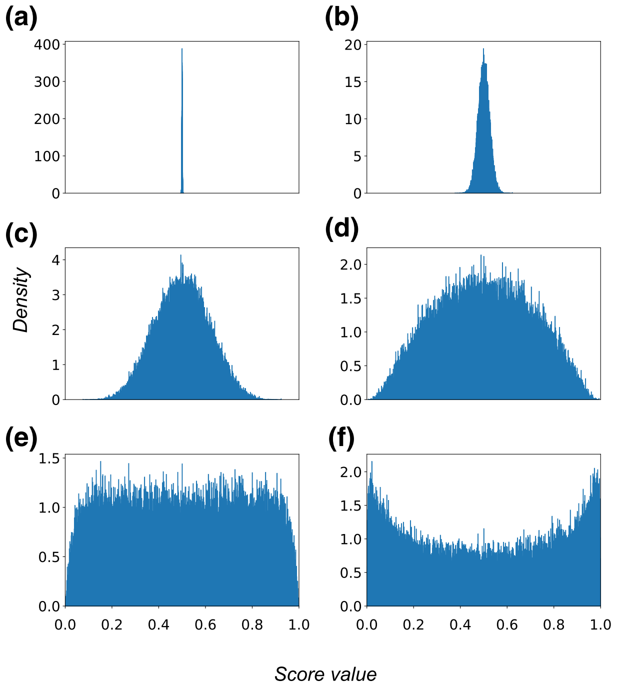

FIGURE 8 Transformed Blur score distributions. (a) $\alpha = 0$, (b) $\alpha = 1$, (c) $\alpha = 5$, (d) $\alpha = 10$, (e) $\alpha = 16$ (similar to uniform distribution), and (f) $\alpha = 22$

issue since our target is only to map the quality scores to the required range according to the order, which is specified by the original sampling.

We perform our experiments for the Blur score sampling for a set of values $\alpha = \{0, 1, 5, 10, 16, 22\}$(Tables 9 and 10). The Blur quality metrics is chosen since it is the most generic of image characteristics from our list. Also, it is smoothly distributed within its range of values and conveniently transformed in our technique (Equation (5)). The training is performed with similar settings as in Section 4.6; however, the number of epochs was reduced from 20 to 10. Our best results are achieved for the distribution with $\alpha = 5$ for both of the benchmarks. The standard deviation of that distribution corresponds to 0.11.

We conclude that for achieving the better performance and robustness in our strategy, sampling should be performed carefully in a Gaussian-like manner and constrained within a low range of score values.

5 | CONCLUSIONS

In this work, we propose a novel approach of adapting deep learning face recognition methods for document security applications. We introduce a sophisticated sample mining strategy that regularises the training process by careful emphasising the

TABLE 9 FNMR@FMR for the models with $m_0 = 0.4$, $m_1 = 0.1$ with different types of Blur sampling score distributions. *Wild* and *Strict* benchmarks

	***Wild* benchmark**				***Strict* benchmark**				
Method	**1e-2**	**1e-3**	**1e-4**	**AUC**	**1e-3**	**1e-4**	**1e-5**	**1e-6**	**AUC**
ArcFace (~ QualFace blur $\alpha = 0$)	0.09576	0.19103	0.31089	0.985109	0.000330	0.00249	0.00840	0.01574	0.9999984
QualFace blur $\alpha = 1$	0.08820	0.17005	0.30286	0.986254	0.000019	0.00045	0.00138	0.00882	0.99999978
QualFace blur $\alpha = 5$	**0.08536**	**0.16836**	**0.28994**	**0.986257**	0.000019	**0.00022**	**0.00061**	**0.00212**	**0.99999989**
QualFace blur $\alpha = 10$	0.08860	0.17186	0.30164	0.986059	0.000038	0.00098	0.00484	0.00658	0.99999957
QualFace blur $\alpha = 16$	0.08818	0.17696	0.30743	0.986327	0.000038	0.00042	0.00265	0.00570	0.99999973
QualFace blur $\alpha = 22$	0.09047	0.17944	0.29422	0.985855	**0.000009**	0.00041	0.00328	0.00656	0.99999974

Note: Bold numbers indicate the best performance.

TABLE 10 FNMR@FMR for the models with $m_0 = 0.4$, $m_1 = 0.1$ with different types of Blur sampling score distributions. LFW, CALFW, CPLFW, and XQLFW benchmarks

	LFW			**CALFW**			**CPLFW**			**XQLFW**		
Model	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**	**1e-2**	**1e-3**	**AUC**
ArcFace (~ QF blur $\alpha = 0$)	0.0260	0.0480	0.9896	0.2666	0.4750	0.9523	0.4340	0.6280	0.9183	0.4683	0.6413	0.9195
QualFace blur $\alpha = 1$	0.0246	0.0336	0.9896	0.2223	0.3440	0.9554	0.4086	0.6280	0.9177	0.4553	0.5883	0.9287
QualFace blur $\alpha = 5$	0.0250	0.0380	0.9891	0.2209	0.3269	0.9551	0.3990	0.6240	0.9199	0.4170	0.6026	0.9327
QualFace blur $\alpha = 10$	0.0240	0.0376	0.9891	0.2189	0.3646	0.9528	0.4016	0.5613	0.9164	0.4403	0.6576	0.9317
QualFace blur $\alpha = 16$	0.0253	0.0446	0.9895	0.2136	0.3983	0.9560	0.4060	0.6546	0.9187	0.3863	0.5720	0.9286
QualFace blur $\alpha = 22$	0.2466	0.0356	0.9894	0.2400	0.3683	0.9536	0.4496	0.6826	0.9155	0.4479	0.6796	0.9269

impact of samples that 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 baseline marginal loss function prove the superiority of adapted models against the default ones in tests with ID-compliant images and allow to understand better the impact of quality sampling. In most of our experiments, quality sampling allows to retain the performance (or sometimes improve it) in the non-target, unconstrained (wild) verification scenario. Namely, it evokes the idea that any type of sampling can benefit, acting as a stimulus of reordering samples allowing to generally attain more compact class representation in the feature domain. At the same time, the character of that sampling allows achieving better performance in the required scenario. The results of our work give some insights on finding a better sampling strategy. Our strategy indeed is not only limited to the loss function, which is used in our experiments but can be adapted to other loss metrics. However, the straightforward application will require repeating the ablation study for finding the suitable hyper-parameters in each case. That is why, this generalisation performance will be investigated in further work.

ACKNOWLEDGEMENTS

The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics-the 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.

CONFLICT OF INTEREST

Iurii Medvedev is familiar with Ana Filipa Sequeira as she was the jury of his PhD project.

PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES

None.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Iurii Medvedev https://orcid.org/0000-0003-2372-9681
João Tremoço https://orcid.org/0000-0001-5595-6657

REFERENCES

1. Medvedev, I., Gonçalves, N., Cruz, L.: Biometric system for mobile validation of ID and travel documents. In: 2020 International Conference of the BIOSIG, pp. 1–5 (2020)
2. Medvedev, I., et al.: Towards facial biometrics for ID document validation in mobile devices. Appl. Sci. 11(13), 6134 (2021). https://doi.org/10.3390/app11136134
3. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: a unified embedding for face recognition and clustering. In: 2015 IEEE Conference

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)

4. Sun, Y., et al.: Deep Learning Face Representation by Joint Identification-Verification. NIPS (2014)
5. Deng, J., et al.: Additive angular margin loss for deep face recognition. In: 2019 IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4685–4694 (2019)
6. Liu, W., et al.: Deep hypersphere embedding for face recognition. In: IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6738–6746 (2017)
7. Wang, H., et al.: CosFace: large margin cosine loss for deep face recognition. In: 2018 IEEE/CVF Conference on Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5265–5274 (2018)
8. Cao, Q., et al.: 'VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on AFGR, 2018, pp. 67–74
9. Guo, Y., et al.: A dataset and benchmark for large-scale face recognition. In: 2016 ECCV, vol. 9907, pp. 87–102 (2016)
10. ISO/IEC JTC1 SC17 WG3.: Portrait Quality - Reference Facial Images for MRTD, (2018). version: 1.0 Date – 2018-04, Accessed: 2021-04-04. https://www.icao.int/Security/FAL/TRIP/Documents/TR%20-%20Portrait%20Quality.v1.0.pdf
11. ISO/IEC JTC 1/SC 37 Biometrics.: Information Technology — Extensible Biometric Data Interchange Formats — Part 5: Face Image Data (2019). ISO/IEC 39794-5:2019
12. European Commission.: European Enrolment Guide for Biometric ID Documents (2018). CEN/TC 224
13. European Commission.: General Data Protection Regulation. Official Journal of the European Union (2016)
14. Shi, Y., Jain, A.K.: DocFace: matching ID document photos to selfies. In: 2018 IEEE 9th International Conference on BTAS, pp. 1–8 (2018)
15. Tremoço, J., Medvedev, I., Gonçalves, N.: QualFace: adapting deep learning face recognition for ID and travel documents with quality assessment. In: 2021 International Conference of the Biometrics Special Interest Group, pp. 1–6. (BIOSIG) (2021)
16. Grother, P., et al.: Ongoing Face Recognition Vendor Test (FRVT). Part 5: Face Image Quality Assessment. Gaithersburg (2021)
17. Zeng, D., et al.: NPCFace: A Negative-Positive Cooperation Supervision for Training Large-Scale Face Recognition. CoRR (2020). abs/2007. 10172
18. Meng, Q., et al.: MagFace: a universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14225–14234 (2021)
19. Shi, Y., Jain, A.K.: DocFace+: ID document to selfie matching. IEEE Trans. Biometrics. Behav. Identity Sci. 1(1), 56–67 (2019). https://doi.org/10.1109/tbiom.2019.2897807
20. Schlett, T., et al.: Face Image Quality Assessment: A Literature Survey. ACM Computing Surveys (CSUR) (2022)
21. Bansal, R., Raj, G., Choudhury, T.: Blur image detection using laplacian operator and open-cv. In: 2016 International Conference System Modeling Advancement in Research Trends (SMART), pp. 63–67 (2016)
22. Zhang, L., Zhang, L., Li, L.: Illumination quality assessment for face images: a benchmark and a convolutional neural networks based model. Lect. Notes Comput. Sci. 10636 LNCS, 583–593 (2017)
23. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2155–215509 (2018)
24. Soukupová, T., Cech, J.: Real-time eye blink detection using facial landmarks. In: Proceedings of the 21st Computer Vision Winter Workshop (2016)
25. Hernandez-Ortega, J., et al.: FaceQnet: Quality Assessment for Face Recognition Based on Deep Learning. arXiv (2019)
26. Terhorst, P., et al.: SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5650–5659 (2020)
27. Shi, Y., Jain, A.K., Kalka, N.: Probabilistic face embeddings. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6901–6910 (2019)
28. Xie, W., Byrne, J., Zisserman, A.: Inducing predictive uncertainty estimation for face verification'. In: 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, 7–10. BMVA Press, (2020)
29. Ou, F.Z., et al.: 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), pp. 7670–7679 (2021)
30. Boutros, F., et al.: Cr-fiqa: Face Image Quality Assessment by Learning Sample Relative Classifiability (2021)
31. Fu, B., et al.: A deep insight into measuring face image utility with general and face-specific image quality metrics. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1121–1130 (2022)
32. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Trans. Image Process. 21(12), 4695–4708 (2012). https://doi.org/10.1109/tip.2012.2214050
33. Deng, J., et al.: Retinaface: Single-Shot Multi-Level Face Localisation in the Wild, pp. 5202–5211 (2020)
34. He, K., et al.: Identity mappings in deep residual networks. In: Proceedings of ECCV 2016, pp. 630–645. Springer International Publishing (2016)
35. Maze, B., et al.: IARPA janus benchmark - C: face dataset and protocol. In: 2018 International Conference on Biometrics (ICB), pp. 158–165 (2018)
36. Phillips, P.J., et al.: Overview of the face recognition grand challenge. In: 2005 IEEE Computer Society Conference on Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 947–954 (2005)
37. Huang, G.B., LearnedMiller, E.: Labeled Faces in the Wild: Updates and New Reporting Procedures. University of Massachusetts. Amherst (2014). UM-CS-2014-003
38. Zheng, T., Deng, W., Hu, J.: Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. CoRR (2017). abs/1708. 08197
39. Zheng, T., Deng, W.: Cross-pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments, pp. 18–01. Beijing University of Posts and Telecommunications (2018)
40. Knoche, M., Hoermann, S., Rigoll, G.: Cross-quality LFW: a database for analyzing cross- resolution image face recognition in unconstrained environments. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 1–5 (2021)

How to cite this article: Medvedev, I., et al.: Towards understanding the character of quality sampling in deep learning face recognition. IET Biome. 11(5), 498–511 (2022). https://doi.org/10.1049/bme2.12095