On Evaluating Pixel-Level Face Image Quality Assessment

Marco Huber[1,2], Philipp Terhöst[1,3], Florian Kirchbuchner[1], Naser Damer[1,2],
Arjan Kuijper[1,2]
[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
[3] Norwegian University of Science and Technology, Gjøvik, Norway

***Abstract*—A decisive factor for face recognition performance is face image quality (FIQ). It describes the utility of face images for automatic recognition. While this FIQ has conventionally been considered as a scalar for the whole image, emerging works suggest assessing pixel-level FIQs to provide higher explainability. However, the value of pixel-level qualities as a measure of utility (value for recognition) is not yet investigated. In this work, we address this by presenting two evaluation schemes, deletion evaluation curve (DEC) and insertion evaluation curve (IEC). The DEC investigates the change in recognition performance as pixels are deleted based on their quality. Complementary, the IEC reports the change in recognition performance as pixels are inserted based on their quality into a blurred image. Since pixel-level face image quality assessment (PLFIQA) methods assign high values to pixels that contain discriminant information, the recognition performance should decrease or increase when they are removed or added, respectively. We have successfully demonstrated the proposed evaluation scheme on two face recognition solutions by comparing a recently proposed PLFIQA method to a random baseline. With the growing interest in explainable face recognition, the proposed metrics will enable adequate comparison of future advances in pixel-level quality assessment.**

***Index Terms*—face recognition, explainablity, face image quality, evaluation metrics, biometrics**

I. Introduction

Face recognition (FR) systems based on deep learning have proven to be a powerful tool for incorporating high-performing face recognition into our daily lives [1]–[3]. Despite the significant advances in performance in recent years, today's FR systems are still challenged by unconstrained scenarios. In these scenarios, the image acquisition process is not controlled, and factors such as illuminations, pose, and occlusions cannot be controlled. These larger variabilities might result in defective matching decisions [4]. One approach to measure the impact of these variabilities and to take them into account during the matching process is the determination of the face image quality (FIQ). FIQ describes the utility of the image for the purpose of recognition [5]–[7], which does not necessary reflect the perceived image quality [8]. Whether for developers or end-users, it is important to provide more explainable decisions for the different components of the face recognition system, including FIQ assessment.

A recent work [9] was the first to propose face image quality, not on image-level, but on the pixel-level to introduce a higher degree of explainability and interpretability of the estimated face image quality. The specification of pixel qualities rather than scalar image qualities allows humans to identify areas of low and high qualities given by an FIQ assessment system. This explanation enables the user to react to low-quality areas and provides visual and interpretable instructions on how to increase the FIQ. Two example images of what those pixel-level quality (PLQ) maps look like are shown in Fig. 1. Although [9] has investigated the usefulness of the approach based on synthetic improvement and degradations, they did not evaluate the assessed PLQs as a measure of the utility of these pixels in FR. Such an evaluation scheme was not previously presented in the literature and is the main contribution of this work. In the field of evaluating classification decision, [10] and [11] proposed to use pixel insertion and deletion metrics as a causal evaluation metrics for attention map correctness. In these metrics, pixels are iteratively inserted or deleted based on the importance determined by the attention maps, and the classifier's prediction is then analyzed. Our contribution is motivated by these works and builds on them toward evaluating pixel-level biometric sample quality assessment methods.

In this work, we propose two evaluation schemes to investigate the performance of pixel-level face image quality assessment approaches. First, the deletion evaluation curve, the pixel-level face image quality assessment performance, is investigated in terms of the change in recognition performance depending on the fraction of highest quality pixels that are removed. Second, the insertion evaluation curve, the pixel-level face image quality assessment performance, is investigated by monitoring the change in recognition performance as the highest quality pixels are inserted into a highly blurred version of the image. In the experiments, we compare the approach proposed in [9] on two different FR models, ArcFace [12] and CurricularFace [13] to a random baseline approach. The experiments show the usefulness of the proposed evaluation scheme and its suitability to determine the performance of pixel-level face image quality assessment methods.

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. This work was also partially carried out during the tenure of an ERCIM 'Alain Bensoussan' Fellowship Programme.

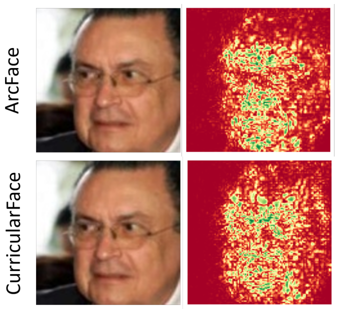


Fig. 1. **Examples of the PLQ maps as proposed by [9]**. Red indicates low pixel quality, and green indicates high pixel quality. The green and yellow pixels are mainly distributed over the face, where the highly discriminant information is located. Our proposed evaluation scheme represents the correctness of such pixel-level quality as a measure of the utility of these pixels for FR.

II. Related Work

Different standards mentioned restrictions regarding face image acquisition to achieve high quality and thus usability for face recognition [14], [15]. Since the face image acquisition process cannot be influenced in unconstraint scenarios this results in a wide range of face image variations. This has driven the need for face image quality assessment (FIQA) approaches that assign face image utility estimations, i.e. the usability of the face images for recognition [7]. Best-Rowden and Jain [5] proposed to utilize human assessments of face image quality and quality values based on similarity scores to train a quality assessment network. FaceQNet [6] in comparison, fine-tunes a face recognition network in a regression task to predict face image quality values. Meng et al. [16] proposed to encode the FIQ into the face representation by using a magnitude-aware angular margin loss function. CR-FIQA [17] presents a novel learning paradigm that estimates the FIQ by predicting the relative classifiability. SER-FIQ [18], on the other hand, exploits the robustness of FR models to variations in dropout patterns to determine the FIQ.

Explaining face image quality on a more fine-grained level to increase the explainability and interpretability was recently proposed in [9]. The proposed PLQ maps are based on back-propagated quality-dependent gradients and provide quality estimates on pixel-level. Their work investigated the suitability of the pixel-level qualities in a qualitative and quantitative analysis based on synthetic improvements of low-quality areas and degradations of high-quality areas. However, they did not analyze the recognition performance in relation to the pixel-level estimated quality, i.e. utility. This evaluation gap and the foreseen emergence in explainable and spatially-defined quality estimation is the main motivation behind this work. An evaluation scheme of pixel-level qualities in recognition performance is important and has not yet been presented.

III. Evaluation Methodology

In this section, we propose our pixel-level face image quality assessment (PLFIQA) evaluation scheme, which can be perceived as an evaluation of explainable face image quality assessment. Even though explainable machine learning has been receiving a lot of attention lately, there is so far no consensus on how such methods should be evaluated [19]. Since evaluating explainability by humans is costly and time-consuming, automated methods should not be neglected. The evaluation without humans also has the advantage that the explanation corresponds more to the view of the model than to that of the human, which reduces human bias [11].

We propose two evaluation schemes: *deletion evaluation curves (DEC)* and *insertion evaluation curves (IEC)*. The motivation behind both approaches is that if pixels of high quality, i.e. pixels with identity information, are considered, face recognition performance should increase. If this is not the case then the PLQ values do not reflect the usability for face recognition. This is partially inspired by the work of Petsiuk et al. [11] where they address the evaluation of activation maps. They propose an insertion and deletion evaluation scheme inspired by [10] to evaluate attention maps that try to explain object classification decisions.

Since the calculated PLQs are ideally not only meaningful within one face image but comparable across multiple face images, i.e. the distribution of pixel qualities can vary between images, we consider them across the entire database. To avoid higher-level complexities of the variations in pixel qualities between the reference and probe images, we restrict the modification, and thus the evaluation, to the probe image only and leave the reference image unchanged.

A. Deletion Evaluation Curve (DEC)

The intuition of the deletion evaluation curve (DEC) is that removing high-quality pixels from the original images by masking them with a constant value should quickly lead to a decrease in recognition performance as they contain relatively high-utility information. The masking in comparison to other approaches, e.g. blurring, is motivated by removing discriminant information. For the constant value, we chose zero and therefore changed each deleted pixel's color to black. The DEC is created by evaluating the recognition performance at each iteration of removing a fraction of pixels depending on their quality and plotting the performance depending on the removed fraction.

B. Insertion Evaluation Curve (IEC)

A contrary approach is taken by the insertion evaluation curve (IEC). The intuition is that adding unchanged pixels of high quality to a highly blurred version of the original image should quickly lead to an increase in face recognition performance. By adding high utility pixels, more discriminant information is added, and the model can better distinguish identities. Similar to the approach for evaluating classification decisions by [11] we start by a highly blurred image rather than a constant canvas, which in our case is motivated by the idea to reduce the impact of artifacts on the face recognition models. The IEC is created by evaluating the recognition performance at each iteration of inserted pixels based on their

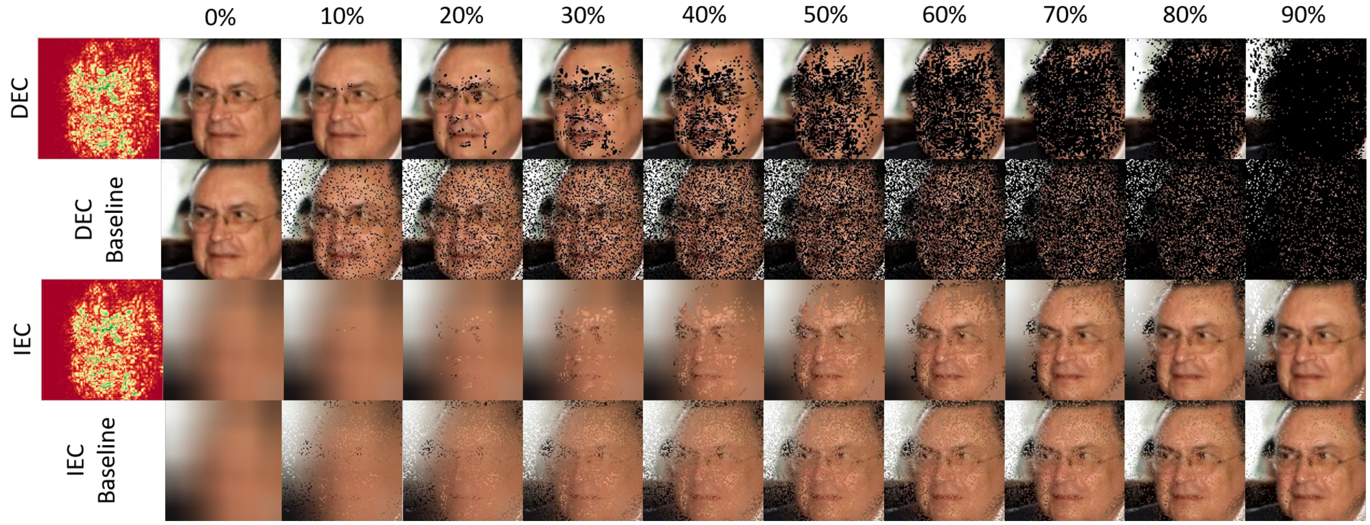


Fig. 2. **Examples of the image manipulations used in the proposed evaluation scheme**. First row (DEC): Depending on the calculated PLQ maps for the images of the entire dataset and the calculated quality values of the pixels, a fraction of pixels are masked. Second row (DEC-Baseline): The same fraction of pixels of the dataset are randomly selected and masked. Third row (IEC): Based on the pixel qualities of the dataset, a fraction of pixels are inserted into a highly blurred version of the original image. Last row (IEC-Baseline): The same fraction of pixels over the entire dataset are randomly selected and inserted into the blurred version.

quality and then plotting the performance depending on the removed fraction.

C. *Area under the Curve (AUC)*

The area under the curve (AUC) is often used to investigate receiver operating characteristic (ROC) curves further. We utilize the AUC to quantify the proposed evaluation curves above and simplify the comparisons of multiple pixel-level quality assessment approaches. Since many pixels on the face images also contain background (usually low-quality pixels), we suggest two ways of observation. The first one calculates the AUC over the whole curve, while the second one considers only the highest 20% of removed or inserted pixels. Limiting the consideration to the first 20% indicates how well the approach can identify the crucial pixels for identification. This percentage can be selected by the evaluator. It is also important to note that for the IEC, a lower AUC-IEC reflects a better-performing PLFIQA, and for DEC, a higher AUC-DEC reflects a better-performing PLFIQA.

D. *Evaluation Process*

The evaluation process proceeds as follows. First, the pixel-level qualities are determined on the entire database using a PLFIQA method. Then, depending on the chosen evaluation curve, these calculated pixel-level qualities are used to iteratively mask the original image or add pixels to the blurred version of the image. After each evaluation step, the altered images are processed by an FR system. The embedded altered images are then compared against unaltered reference images and the performance is evaluated. An example of the created images is shown in Fig 2.

IV. Experimental Setup

A. *Face Recognition Models*

We utilize two face recognition models in the experiments: ArcFace [12] and CurricularFace [13]. Both are high-performing FR solutions and are widely used in public. Both models follow the ResNet-100 architecture [20] and were trained by the corresponding authors on the MS1MV2 dataset [12]. The face images are pre-processed as proposed by the corresponding authors. Both FR models are used to extract face embeddings, feature representations of the faces, and the embeddings are then compared using cosine similarity.

B. *Benchmarks*

Two widely used benchmarks were selected for evaluation: LFW [21] and AgeDB-30 [22]. LFW [21] is an unconstrained face verification benchmark containing 13k images of over 5k identities. AgeDB [22] is an in-the-wild dataset for age-invariant face verification. We use the most reported and most challenging scenario AgeDB-30 with an age gap of 30 years between the images of the individuals. Both standard protocols of the face verification benchmarks provide 6000 pre-defined comparison pairs. We treat the first image as the probe image and the second image of each pair as the reference image. In the experiments, the probe image is always manipulated (insertion and deletion), while the reference images remain unchanged.

C. *Pixel-level Quality Maps*

To determine the face image quality at pixel-level, we follow the approach in [9], as it is the first and only (so far) PLFIQA approach. To obtain the PLQ maps, first, the model-dependent overall face image quality is estimated using the SER-FIQ [18] approach. The estimated quality scalars are then used to extend the FR model to a quality-regression model and quality-based gradients are back-propagated and post-processed to obtain the final PLQ maps. For the parameters of the PLQ-map generation process, we set the parameters for the ArcFace model, as proposed by [9], to $\alpha_{AF} = 130$, $r_{AF} = 0.88$, and $\gamma_{AF} = 7.5$. The parameters for the CurricularFace model were experimentally obtained on the Adience [23] dataset as $\alpha_{CF} = 100$, $r_{CF} = 0.88$, and $\gamma_{CF} = 3.5$.

D. Evaluation, Baseline & Evaluation Criteria

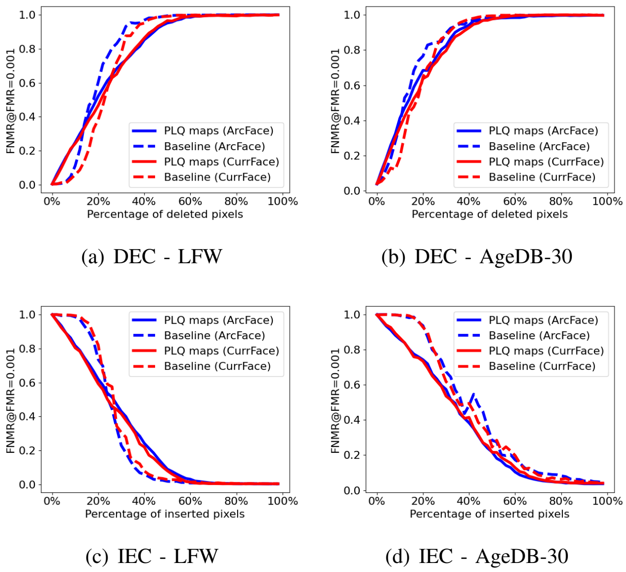
We investigate both, DEC and IEC on both datasets utilizing both FR models. We evaluate the recognition performance after iteratively inserting or deleting 2% of the pixels in the dataset per iteration. We apply a normalized box filter with kernel size [50,50] for the blurring of the images used during the IEC evaluation. For the baseline, we consider a random insertion and deletion approach. In this approach, we randomly selected 2% of the remaining pixels of the entire dataset to be inserted or deleted at each iteration.

For the evaluation, we investigate the false non-match rate (FNMR) at the false match rate (FMR) of 0.1%. This has been proposed as the best practice evaluation operation point for high-security scenarios, e.g. automatic border control systems by the European Border and Coast Guard Agency (Frontex) [24].

For the overall assessment of the suitability of the process for determining the pixel qualities, we utilize the calculation of the AUC as mentioned above. For both evaluation curves, DEC and IEC, we compute the area under the computed curves, which map the performance depending on the number of pixels removed using the composite trapezoidal rule. For the DEC, a lower recognition performance value indicates a better performing PLFIQA method, as the algorithms remove the most discriminant pixel regarding a correct decision first. For IEC, a higher recognition performance indicates a better PLFIQA method, as the most discriminant pixels that lead to a correct decision are inserted first. The same applies to AUC-DEC and AUC-IEC.

V. Results

In this section, we present the results of our proposed evaluation methodology applied to the PLQ maps of [9]. We compare them to a baseline based on random pixel selection on two different FR models, ArcFace [12] and CurricularFace [13] on two different datasets. The results are shown in Fig. 3. For the DEC, it can be seen that the masking of pixels depending on their calculated quality leads to an earlier decrease in performance in comparison to the baseline, and as more pixels are removed, the recognition performance keeps decreasing. Furthermore, we investigate that at some points, the noise introduced by randomly masking pixels in the baseline approach leads to a higher decrease than the masking based on the calculated pixel-level qualities. This might be caused by the higher distribution of the masked pixels by the random approach in comparison to the PLQ maps as it can also be observed in Fig 2. In the evaluation of the PLQ maps, images of an overall lower quality stay unaltered as the pixel qualities of overall high-quality images are higher. In the random approach, in comparison, no distinction is made and noise is randomly introduced into all images.



(a) DEC - LFW

(b) DEC - AgeDB-30

(c) IEC - LFW

(d) IEC - AgeDB-30

Fig. 3. **DEC and IEC:** The curves show that the PLQ maps assign pixels that are important for face recognition higher quality values than pixels that are less important. This is demonstrated by the rapidly rising DEC and dropping IEC as more pixels are deleted and inserted, respectively.

For the IEC evaluation that is shown in Fig. 3 (c) and (d), the drop regarding the FNMR shows that the recognition performance increase as more pixels are added to the blurred version of the image. The insertion of pixels based on their pixel quality as calculated using the PLQ maps leads to an earlier increase in recognition performance than random selection of the pixels inserted. At some point on the LFW dataset, similar to the observations on the DECs, the random baseline approach outperforms the PLQ maps in some cases. This might also be caused by the fact that low-quality face images remain longer blurred as the high-quality pixels are distributed over high-quality faces while the random approach more equally distributes the inserted pixels over the whole dataset.

In Table I the AUC-DEC and AUC-IEC are shown. The AUCs allow a more quantified comparison of the performance of the different approaches. We investigate the overall AUC (100%) and the AUC for the first 20% as motivated in the previous section. For the AUC-DEC a higher value indicates better performance. Therefore, the random baseline outperforms the PLQ maps approach in some cases if we consider the whole evaluation curve. If we limit the evaluation to the first 20% of the pixels, the PLQ approach shows clearer superiority than the baseline. Using the AUC-IEC, a lower value indicates a better-performing PLFIQA approach. In all but one case, the PLQ maps perform better than the baseline approach, especially with the limitation to 20%.

In summary, our proposed evaluation methods, DEC and

TABLE I

The Area under the curve (AUC) for the DEC or IEC is based on the whole evaluation curve and limited to the range of the first 20%. For DEC, a higher value indicates superior PLFIQA performance, for the IEC a lower value indicates superior PLFIQA performance. On the 20% range of DEC and IEC, the PLQ maps outperform the baseline approach in all cases. On the whole evaluation curve, the random baseline performs better in some cases, probably due to the more equally distributed pixels.

AUC-DEC		ArcFace		CurricularFace	
		100%	20%	100%	20%
LFW	Baseline	39.45	1.47	37.37	0.72
	PLQ maps	37.90	2.07	37.67	1.96
AgeDB-30	Baseline	41.59	3.19	40.18	2.10
	PLQ maps	40.71	3.21	40.13	2.89
AUC-IEC					
LFW	Baseline	12.90	8.69	13.77	8.83
	PLQ maps	13.76	7.55	13.10	7.46
AgeDB-30	Baseline	21.57	8.92	21.02	8.94
	PLQ maps	17.42	7.85	17.62	7.87

IEC, evaluate how well the assessed pixel qualities influence the recognition performance of face recognition algorithms, i.e. utility. Using the proposed AUCs of these calculated curves, we can further quantify the result and simplify the comparison between different approaches. Calculating the AUC-DEC and AUC-IEC on the entire evaluation curve is beneficial for the baseline approach, as the randomly selected pixels are more equally distributed over the images than the calculated pixel-level qualities of the PLQ maps. The limitation to the first 20% shows its advantages here because mainly the decisive pixels are considered with less background and less informative pixels.

VI. Conclusion

In this paper, we proposed the first evaluation scheme for pixel-level face image quality assessment (PLFIQA) methods. Our proposed deletion evaluation curve (DEC) evaluates the decrease in recognition performance when pixels are iteratively masked based on their assigned quality value. On the other hand, our proposed insertion evaluation curve (IEC) investigates the increase in recognition performance when pixels are iteratively inserted into a highly blurred version of the original images. Calculating the area under the curve for both of these curves (AUC-DEC, AUC-IEC) allows further quantification of PLFIQA methods' performance. In the experiments, we investigated the suitability of our proposed evaluation methodology by looking at the PLQ maps of [9] and a baseline approach, based on random pixel selection. Our proposed evaluation method does not include human interaction and is more easily scalable regarding time and resources. It also does not take into account subjective human views but is restricted to the utility of the pixels to the recognition model. The proposed approach is not limited to the face modality and can be utilized to evaluate any biometric modality's pixel-level quality assessment methods.

References

[1] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[2] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," in *CVPR Workshops*. Computer Vision Foundation / IEEE, 2022.

[3] M. Huber, F. Boutros, F. Kirchbuchner, and N. Damer, "Mask-invariant face recognition through template-level knowledge distillation," in *FG*. IEEE, 2021, pp. 1–8.

[4] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ACM Comput. Surv.*, 2021.

[5] L. Best-Rowden and A. K. Jain, "Learning face image quality from human assessments," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 12, pp. 3064–3077, 2018.

[6] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality assessment for face recognition based on deep learning," in *ICB*. IEEE, 2019, pp. 1–8.

[7] ISO/IEC JTC1 SC37 Biometrics, "ISO/IEC 29794-1:2016 Information technology - Biometric sample quality - Part 1: Framework." International Organization for Standardization, 2016.

[8] B. Fu, C. Chen, O. Henniger, and N. Damer, "A deep insight into measuring face image utility with general and face-specific image quality metrics," in *WACV*. IEEE, 2022, pp. 1121–1130.

[9] P. Terhörst, M. Huber, N. Damer, F. Kirchbuchner, K. B. Raja, and A. Kuijper, "Pixel-level face image quality assessment for explainable face recognition," *CoRR*, vol. abs/2110.11001, 2021.

[10] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*. IEEE Computer Society, 2017, pp. 3449–3457.

[11] V. Petsiuk, A. Das, and K. Saenko, "RISE: randomized input sampling for explanation of black-box models," in *BMVC*. BMVA Press, 2018, p. 151.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[13] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: Adaptive curriculum learning loss for deep face recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 5900–5909.

[14] I. O. for Standardization, "Iso/iec 19794 information technology — biometric data interchange formats — part 5: Face image data," 2011.

[15] ISO/IEC JTC1 SC17 WG3, "Portrait Quality - Reference Facial Images for MRTD." International Civil Aviation Organization, 2018.

[16] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 14 225–14 234.

[17] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "CR-FIQA: face image quality assessment by learning sample relative classifiability," *CoRR*, vol. abs/2112.06592, 2021.

[18] 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 *CVPR*. Computer Vision Foundation / IEEE, 2020, pp. 5650–5659.

[19] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach, "Manipulating and measuring model interpretability," in *CHI*. ACM, 2021, pp. 237:1–237:52.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE Computer Society, 2016, pp. 770–778.

[21] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[22] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," in *CVPR Workshops*. IEEE Computer Society, 2017, pp. 1997–2005.

[23] E. Eidinger, R. Enbar, and T. Hassner, "Age and gender estimation of unfiltered faces," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 12, pp. 2170–2179, 2014.

[24] Frontex, "Best practice technical guidelines for automated border control (abc) systems," 2015.