On the Impact of Face Image Quality on Morphing Attack Detection

On the Impact of Face Image Quality on Morphing Attack Detection

Annalisa Franco[1], Matteo Ferrara[1], Chengcheng Liu[2], Christoph Busch[3], Davide Maltoni[1]
[1]Dipartimento di Informatica - Scienza e Ingegneria, University of Bologna
[2]Xi'an Jiaotong University
[3]Norwegian University of Science and Technology and Hochschule Darmstadt
annalisa.franco, matteo.ferrara, davide.maltoni@unibo.it,
liuchengcheng_xj@stu.xjtu.edu.cn,christoph.busch@ntnu.no

Abstract

The morphing attack is widely acknowledged as an important security threat to face recognition systems in the context of electronic machine readable travel documents and several possible countermeasures have been recently proposed. Among the existing solutions, differential Morphing Attack Detection (MAD) algorithms, based on the comparison of the document image (possibly morphed) and a trusted live capture, proved to be quite effective and robust in detecting this kind of attack. However, deploying such solutions in a real-world operational scenario requires the capability of dealing with images of variable quality in terms of illumination, pose, focus, etc. This paper analyzes the impact of face image quality on MAD performance through an extensive image quality assessment, carried out on a large and realistic operational dataset using different state-of-the-art algorithms, thus providing useful insights for the development of more robust MAD systems.

1. Introduction

Face Image Quality Assessment (FIQA) is a very relevant topic in the context of face recognition and a number of studies have been reported in the literature to propose new effective approaches for quality assessment or to evaluate the impact of image quality on face recognition accuracy. In the context of face morphing [6], face image quality can play an important role too. A first aspect to consider is that the morphing process, either landmark- or GAN-based, typically leaves some traces on the generated morphed image, such as artifacts in specific face regions or typical GAN generated artifacts, respectively. Such traces could be reflected in the image quality score that could be exploited to prove the quality of the morphing generation process. Furthermore, during the morphing process, the blending phase could increase slightly the blurriness and be observable as a lack of sharpness. Moreover, some works in the literature propose the possibility of exploiting quality scores for MAD [10, 9]. Another quality-related aspect to analyze is the impact that image quality might have on general MAD systems. For instance, in the Differential Morphing Attack Detection (D-MAD) task, the trusted gate image acquired by an Automated Border Control (ABC) gate is not strictly controlled in terms of pose and illumination, and such factors could have a strong impact on the D-MAD results. To the best of our knowledge, this aspect has not yet been addressed in the literature, even because this kind of analysis is not easy to carry out since public datasets generally used for the evaluation of MAD systems fail to represent realistic operational conditions, being the images mostly acquired in different contexts. In this work, a big effort has been done to collect a dataset, that aims to reproduce a very realistic testbed, thanks to the availability of images acquired at real airport border gates or at very realistic laboratory simulations of the gate environment (i.e., images acquired with real gate equipment but in a laboratory setting). The data are available in the BOEP platform[1] as sequestered test dataset. The quality of the dataset images is in this work evaluated using different FIQA Algorithms (FIQAAs), and the correlation between image quality and MAD performance is analyzed in a real operational scenario. We believe that the outcomes of this study can provide interesting insights about possible improvements of existing MAD approaches aimed at increasing the robustness to adverse image acquisition conditions. Moreover, stricter requirements to control face image quality can be justified with the findings in this work.

2. Face Image Quality Assessment

Standardized quality measures have been established in the past for some biometric traits (e.g., NFIQ 2 [16] [26] for fingerprints). As to face, the standardization process

[1]https://biolab.csr.unibo.it/fvcongoing/UI/Form/BOEP.aspx

for interoperable image quality measures is still ongoing [17], and in our analysis, we took into account the main approaches used in the literature for face image quality evaluation.

2.1. FIQA Algorithms

Most of the existing methods have been proposed for general face recognition application scenarios where the image acquisition is not strictly controlled or even uncontrolled. A recent survey [25] provides a comprehensive review of the recent literature. Some of the most recent and widely used approaches exploit deep-learning based models to assign a unified quality score to a face image. From this category, we consider in particular:

- FaceQNet [13] - A well-established approach for face image quality estimation in a variety of conditions. The proposed framework aims at attributing to ISO/ICAO compliant images top quality scores, and adopts the BioLab-ICAO framework [5] to produce the ground truth quality score used for model training.

- MagFace [21] - A face recognition approach trained by a loss function defined to learn a universal feature embedding whose magnitude can measure the quality of the given face image. The magnitude of the feature embedding monotonically increases if the subject is more likely to be recognized. In addition, MagFace introduces an adaptive mechanism to learn a well-structured within-class feature distributions by pulling easy samples to class centers while pushing hard samples away. This prevents models from overfitting on noisy low-quality samples and improves face recognition in the wild. For our experiments, the magnitude is normalized according to the minimum and maximum values obtained on the dataset. MagFace is the candidate unified quality scoring algorithm in the new international standard ISO/IEC 29794-5 and is deployed within the reference implementation Open Source Face Image Quality (OFIQ) software[2].

- CR-FIQA [2] - A recent method that estimates the face image quality of a sample by learning to predict its relative classifiability. This classifiability is measured based on the allocation of the training sample feature representation in angular space with respect to its class center and the nearest negative class center. To predict the classifiability property of a facial image, the model is trained simultaneously with a face recognition model.

- SER-FIQ [27] - The quality score is established in an unsupervised fashion, based on the relative robustness of deeply learned embeddings of the image, rather than on a predefined ground truth derived from human labeling or face comparison scores that could provide inaccurate information. By determining the embedding variations generated from random subnetworks of a face model, the robustness of a sample representation, and thus its quality, is estimated.

- Quality Regressor [8] - A regressor trained starting from a number of single quality components, specifically designed to encode in a quality score the utility of a given face sample for subsequent face recognition. Differently from the previous approaches, this method has been explicitly designed to be applied to high-quality ISO/ICAO compliant images, assigning a meaningful (and varied) quality score within the small range of variability allowed by the standard.

Besides the above mentioned FIQAAs, we also take into account some specific quality components that are particularly relevant for the analyzed scenario. Specifically, we analyze those components, which are also considered relevant in ISO/IEC 29794-5, eventhough the definition in the recently posted draft standard is close yet not identical to our work. In particular:

- *Illumination Uniformity* (IU), measuring the difference in illumination on the left and right side of the face. The IU score is determined as the intersection of the normalized luminance histograms $HL = \{hl_i, i = 1, .., n\}$ and $HR = \{hr_i, i = 1, ..n\}$ computed on the left and right side of the face region, respectively: $IU = \sum_{i=1}^{n} \min(hl_i, hr_i)$.

- *Defocus* (DF), that analyzes the level of sharpness. The DF score is computed as the difference between the face region F and the smoothed version of the same face region ($\tilde{F} = F * g$) obtained through a convolution of the image with a 3×3 mean filter g: $DF = |F - \tilde{F}|$.

- *Pose*, analyzing the head roll, pitch and yaw. In particular, given the orientation of the head relative to the optical axis; the three angles corresponding to Roll (ϕ_R), Pitch (ϕ_P) and Yaw (ϕ_Y), are estimated using the 6DRepNet [12] and each score for a given angle ϕ is computed as: $P = \max(0, 100 \cdot \cos\phi)$

- Shadows across face, computed as described in [5].

3. Dataset acquisition and composition

In the context of the iMARS European project [14], a new dataset, called Mixed-Quality (MQ) database, has been collected in six sites in different European countries, including two airports (in Lisbon and Athens) and four research

[2]https://github.com/BSI-OFIQ/OFIQ-Project

laboratories, where images were acquired under real border control conditions using real ABC capture devices.

A total of 60 different subjects have been involved in the acquisition and some of them have been acquired across multiple sites: four subjects across three sites, ten subjects across two sites and the other 46 subjects were acquired at a single site. The MQ dataset consists of:

1. **Bona fide enrolment images**, taken in a capture setup, which meets the requirements [18] for a document image in a passport application (see Figure 1(a)). For each of the 60 subjects in the database, a varying number of bona fide enrolment images were captured using a high-quality studio acquisition setup, reflecting the real-life passport photo capture process. Given the context of this work reflecting an operational border control scenario, we have taken care to ensure that all images are ISO/ICAO complaint [18]. The database comprises a total of 205 bona fide enrolment images that have been cropped to remove the background and resized in order to follow the same inter-eye distance distribution of the morphed images, so that it is not possible to infer the image class from its size or background properties.

2. **Morphed enrolment images**: morphed images created from the pool of bona fide enrolment images, as described in Section 3.1, to simulate the attack (see Figure 1(b)). In total a set of 7652 morphed enrolment images have been generated using 12 morphing algorithms (both landmark- and GAN-based) and different morphing factors.

3. **Gate images**: bona fide face images captured live with a face capture system in an ABC gate (see Figure 1(c-f)). The database contains multiple gate images captured from each subject (overall 612 images) obtained across various locations under real border control conditions using authentic ABC devices.Thus, given six different setups of ABC gates, the probe-set provides a variation for benchmarking different D-MAD algorithms. This necessitates the algorithms to exhibit robustness to varying conditions.

Some examples of bona fide, morphed and gate images from the MQ dataset are given in Figure 1).

3.1. Selecting the morph pairing candidates

An essential aspect of creating a successful morph attack is the selection of subject pairs that closely resemble each other. In line with the methodology in [7], the morphed images were created by selecting the morph pairing candidates with high comparison scores from three Commercial-Off-The-Shelf (COTS) FRSs – Neurotechnology Verilook [23], Cognitec FaceVACS [3] and Innovatrics IFace [15] SDKs. All bona fide enrolment images of each subject (i.e., the criminal) are compared with all bona fide enrolment images of other subjects of the same gender (i.e., possible accomplices) even if acquired at different sites. Images with glasses were excluded from the comparison, to prevent visible artifacts in the resulting morphed images. Given the three FRSs, a unique value $v(i, j, u, w)$ was computed:

$$v(i, j, u, w) = \frac{1}{3} \cdot \sum_{k=1}^{3} \frac{\tau_k - s_k(i,j,u,w)}{\tau_k}$$

where:

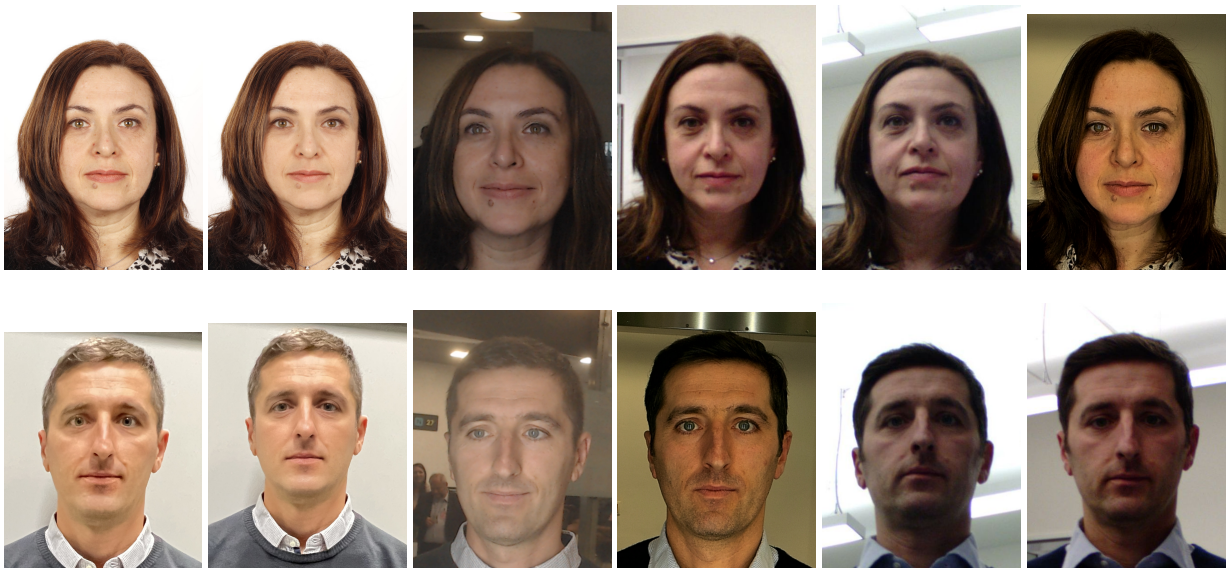
- $s_k(i, j, u, w)$ is the similarity score between bona fide image u of subject i and bona fide image w of subject j provided by the FRS k;
- τ_k is the score threshold recommended by the k^{th} FRS corresponding to a False Match Rate (FMR) of 0.1%.

The value $v(i, j, u, w)$ indicates how far the verification scores are, on average, from the FMR=0.1% thresholds; lower values imply higher similarity between images. For each criminal subject i, the potential accomplices j are sorted in increasing order of $v(i, j, u, w)$, and the top five subject candidates are selected. Given a selected pair $v(i, j, u, w)$, a morphed image is generated by combining the bona fide enrolment image u of subject i with the bona fide enrolment image w of subject j. This decision aims at maximizing the likelihood of fooling FRSs at the gate. Note that, none of the selected image pairs was able to fool all the three FRSs at the same time. Since one subject wears glasses in all the enrolment images, it was excluded from the generation process. Each of the remaining 59 subjects is paired with five other subjects, obtaining a total of 295 image pairs for morphing.

4. Experimental evaluation

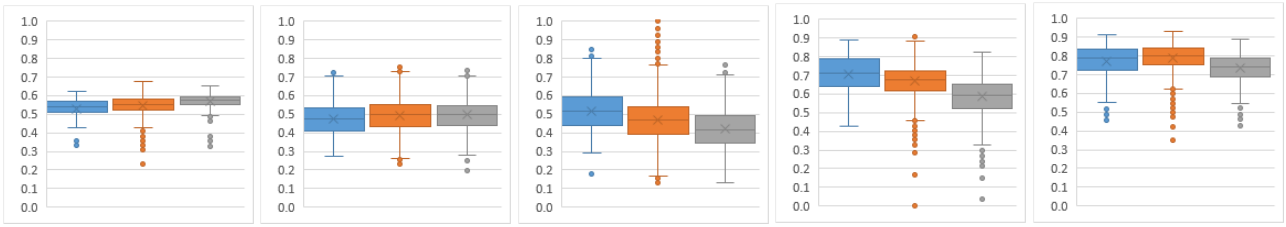
4.1. Dataset quality assessment: enrolment images

From the visual point of view, morphed images in the MQ dataset present different quality levels, since some morphing algorithms produce quite visible artifacts while other are able to generate good face morphs with limited presence of artifacts. In this section, we analyze the quality scores of the enrolment images in the MQ dataset. We report the results computed using different FIQAAs separately for the bona fide and the morphed images generated using landmark- or GAN-based morphing algorithms. In particular, Figure 2 shows the box plots of the quality scores computed over the enrolment images using the Quality Regressor, FaceQNet, OFIQ, CR-FIQA and SER-FIQ, respectively. The graphs show that some approaches tend to assign uniform quality scores to the enrolment images on average; in particular, no evident differences are visible between bona fide and morphed images for some approaches



(a) (b) (c) (d) (e) (f)

Figure 1. Example of images contained in the iMARS MQ database for two different subjects. For each row, bona fide, morphed and gate images are reported in the first (a), second (b) and last four (c-f) columns, respectively.



(a) Regressor (b) FaceQNet (c) OFIQ (d) CR-FIQA (e) SER-FIQ

Figure 2. Quality distributions of enrolment images, reported separately for Bona fide (blue), landmark-based morphs (orange) and GAN-based morphs (gray), for different FIQAAs.

(e.g., FaceQNet). For others, the quality scores assigned to bona fide images are slightly higher than those measured for the morphed images, even if the differences are limited and the overlap in terms of score distribution is significant.

Most of the FIQAAs are able to capture such differences by assigning generally lower quality scores when strong artifacts are visible. Concerning this aspect, it is worth noting that most of the tested FIQAAs first detect and crop the face so the analysis is mainly focused on the inner face region and the artifacts surrounding the face (e.g., shadows in the hair region), produced by some morphing algorithms, are not taken into account. The Quality Regressor takes into account the whole image to assign a quality score but, even in this case, most of the quality components contributing to the quality score computation focus on specific face parts and are not explicitly designed to analyze the presence of artifacts. Anyway, in all cases, the morphing algorithms that produce more artifacts received low quality scores. An interesting aspect to note is that, for some FIQAAs, a noticeable difference can be observed between landmark- and GAN-based morphing algorithms. Specifically, GAN-based algorithms achieve, in most cases, lower quality scores w.r.t. landmark-based approaches. This can be reasonable if we consider that the generated images do not present the typical artifacts of landmark-based approaches but are usually characterized by a blurred texture that is not always "natural" and an image style quite different from real images.

4.2. Dataset quality assessment: gate images

The quality of the gate images in the MQ dataset has been evaluated using FaceQNet, OFIQ, CR-FIQA and SER-FIQ; the results are reported in Figure 3.

First of all, it is worth noting that the gate images used for the D-MAD evaluation have been selected from video streams either by the specific face verification SDK installed in the acquisition device when feasible, or manually. Despite of this selection, they still present some pose variations due to the free movement of the subject during acquisition (potentially causing motion blur) or to the position

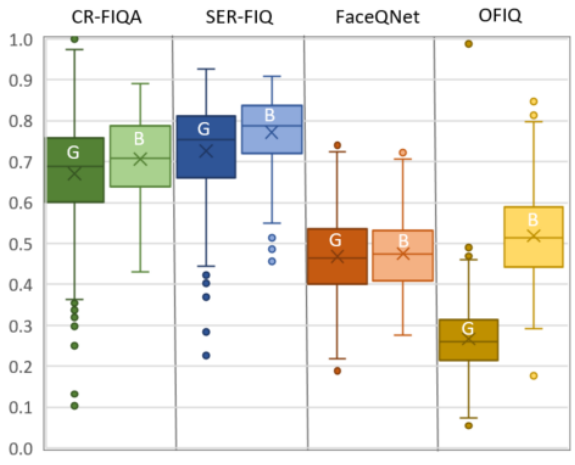


Figure 3. Comparison of quality measures for gate (G) and bona fide enrolment images (B), using different FIQAAs.

of the camera, changes in illumination or slight blurring. A first aspect to note is that, for the different FIQAAs, the bona fide images achieve quality scores mostly comparable to those of gate images (see Figure 3); the only exception is OFIQ, which generally assigns higher scores to bona fide images. We believe that this is due to several factors. First of all the methods used are deep-learning based, and the models used for quality assessment take as input fixed-size images; this implies a face detection stage (which removes the face surrounding area) and a resize that causes a loss in terms of resolution in case of high-quality ISO/ICAO compliant images. Moreover, the pre-selection done on the gate images certainly allowed to exclude the most extreme cases; the visual quality of the gate images is therefore reasonable, despite of the above mentioned variations. Finally, we observe that the range of quality scores is quite different for the analyzed FIQAAs, with average values quite high for some FIQAAs such as CR-FIQA and quite low for others (e.g., FaceQNet). This behaviour clearly shows that the comparison of quality scores computed with different approaches is critical and supports the need for a standardized approach able to guarantee system interoperability.

5. Impact of image quality on D-MAD performance

D-MAD approaches can be broadly organized in two main categories. The first one includes approaches aimed at inverting the morphing process and exploiting the resulting image for MAD detection. The first example of this category is Face Demorphing [7], where the typical processing pipeline adopted for landmark-based morphing is exploited. This method, being based on a combination of shape warping and texture blending, produces in general better results when the quality of the input images is good, i.e., well-aligned frontal images with uniform illumination. The second category includes approaches focusing on identity features, extracted from both the enrolment and gate images and compared to classify the enrolment image as morphed or not. In this category, we find two interesting recent approaches, ArcFace [4] and MagFace [21], which currently represent the state-of-the-art in D-MAD based on the results achieved in NIST FATE-MORPH[22] and BOEP [1]. The two techniques exploit the same basic idea of using a Deep Neural Network, originally trained for the face recognition task, to extract feature embeddings from the two input images. According to the authors, the network is pre-trained and no additional training is performed specifically for the morphing detection task, thus avoiding any kind of overfitting with training datasets limited in size and variety. The extracted features are then subtracted and the resulting feature vector is given as input to an SVM classifier for the final classification. For ArcFace [24, 20], a ResNet-50 [11] network is used as backbone, trained through the ArcFace loss [4]; as to MagFace, the same backbone is trained through the MagFace [21] loss function, an adaptive mechanism to learn a well-structured within-class feature distribution relying on the magnitude of vectors.

5.1. Analysis of the results

We analyze here the impact of image quality on the D-MAD performance. In particular, we conducted the D-MAD experiments on the MQ dataset described in Section 3, consisting of 2187 bona fide and 158987 morphed attempts. Figure 4 reports the results obtained by the different D-MAD algorithms on the MQ dataset in a DET plot where the Bona fide Classification Error Rate (BPCER) is given as a function of the Morphing Attack Classification Error Rate (MACER) [19].

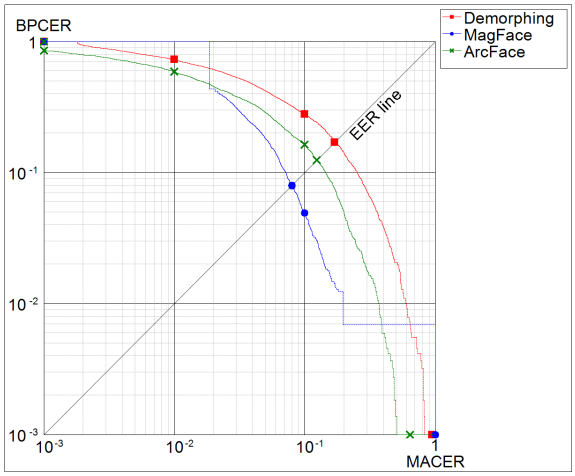


Figure 4. DET curves of the tested D-MAD approaches on the MQ dataset.

The D-EER of Demorphing is quite high, 17.01% on the

whole MQ dataset, confirming that this dataset represents a realistic and challenging benchmark compared to existing public evaluation datasets (where Demorphing achieved significantly lower error rates). Slightly better results are achieved by ArcFace, whose D-EER on the whole set is 12.4%; MagFace achieved the best performance with a D-EER of 7.73%.

5.1.1 Impact assessment based on unified quality score

For quality impact assessment, we organized the D-MAD attempts (both bona fide and morphed) in different subsets based on the quality of the enrolment and gate images involved in each attempt. Specifically, the quality scores of enrolment images are sorted in increasing order and the 1^{st}, 2^{nd} and 3^{rd} quartiles are computed and used as thresholds to organize the enrolment images into four equally-sized subsets of increasing quality ([$< 1^{st}$], [1^{st} -2^{nd}], [2^{nd} - 3^{rd}], [$>3^{rd}$]). The same is done for the gate images, thus identifying 16 possible combinations of enrolment/gate image quality ranges. Each attempt is therefore assigned to one of these 16 subsets according to the quality of the enrolment and gate image used for the attempt. We then computed the D-EER for the single subset of attempts to analyze the possible impact of image quality. The full results are given in the supplemental material. To better quantify the impact of quality, for each enrolment subset, we also computed the relative difference (percentage) in the D-EER between the gate images of bad quality ($< 1^{st}$ quartile) and those of good quality ($> 3^{rd}$ quartile):

$$\Delta_{\text{D-EER}} = \frac{\text{D-EER}(4) - \text{D-EER}(1)}{\text{D-EER}(1)} \quad (1)$$

A negative value is desirable for this indicator, since it corresponds to a reduction of the D-EER when the quality of the images is good w.r.t. bad quality images.

The results in terms of $\Delta_{\text{D-EER}}$ for Demorphing, MagFace and ArcFace are reported for different FIQAAs in Table 1. For each D-MAD approach, the table reports the $\Delta_{\text{D-EER}}$ values computed according to Eq. 1 for different FIQAAs. In the first four results columns, the Quality Regressor is used to estimate the quality of the enrolment image and one of the other four FIQAAs to compute the gate image quality. In the remaining columns, the four considered FIQAAs are used to compute the quality of both enrolment and gate images.

The results obtained show an interesting correlation between image quality and D-MAD performance of the Demorphing algorithm (see also Table 4). In general, we can confirm that, as expected, the quality of the gate image has a stronger impact on the results as compared to the enrolment image. When both enrolment and gate images have top quality scores – element (4, 4) in the matrices in the supplemental material - the D-EER measured is significantly lower: around 8-13% compared to about 17% on the whole dataset. Analogously, when both images have poor quality – element (1, 1) in the matrices in the supplemental material – significantly higher error rates are observed (around 18-20% or more). The trend is highlighted quite well by the $\Delta_{\text{D-EER}}$ values that are negative in most cases, confirming a reduction (often noticeable) of the D-EER when good quality gate images can be exploited. As to the single FIQAA, the general trend described above is more evident for some approaches. For instance, in this specific evaluation, OFIQ and SER-FIQ, as well as their combination with the Quality Regressor, seem to produce more coherent results.

This outcome is confirmed by the results obtained for some specific quality assessment algorithms also for MagFace (see Table 5 for details), even if the trend is a bit less evident in this case. In particular, the results observed with SER-FIQ and OFIQ confirm more clearly the impact of gate image quality on MAD. In general gate images of very low quality produce worst MAD performance (first column of all tables); the correlation is confirmed also in this case by the mostly negative values of $\Delta_{\text{D-EER}}$. Among the different D-MAD systems analyzed, ArcFace seems to present a lower correlation with image quality (see also Table 6), but even in this case an impact of quality can be appreciated when specific FIQAAs are used (e.g., SER-FIQ and OFIQ), as shown by the negative $\Delta_{\text{D-EER}}$ values.

The higher impact of quality on Demorphing is perfectly reasonable since the Demorphing approach is based on the inversion of the morphing process and produces better results when the two images are frontal, well-aligned and with good illumination. MagFace and ArcFace approaches are different from Demorphing and mainly rely on identity information extracted through DNN models, specifically trained to be more robust to this kind of variations.

As a further result, the impact of quality for enrolment and gate images separately has been quantified by reporting the $\Delta_{\text{D-EER}}$ in Table 2 for different FIQAAs and different D-MAD approaches. The results confirm a generally high impact of gate images quality while the effect of the enrolment image quality is less relevant. SER-FIQ and OFIQ confirm to be reliable FIQAAs, being able to predict quality scores correlated to D-MAD accuracy.

5.1.2 Impact assessment based on specific quality components

The results reported in the previous section show that the quality of gate images has an impact on MAD performance. In order to have more insights on the specific quality aspect influencing the performance, we decided to perform further investigations using some of the quality components suggested in the new version of the international standard ISO/IEC 29794-5 [17]. The analysis focuses on the gate im-

D-MAD	Enrolment quality	Regressor CR-FIQA	Regressor SER-FIQ	Regressor FaceQNet	Regressor OFIQ	CR-FIQA	SER-FIQ	FaceQNet	OFIQ
Demorph.	1	8.7%	16.5%	**-35.2%**	**-12.0%**	**-97.4%**	**-90.2%**	**-41.6%**	**-64.0%**
	2	**-30.8%**	**-43.4%**	**-40.8%**	**-43.1%**	92.3%	**-30.7%**	**-53.0%**	**-47.1%**
	3	**-9.6%**	**-51.6%**	**-20.9%**	**-13.2%**	**-1.5%**	**-31.3%**	**-55.1%**	**-41.1%**
	4	**-17.1%**	**-21.2%**	**-33.1%**	**-50.7%**	33.8%	49.0%	121.8%	**-17.7%**
MagFace	1	6.2%	**-10.3%**	**-33.3%**	**-42.8%**	**-83.1%**	**-61.9%**	**-89.2%**	**-82.5%**
	2	**-24.4%**	**-49.5%**	**-30.6%**	**-47.1%**	9.8%	**-13.4%**	**-25.7%**	**-44.7%**
	3	**-6.7%**	**-52.7%**	**-10.3%**	12.3%	**-10.8%**	**-41.9%**	**-46.9%**	**-33.8%**
	4	**-1.3%**	**-15.4%**	2.0%	**-24.6%**	71.3%	35.7%	30.0%	**-39.6%**
ArcFace	1	13.8%	**-0.9%**	**-15.1%**	**-38.3%**	**-80.3%**	**-41.5%**	**-85.5%**	**-24.4%**
	2	14.8%	**-32.7%**	28.1%	**-43.1%**	**-57.4%**	35.7%	12.6%	4.0%
	3	33.8%	4.8%	45.8%	78.3%	**-26.6%**	**-4.5%**	**-26.6%**	**-5.9%**
	4	12.2%	**-31.6%**	**-10.1%**	**-16.8%**	159.9%	33.3%	39.1%	**-24.8%**

Table 1. Demorphing, MagFace and ArcFace $\Delta_{\text{D-EER}}$ for different enrolment image quality (rows) and FIQAAs (columns).

	Measure	*Demorphing*	*MagFace*	*ArcFace*
Enrolment	Regressor	**-0.83%**	**-3.07%**	24.18%
	CR-FIQA	**-2.18%**	42.28%	43.46%
	SER-FIQ	2.93%	**-29.82%**	**-26.23%**
	FaceQNet	**-10.50%**	**-14.95%**	10.44%
	OFIQ	6.09%	**-9.74%**	56.67%
Gate	CR-FIQA	**-13.88%**	**-0.30%**	20.62%
	SER-FIQ	**-33.53%**	**-34.91%**	**-23.40%**
	FaceQNet	**-37.80%**	**-19.25%**	13.29%
	OFIQ	**-28.21%**	**-30.08%**	**-13.98%**

Table 2. Impact of enrolment and gate image quality on D-MAD performance, expressed in terms of $\Delta_{\text{D-EER}}$ representing the relative D-EER variation between the fourth and the first quality bins.

ages, which present a higher variability in terms of quality with respect to enrolment images. The quality measures are computed on the raw captured images, without any kind of cropping applied. The main focus is on illumination, possible blurring, presence of shadows and pose variations, computed as explained in Section 2.1.

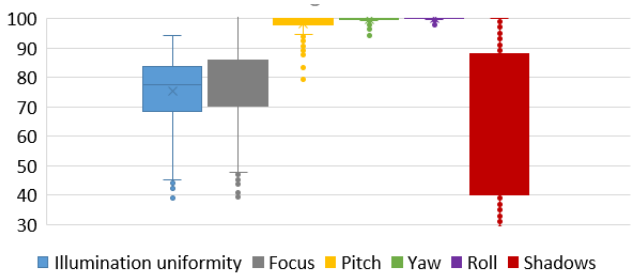
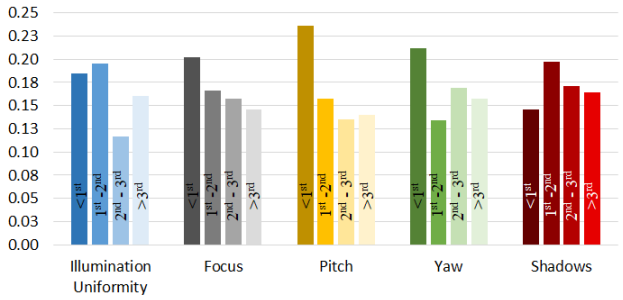


Figure 5. Quality score distribution in gate images for different quality components.

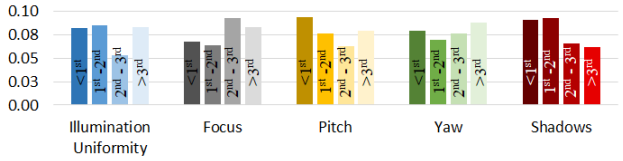
Each quality measure is defined in the range [0-100] in this case. The variations of the quality components in the set of gate images are represented in Figure 5. Gate images present some noticeable variations in terms of illumination uniformity and presence of shadows, due to the variable lightning conditions used in different acquisition sites and environments. The focus level is also quite variable, images are taken while the subject is moving so a blurring effect is visible in some cases. Finally the pose presents limited variations, even if there are some differences for the three axes. In particular, roll angles are always well controlled as expected, some limited variations are observed in terms of yaw, while pitch is the pose indicator presenting a higher variability. We believe the latter is related to the acquisition condition that characterizes some sites where the acquisition device is placed not exactly in front of the subject and images are acquired from a slightly lower (or higher) vantage point. In our evaluation, we consider only pitch that might have an impact, especially on Demorphing results.

We finally performed an analysis of the correlation between the gate image quality for the different components and the D-MAD performance. Also in this case, for each quality component, we identified the different quartiles and used them to define four subsets of gate images of varying quality. The bona fide and morphing attempts have been subsequently organized into four subsets, based on the quality of the gate images computed according to a specific quality component, and the corresponding D-EER is computed. The results are given for illumination uniformity, focus, pitch, yaw and shadows in Figures 6a, 6b and 6c for Demorphing, MagFace and ArcFace, respectively. Moreover, Table 3 reports the corresponding $\Delta_{\text{D-EER}}$.

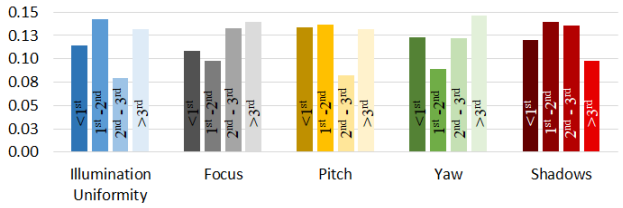
The interesting aspect to analyze in the graphs for Demorphing is the trend observed for each quality component in the different subsets of images. For illumination uniformity, focus, and pitch a very clear decreasing trend in the error rates is visible as the specific quality value increases. For shadows the trend is less evident, probably due to the equalization procedure applied by Demorphing. Pitch exhibits the most evident impact: the D-EER measured on the images with higher angles ($< 1^s t$ quartile in terms of quality)



(a) Demorphing



(b) MagFace



(c) ArcFace

Figure 6. D-EER for the four quality bins computed for different quality components on the gate images.

	Measure	*Demorphing*	*MagFace*	*ArcFace*
Quality components	Illumination	**-12.60%**	1.95%	15.57%
	Defocus	**-27.68%**	22.86%	28.89%
	Pitch	**-41.11%**	**-14.96%**	**-1.49%**
	Yaw	**-25.74%**	10.97%	18.45%
	Shadows	11.13%	**-31.62%**	**-18.72%**

Table 3. Impact of specific quality components on D-MAD performance, expressed in terms of $\Delta_{\text{D-EER}}$.

is around 24% while in the last two subsets, characterized by very limited angles, the D-EER drops to around 14%. The impact of the yaw angle, that is generally limited in the gate images, is slightly lower on D-EER; for higher yaw angles ($< 1^{st}$ quartile in terms of quality), the D-EER is in fact around 21%. Overall, the results show that, as expected, illumination conditions, focus and pose represent important challenges for Demorphing, with pose being the most important factor to consider. MagFace and ArcFace D-MAD approaches (Figures 6b and 6c) confirm a higher robustness to face image variations, generally achieving lower error rates, not strongly impacted by varying quality scores for the different quality components. Some correlations between quality and D-EER are observed for shadows and pitch. On the contrary, the trend for focus is increasing, suggesting that lower error rates are observed for slightly blurred images; this is a little counterintuitive, but it is probably justified by the fact that for MagFace the face image is resized to the fixed dimension required as input by the network for embedding extraction thus limiting the impact of defocus; moreover, a slight blurring could reduce the presence of possible artifacts in the face image, thus allowing to reliably extract identity features.

6. Conclusions

Face image quality plays an important role in face recognition; in the context of electronic ID documents, it represents a valuable instrument for the selection of document images or specific frames from the video stream to be used for face verification. This work focuses on the correlation between image quality and MAD performance. The dataset analyzed represents an important resource for this assessment since it depicts a real operational scenario, where the typical face appearance variations are well represented; the lack of realistic probe images is generally an important limitation of the existing public datasets used for D-MAD algorithms evaluation. The analysis highlights some interesting aspects, showing that D-MAD performance mostly depends, as expected, on the quality of the gate image. In particular, illumination, focus and pose have an impact on the Demorphing performance. Pose is the most critical factor since Demorphing requires an alignment of the document and the gate images and this alignment is not optimal when the pose is not frontal. MagFace and ArcFace showed to be more robust to most of these specific variations, even if in general a correlation between unified quality scores of gate images and D-MAD performance is confirmed. Identifying the most relevant quality factors influencing D-MAD performance and identifying at the same time appropriate metrics able to quantify those factors is a valuable outcome of this study. We believe that selected quality indicators could be effectively coupled with D-MAD approaches in different ways, to pre-process the gate image, if necessary (for instance to adjust lighting or other factors impacting D-MAD performance) or to select "good" frames from the video stream acquired at the gate for reliable D-MAD. Finally, quality scores could be used to modulate the morphing score, thus better taking into account the probe image quality and avoiding false alarms for low quality images. These observations provide the basis for our future research.

Acknowledgment

This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356. This text reflects only the author's views, and the commission is not liable for any use that may be made of the information contained therein.

References

[1] Biolab. BOEP. https://biolab.csr.unibo.it/fvcongoing/UI/Form/BOEP.aspx. Accessed: 2024-03-27.

[2] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. CR-FIQA: Face image quality assessment by learning sample relative classifiability. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5845, 2023.

[3] Cognitec. FaceVACS. https://www.cognitec.com/facevacs-technology.html. Accessed: 2024-03-27.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

[5] M. Ferrara, A. Franco, D. Maio, and D. Maltoni. Face image conformance to ISO/ICAO standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security*, 7(4):1204–1213, 2012.

[6] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics IJCB*, pages 1–7, 2014.

[7] M. Ferrara, A. Franco, and D. Maltoni. Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017, 2018.

[8] A. Franco, A. Magnani, D. Maltoni, D. Maio, L. Odorisio, and A. De Maria. Face image quality assessment in electronic ID documents. *IEEE Access*, 10:77744–77758, 2022.

[9] B. Fu and N. Damer. Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality. *IET Biometrics*, 11(5):359–382, 2022.

[10] B. Fu, N. Spiller, C. Chen, and N. Damer. The effect of face morphing on face image quality. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2021.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022.

[13] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[14] iMARS. iMARS european project. https://imars-project.eu/. Accessed: 2024-06-27.

[15] Innovatrics. IFace. https://www.innovatrics.com/. Accessed: 2024-03-27.

[16] ISO/IEC 29794-4:2017 Information technology - Biometric sample quality Part 4: Finger image data. Standard, International Organization for Standardization, September 2017.

[17] ISO/IEC 29794-5 — Information technology — Biometric sample quality — Part 5: Face image data. Standard, International Organization for Standardization, under development.

[18] ISO/IEC 39794-5 — Information technology — Extensible biometric data interchange formats — Part 5: Face image data. Standard, International Organization for Standardization, 2019.

[19] ISO/IEC CD 20059.2 Methodologies to evaluate the resistance of biometric recognition systems to morphing attacks. Standard, International Organization for Standardization, 2023.

[20] R. Kessler, K. Raja, J. Tapia, and C. Busch. Towards minimizing efforts for morphing attacks—deep embeddings for morphing pair selection and improved morphing attack detection. *PLOS ONE*, 19:1–29, 05 2024.

[21] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A universal representation for face recognition and quality assessment. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14220–14229, 2021.

[22] National Institute of Standards and Technology. NIST FATE Morph. https://pages.nist.gov/frvt/html/frvt_morph.html. Accessed: 2024-03-20.

[23] Neurotechnology. VeriLook. https://www.neurotechnology.com/verilook.html. Accessed: 2024-03-27.

[24] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch. Deep face representations for differential morphing attack detection. *IEEE Transactions on Information Forensics and Security*, 15:3625–3639, 2020.

[25] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. Face image quality assessment: A literature survey. *ACM Comput. Surv.*, 54(10s), sep 2022.

[26] E. Tabassi, M. Olsen, O. Bausinger, C. Busch, A. Figlarz, G. Fiumara, O. Henniger, J. Merkle, T. Ruhland, C. Schiel, and M. Schwaiger. NFIQ 2 NIST Fingerprint Image Quality. Nist technical report, National Institute of Standards and Technology, 2021.

[27] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. 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)*, pages 5650–5659, 2020.

Supplemental material

Regressor - CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.153	.143	.135	.167	8.7%
2	.164	.163	.105	.114	**-30.8%**
3	.206	.193	.155	.186	**-9.6%**
4	.155	.133	.123	.128	**-17.1%**

Regressor - SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.108	.205	.185	.126	16.5%
2	.171	.180	.136	.097	**-43.4%**
3	.261	.246	.211	.126	**-51.6%**
4	.186	.161	.102	.147	**-21.2%**

Regressor - FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.181	.156	.134	.117	**-35.2%**
2	.186	.145	.138	.110	**-40.8%**
3	.230	.174	.161	.182	**-20.9%**
4	.152	.140	.161	.101	**-33.1%**

Regressor - OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.157	.205	.103	.138	**-12.0%**
2	.176	.167	.104	.100	**-43.1%**
3	.231	.192	.140	.200	**-13.2%**
4	.164	.173	.102	.081	**-50.7%**

CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.176	.111	.036	.005	**-97.4%**
2	.140	.128	.091	.269	92.3%
3	.196	.207	.133	.193	**-1.5%**
4	.118	.188	.171	.158	33.8%

SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.207	.175	.131	.020	**-90.2%**
2	.154	.179	.161	.107	**-30.7%**
3	.209	.218	.143	.144	**-31.3%**
4	.095	.267	.213	.142	49.0%

FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.201	.165	.128	.118	**-41.6%**
2	.185	.101	.146	.087	**-53.0%**
3	.197	.176	.157	.089	**-55.1%**
4	.070	.170	.160	.155	121.8%

OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.213	.163	.078	.077	**-64.0%**
2	.193	.247	.111	.102	**-47.1%**
3	.189	.185	.127	.112	**-41.1%**
4	.181	.207	.132	.149	**-17.7%**

Table 4. Demorphing D-EER for 16 quality bins (4 for enrolment, 4 for gate images) determined according to different FIQAAs. Rows and columns refer to enrolment and gate image quality, respectively. The $\Delta_{\text{D-EER}}$ is also reported in the last column of each table, representing the relative D-EER variation between the 4^{th} and the 1^{st} gate quality bins.

Regressor - CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.087	.073	.065	.092	6.2%
2	.086	.074	.063	.065	**-24.4%**
3	.095	.054	.055	.089	**-6.7%**
4	.093	.065	.061	.092	**-1.3%**

Regressor - SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.074	.107	.082	.067	**-10.3%**
2	.096	.090	.076	.049	**-49.5%**
3	.111	.073	.063	.053	**-52.7%**
4	.086	.086	.077	.073	**-15.4%**

Regressor - FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.116	.056	.074	.077	**-33.3%**
2	.109	.055	.074	.075	**-30.6%**
3	.103	.053	.051	.092	**-10.3%**
4	.071	.077	.097	.072	2.0%

Regressor - OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.099	.097	.083	.057	**-42.8%**
2	.104	.066	.084	.055	**-47.1%**
3	.094	.046	.066	.106	12.3%
4	.079	.091	.102	.060	**-24.6%**

CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.092	.032	.018	.016	**-83.1%**
2	.105	.055	.034	.116	9.8%
3	.080	.062	.060	.071	**-10.8%**
4	.059	.112	.079	.102	71.3%

SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.127	.089	.099	.048	**-61.9%**
2	.081	.064	.066	.070	**-13.4%**
3	.089	.105	.070	.052	**-41.9%**
4	.048	.084	.078	.065	35.7%

FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.102	.083	.053	.011	**-89.2%**
2	.082	.039	.095	.061	**-25.7%**
3	.117	.049	.089	.062	**-46.9%**
4	.055	.046	.052	.072	3.0%

OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.119	.056	.101	.021	**-82.5%**
2	.067	.072	.097	.037	**-44.7%**
3	.101	.073	.083	.067	**-33.8%**
4	.109	.092	.087	.066	**-39.6%**

Table 5. MagFace D-EER for 16 quality bins (4 for enrolment, 4 for gate images) determined according to different FIQAAs. Rows and columns refer to enrolment and gate image quality, respectively. The $\Delta_{\text{D-EER}}$ is also reported in the last column of each table.

Regressor - CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.103	.114	.075	.118	13.8%
2	.099	.092	.090	.114	14.8%
3	.082	.103	.103	.110	33.8%
4	.135	.106	.098	.151	12.2%

Regressor - SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.103	.144	.113	.102	**-0.9%**
2	.122	.123	.086	.082	**-32.7%**
3	.088	.125	.122	.092	4.8%
4	.144	.169	.106	.099	**-31.6%**

Regressor - FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.133	.082	.118	.113	**-15.1%**
2	.107	.066	.114	.138	28.1%
3	.104	.083	.091	.152	45.8%
4	.137	.127	.157	.123	**-10.1%**

Regressor - OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.123	.137	.140	.076	**-38.3%**
2	.099	.118	.123	.056	**-43.1%**
3	.083	.102	.096	.148	78.3%
4	.145	.138	.126	.121	**-16.8%**

CR-FIQA

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.089	.035	.029	.018	**-80.3%**
2	.125	.103	.061	.053	**-57.4%**
3	.119	.117	.067	.088	**-26.6%**
4	.058	.145	.130	.150	159.9%

SER-FIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.150	.140	.131	.088	**-41.5%**
2	.088	.087	.123	.120	35.7%
3	.088	.181	.083	.084	**-4.5%**
4	.063	.147	.102	.084	33.3%

FaceQNet

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.124	.099	.085	.018	**-85.5%**
2	.146	.073	.163	.164	12.6%
3	.105	.065	.119	.077	**-26.6%**
4	.095	.091	.078	.132	39.1%

OFIQ

	1	2	3	4	$\Delta_{\text{D-EER}}$
1	.076	.069	.145	.057	**-24.4%**
2	.044	.072	.111	.046	4.0%
3	.138	.134	.127	.130	**-5.9%**
4	.131	.174	.117	.098	**-24.8%**

Table 6. ArcFace D-EER for 16 quality bins (4 for document, 4 for gate images) determined according to different FIQAAs. Rows and columns refer to enrolment and gate image quality, respectively. The $\Delta_{\text{D-EER}}$ is also reported in the last column of each table.