A Deep Insight into Measuring Face Image Utility with General and Face-specific Image Quality Metrics

Biying Fu[1], Cong Chen[1], Olaf Henniger[1], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany
Email: biying.fu@igd.fraunhofer.de

Abstract

Quality scores provide a measure to evaluate the utility of biometric samples for biometric recognition. Biometric recognition systems require high-quality samples to achieve optimal performance. This paper focuses on face images and the measurement of face image utility with general and face-specific image quality metrics. While face-specific metrics rely on features of aligned face images, general image quality metrics can be used on the global image and relate to human perceptions. In this paper, we analyze the gap between the general image quality metrics and the face image quality metrics. Our contribution lies in a thorough examination of how different the image quality assessment algorithms relate to the utility for the face recognition task. The results of image quality assessment algorithms are further compared with those of dedicated face image quality assessment algorithms. In total, 25 different quality metrics are evaluated on three face image databases, BioSecure, LFW, and VGGFace2 using three open-source face recognition solutions, SphereFace, ArcFace, and FaceNet. Our results reveal a clear correlation between learned image metrics to face image utility even without being specifically trained as a face utility measure. Individual handcrafted features lack general stability and perform significantly worse than general face-specific quality metrics. We additionally provide a visual insight into the image areas contributing to the quality score of a selected set of quality assessment methods.

1. Introduction

Face recognition (FR) has gained high user acceptance due to the convenience and high accuracy [3]. Commercial products, such as Google Pay [28] and Apple Pay [31] integrate FR for user authentication to make digital payment easier and more secure. The automatic FR on border control further reduces the workload of border officers and accelerates the process. The widespread use of FR solutions are only possible due to the recent advances made in deep-learning based FR algorithms.

The quality of a face image directly affects the performance of the underlying FR systems. Face image quality assessment (FIQA) assigns a quality score to an input face image expressing its utility to support a correct outcome of an FR decision. The term "utility" considered in this work is based on the definition in ISO/IEC 29794-1 [26] relating the quality score (QS) to biometric performance. The output of a specific FIQA algorithm may depend on a specific FR system used for training. This entails that the face image utility is conditioned on both, the face image and the specific FR system. Such metrics should preferably be indifferent to the FR solution and work in general.

In contrast to FIQA, image quality assessment (IQA) works more generalized but depends on the subjective viewpoints of the end-users. IQA algorithms link the quality score (QS) of an image with the mean opinion scores (MOS) assigned by multiple human observers. Popular subjective IQA databases used for training IQA algorithms are e.g., LIVE [46], TID2008 [42], and TID2013 [41], each containing different types of visible image distortions, such as additive noise, compression errors, blurring, and contrast changes. Trained IQA methods as in [2, 29, 35] predict image quality without a reference image.

FIQA and IQA algorithms have been developed as two separate streams. Only limited research work can be found studying the relationship or the possible interoperable use between these two research streams as in [10, 1, 11]. Natural image statistics like BRISQUE, NIQE, and PIQE are sometimes used as baselines to emphasize the strong performance of learned face-specific quality metrics such as in [36, 49]. In this paper, we investigate IQA and FIQA algorithms to provide a deep insight into their role in face recognition systems. In addition, based on the utility conditioned on the FR performance for specifically designed experiments, we aim to answer questions related to the correlation of both, the general image quality metrics and the

face-specific quality metrics, to the face image utility.

To address these research challenges, we build our take-home conclusions based on the evaluation of 25 different quality metrics, which can be categorized into four groups: 1) general image quality measures, 2) handcrafted image quality measures, 3) face related handcrafted measures, and 4) learned face utility measures. The paper is structured as follows: In Section 2, we show the high-level progress made in the fields of IQA and FIQA methods. In Section 3 and Section 4, we introduce selected algorithms from both research domains to draw comparisons by conducting experiments (in Section 5) specifically designed to address three research questions, followed by the results, analysis and a visual interpretation (Section 6) to emphasize our findings. Section 7 recapitulates the quintessence of the paper.

2. Related work

Perceptual IQA assesses the distortion and degradation on the visual material, such as compression, white noise (WN), or Gaussian blur (GB). These types of distortion can also cause the face image utility to drop. The research domain of IQA is sub-divided into full-reference, reduced-reference, and no-reference IQA. Full-reference IQA algorithms require an original image for comparison, e.g., peak signal-to-noise ratio (PSNR) [24], and structural similarity index measure (SSIM) [52]. In contrast to these methods, the Reduced-reference IQA algorithms provide a solution for image quality estimation where only partial information is accessible e.g., in [43, 7, 54, 30].

Usually one does not have a reference image directly to verify the comparison score in an FR system. Therefore no-reference IQA is the most similar case to FIQA. Dutta et al. proposed in [10, 11] to utilize a Bayesian framework to link feature-based FIQA (e.g., focus, pose, and illumination direction) to predict FR performance, but did not consider IQA methods in general. However, to our knowledge, only limited previous works explore the relation between no-reference IQA methods and the face utility.

ISO/IEC TR 29794-5 [27] defines image-level features, such as illumination symmetry, inter-eye distance, blur, sharpness, and person-dependent features such as wearing beards, eyeglasses, or eyes/mouth closed/opened as required in [25]. These features are assumed to have a direct impact on the face utility. While these generated features are explainable and directly connected with image qualities, several current FIQA algorithms [21, 49, 36] rely on methods using end-to-end learning. This trend is due to the significant improvement of deep-learning-based (DL-based) FR solutions in both industry and academic fields. Only very few works focus on relating FIQ with morphed faces [15] or face parts [14] to enhance interpretability of these measures.

3. Face image quality assessment algorithms

We selected six DL-based FIQA methods as they demonstrated state-of-the-art performances and are based on various training and conceptualization strategies. They can be grouped into categories of either supervised on quality labels, e.g., FaceQnet [21] or unsupervised methods, e.g., rankIQ [6], MagFace [36], SER-FIQ [49], PFE [47], and SDD-FIQA [40]. We further examine explainable image-level features as proposed by ISO/IEC TR 29794-5 [27].

3.1. Deep learning-based FIQA Methods

RankIQ [6] is a model trained to assess face utility using a ranking-based approach. The author is inspired by the premise that the quality of facial images cannot be quantified absolutely and is easier to be considered in a relative manner. This method was trained using three databases with varying qualities. The training is a two-stage process. While stage I learns to map the individual face image features (e.g., HoG, Gabor, LBP, and CNN features) to first level rank weights, the stage II maps the learned feature scores to a final normalized quality score by using kernels.

FaceQnet [21] by Hernandez-Ortega et al. is a supervised, and DL-based method trained on VGGFace2 [4] database. We used FaceQnet v2 which is the most recent version of FaceQnet [20][2]. The BioLab-ICAO framework [12] was used to select ICAO-compliant high-quality reference images. FaceQnet then fine-tuned a pre-trained FR base-network (RseNet-50 [19]) and the successive regression layer on top of the feature extraction layers to associate an input image to a utility score. The target utility score is the normalized similarity score between the face image and a high-quality mated reference image.

SER-FIQ [49] is an unsupervised DL-based FIQA method based on a stochastic method applied on face representations. This method mitigates the need for any automated or human labeling. The face image was passed to several sub-networks of a modified FR network by using dropout. Images with high-utility are expected to possess similar face representations resulting in low variance. Therefore, the proposed method linked the robustness of face embeddings directly with face utility. In our study, we used the method finetuned with ArcFace loss [9] using ResNet-100 [19] base architecture trained on MS1M database [17] named the SER-FIQ (on ArcFace) method.

Shi et al. proposed the **Probabilistic Face Embeddings (PFEs)** [47], which represents each face image as a Gaussian distribution in the latent space. The model is trained to maximize the mutual likelihood score of all genuine pairs equally to map distorted input face to its genuine latent space. The mean of the Gaussian represents the most likely features and the variance indicates the uncertainty in the

[2]FaceQnet Github: https://github.com/uam-biometrics/FaceQnet

features values. In this paper, we refer the variance to face utility, where a low quality face image possess a larger uncertainty in the latent embedding space.

MagFace [36] by Meng et al. is a recently developed method to derive both the face representation and the face image quality from calculating the magnitude of the face embedding. They extended the ArcFace loss [9] by an adaptive margin regularization term to further enforce the easily recognizable samples towards the class center and the hard samples further away. The face utility is inherently learned through this loss function during training. The magnitude of the feature vector is proportional to the cosine distance to its class center and is directly related to the face utility.

SDD-FIQA by Ou et al. [40] is a novel unsupervised FIQA method that incorporates Similarity Distribution Distance for FIQA. The method leverages the Wasserstein Distance (WD) between the inter-class samples (Neg-Similarity) and the intra-class samples (Pos-Similarity). The FR model uses ResNet-50 [19] trained on MS1M database to calculate the positive samples and negative samples distributions. This WD metric is used as the quality pseudo-labels to train a regression network for quality prediction with Huber loss.

3.2. FIQA metrics based on Handcrafted Features

We chose eight representative features to extract information from a face image according to the ISO/IEC Technical Report [27]. We aim to relate these individual features to assess the utility of the underlying FR systems. Features derived from the spatial domain are blur, contrast, mean, luminance, sharpness, lighting symmetry, and exposure. They depend on the pixel intensity and its statistical distribution. Inter-eye distance measures the distance between two eye middle points from the original image and is assumed to be directly related to face image utility. The set of features implemented are seen in the legend of Figure 1.

4. Image quality assessment algorithms

In this section, we present ten IQA methods, which can be categorized into: (1) model-based, (2) CNN-based, (3) multi-task learning-based, and (4) rank-based approaches. In our experiment, we associate these IQA metrics to face image utility. These methods are not specifically designed to assess face image utility for face recognition.

4.1. Model-based IQA methods

This IQA category tries to build a model to assess the general image quality based on natural image statistics. Here, we cite three methods **BRISQUE** [37], **NIQE** [38], and **PIQE** [50] that all based on studying the deviation from the general statistics of natural images. This statistical model is derived from the finding by Rudermann [44] that natural scene images have a luminance distribution similar to a normal Gaussian distribution. The degree of deviation from the normal Gaussian distribution relates to the degradation in image quality. While NIQE and PIQE are both opinion-unaware methods without the need for human-rated scores, BRISQUE requires training with human opinion scores. We selected these three methods, as they were used as baseline for performance comparison in FIQA methods like SER-FIQ and MagFace. The results demonstrated on the error vs. reject characteristic curve showed that these three methods are inferior compared to the SER-FIQ algorithm concerning face quality assessment.

4.2. CNN-based IQA algorithms

CNNIQA[29] and DeepIQA [2] are both methods that leverage CNNs in the base layers to automatically extract spatial image features from the input image avoiding generating handcrafted features as opposed to methods in the previous sub-section. While **CNNIQA** [29] used a shallow net with only one convolutional layer, the **DeepIQA** [2] proposed by Bosse et al. applied a deeper structure with multiple stacked convolutional layers. Compared to CNNIQA, the increased model capacity allows the model to deal with more complex and colored images. Both methods are derived from patch-wise quality score where local metrics are aggregated to a global metric resulting in a whole image quality estimate. Although other methods can be based on CNNs, we refer in this category to methods with single-task learning (i.e. CNNs solely trained to estimate a quality value).

4.3. Multi-task learning-based IQA algorithms

Training multiple tasks within one network structure often enhance the network ability in performing one specific task [5]. MEON [35] and DBCNN [57] are both mutlitask-learning-based approaches trained to classify the type of image distortions on the one hand to enhance the ability of estimating the image quality on the other hand in one combined network. **MEON** [35] used a shared network as base structure with two parallel networks added on top, one for identifying the type of distortion (Subtask I) and one for predicting the quality of a given input image (Subtask II). Subtask I can benefit from generating large quantity of synthetic data at low cost. Similar concept is used for **DBCNN** [57] where two CNNs were trained at the base, one on large-scale synthetically generated databases, while the other focuses on the classification network pre-trained on ImageNet [8] for extracting more authentic distortions. Finally, the features from both CNNs are pooled bilinearly into a unified representation for a final quality prediction.

4.4. IQA based on ranked image pairs

Since image quality is not homogeneously quantifiable, the mean opinion score (MOS) from human operators is

Method	Conference, Year	Category	Sub-category	Un-/supervised
RankIQ [6]	IEEE SPL, 2015	FIQA	Feature Fusion	unsupervised
FaceQnet [20]	arXiv preprint, 2020	FIQA	ResNet-50	supervised
SER-FIQ [48]	CVPR, 2020	FIQA	ResNet-100	unsupervised
PFE [47]	ICCV, 2019	FIQA	64-layer CNN	unsupervised
MagFace [36]	CVPR, 2021	FIQA	ResNet-100	unsupervised
SDD-FIQA [40]	CVPR, 2021	FIQA	ResNet-50	unsupervised
BRISQUE [37]	IEEE IP, 2012	IQA	Model-based	supervised
PIQE [50]	NCC, 2015	IQA	Model-based	unsupervised
NIQE [38]	IEEE SPL, 2012	IQA	Model-based	unsupervised
CNNIQA [29]	CVPR, 2014	IQA	CNN-based	supervised
DeepIQA [2]	IEEE TIP, 2018	IQA	CNN-based	supervised
MEON [35]	IEEE TIP, 2018	IQA	Multitask-based	unsupervised
DBCNN [57]	AICCSA, 2020	IQA	Multitask-based	semi-supervised
dipIQ [34]	IEEE TIP, 2017	IQA	Ranking-based	unsupervised
RankIQA [33]	ICCV, 2017	IQA	Ranking-based	unsupervised
UNIQUE [58]	IEEE TIP, 2021	IQA	Ranking-based	supervised

Table 1. Table provides an overview of the used methods in both categories of IQA and DL-based FIQA.

considered a useful metric. However, the process of obtaining MOS is tedious. Hence, using relative comparison without quantifiable metric simplifies the design of IQA metrics. UNIQUE [58], rankIQA [33], and dipIQ [34] are methods that use ranked image pairs to avoid the absolute scale of image quality. Large databases with ranked image pairs can be generated at low cost synthetically as in [34, 33] or assembled as in [58]. **dipIQ** [34] leveraged a two-stream networks trained on quality-discriminative pairs (DIP) of images. Each DIP is associate with a degree of perceptual uncertainty. **RankIQA** [33] used a Siamese network to train on pairs of ranked inputs. Subsequently, the trained Siamese Network can be used to teach a traditional CNN that estimates the image quality using only one single input. **UNIQUE** [58] used a deep neural network structure. The training is based on pairs extracted from multiple IQA database to acquire the MOS score and the corresponding variance. The network training uses the fidelity loss to optimize the DNN over a large number of such image pairs and the hinge loss to regularize the uncertainty estimation during optimization. The network learns the mean and the variance of an input image which represent the quality score and the uncertainty respectively.

An overview of the used methods are found in Table 1, where the previous works in both categories of IQA and DL-based FIQA are introduced in terms of its year of publication, categories, sub-categories, and way of training.

5. Experimental Evaluation

In this section, we first introduce the used databases, FR solutions, and evaluation metrics. Then, the experimental results will be presented in a comprehensible manner.

5.1. Face Image Databases

Three face image databases are used to link the scores estimated by FIQA methods in Section 3 and IQA methods in Section 4 with the face image utility.

The **BioSecure** [39] DB contains face images of 210 subjects with only frontal views and highly controlled quality data. We chose this database for the following reasons: 1) it represented the controlled and collaborative use case scenario relevant for border checks and identity documents (ISO/IEC 19794-5), and 2) it was reported for face quality in ([21] ICB2019). We conducted a series of experiments on this database. However, the results show that due to the high-quality images, the FR systems perform almost perfectly.

The Labeled Faces in the Wild (**LFW**) DB [23] is a widely used standard benchmark for automatic face verification. It contains in total 13233 images from mostly uncontrolled scenarios. We chose this database as it is used in the FIQA methods [49, 36]. However, this database is strongly imbalanced regarding the number of images for each subject.

Similarly, **VGGFace2** [4] test contains 500 subjects. We chose VGGFace2 due to its large variety in quality distribution which can be considered a challenging database. The images have diverse and complex acquisition conditions. To manage the heavy computation due to the large database, we selected a representative subset by randomly choosing 30 out of 300 images from each subject to perform the 1:1 verification task.

The MTCNN framework [56] is used to detect, align, and crop the input face images to a fixed size of 260×260 pixels. The set of used images are subsequently adapted to the input size of each of the used networks.

5.2. Face recognition solutions

Three open-source academic FR solutions are used to derive the face embeddings to perform the verification task.

We chose **Facenet** [45] because it is one of the first FR solutions based on deep CNN structures using inception resnet as backbone. Triplet loss and center loss are used to facilitate the training. The accuracy reported on the LFW DB is 99.63%$\pm$0.09 and on YouTube Faces DB (YTF) [53] is 95.12%%$\pm$0.39.

SphereFace [32] used a 64-CNN layers trained on CASIA-WebFace [55]. We chose this FR model as it achieved a competitive state-of-the-art verification accuracy on LFW DB to 99.42% and YTF DB to 95.0%.

ArcFace [9] is trained on ResNet-100 [18] using the MS1M dataset [17]. The loss function further uses additive angular margin to improve the discriminative power of the FR model. This model is chosen due to its improved accuracy on LFW 99.83% and YTF DB 99.02%.

5.3. Evaluation metric

The evaluation metric used is the error vs. reject characteristic (ERC) [16]. The ERC shows the relative performance when rejecting different ratios of the evaluation data with the lowest error. With a "better" face utility estimation,

the face verification error should strongly decrease when rejecting more low-quality data. In our presented ERC, we show the false non-match rate (FNMR) at different ratios of rejected (low quality) images. The presented FNMR is the FMR1000, i.e. the FNMR at false match rate (FMR) value of 0.1% as recommended for border operations by Frontex [13]. For a well-functioning face utility estimation, the FNMR is expected to go down as the ratio of discarded (rejected) images increases.

For LFW DB, we used the test protocol as reported in [22] to balance the database. The original database contains 5749 subjects, but only 1680 subjects have two or more images. BioSecure and VGGFace2 include already a balanced database and are used in verification scenarios comparing every image with every other image.

To quantitatively represent the correlation between quality estimation methods and categories, we calculate the samples overlap ration between the samples of the lowest quality (10% of the data) between every pair of quality estimation methods, and the same for the 10% of the highest quality. A large overlapping ratio indicates a larger reasoning similarity between the considered pair of methods.

Table 2 provides an overview of the three top-performing methods compared in sub-groups of DL-based FIQA, feature-based FIQA, and learned IQA methods, for all two face image databases (LFW, and VGGFace2) at FMR1000. The best performing three methods in each setup are emphasized in bold. We report the two reject ratios at 20 % and 40 %. The full table including the results of all methods is available in the supplementary material for a more comprehensive study.

5.4. Results on face image databases

This section presents the evaluation results based on three database and three SOTA FR solutions. The results and discussions are divided into answering separate research questions regarding the effect of IQA and FIQA metrics on the face image utility.

The BioSecure DB is a high-quality database with a controlled capture environment. Due to the consistent high-quality of the images in the DB, the ERC curve shows a very low FNMR and does not reveal any apparent changes across different reject ratios. Since all FR systems perform almost perfectly on the entire database, no interesting observations can be found. However, it was essential to include this database as it represents the controlled (passport-like) data presentation scenario. Therefore, we only included the results in the supplementary materials for completeness.

Q1: How does FIQA metrics using handcrafted features correlate to face image utility? Figure 1 shows the ERCs at FMR1000 using handcrafted image features as quality metrics. Observing the results on VGGFace2, the inter-eye distance feature seems to be most effective in selecting high-utility face images. However, inter-eye distance cannot obtain consistent results on all DBs. For LFW at FMR1000 with ArcFace embeddings at the reject ratio of 20%, the mean-feature with an FNMR=0.532% almost halved the error rate compared to the inter-eye distance feature with an FNMR=0.903%. This may be due to the fact that handcrafted features such as the inter-eye distance is naive and assume the images are not further processed after capture (not scaled, scanned, etc.) and that the capture device is consistent (the same amount of pixels corresponds to same quality). These strict assumptions do not apply to a database collected from various sources such as LFW. The data collection process of LFW may undergo special processing. Hence it will affect the performance of handcrafted features. In LFW, no single handcrafted feature excels the others and the changes in the error performance are less obvious. The total reduction in error for handcrafted image features accounts less compared to FIQA methods.

Even though most handcrafted features show high intra-category correlation, e.g., contrast and exposure with an overlapping ratio of 39.53%, or contrast and mean with 30.73%, according to Figure 4, the inter-category correlation is relative low for these metrics. In general, individual features [27] show low impact on selecting high-utility face images. This is confirmed in Figure 4 (left) where an average ratio of less than 1.5% is shown between the FIQA handcrafted features and DL-based FIQA methods. To sum up, these handcrafted features are sensitive to the image capture setups and the post-processing steps performed on the captured images. Therefore, these features might be less useful in comparison with learned FIQA methods.

Q2: How does IQA metrics correlate to face image utility? Figure 2 depicts the ERCs at FMR1000 using the DL-based FIQA methods from Section 3 (solid lines) and learning-based IQA methods from Section 4 (dashed lines). Taking a look at the results on LFW (the left three plots in Figure 2), we observed that error is reduced due to the increasing amount of dropped low-quality images. In contrast to the findings on LFW, ERC on VGGFace2 (the right three plots in Figure 2) exhibits a strong correlation between the learned IQA methods towards image utility. The curves using IQA methods reveal a distinctly decreasing FNMR when discarding bad quality images. The total reduction in error for learning-based IQA methods is more evident compared to FIQA methods or handcrafted quality metrics. Observing the ERC curve for VGGFace2 at FMR1000 with ArcFace embeddings, the error in terms of FNMR for CNNIQA dropped around 36.7% from 20% to 40% reject ratios, whereas FaceQnet only reduced about 19% and the inter-eye-distance feature about 16%.

Table 2 showed that although FIQA metrics outperform

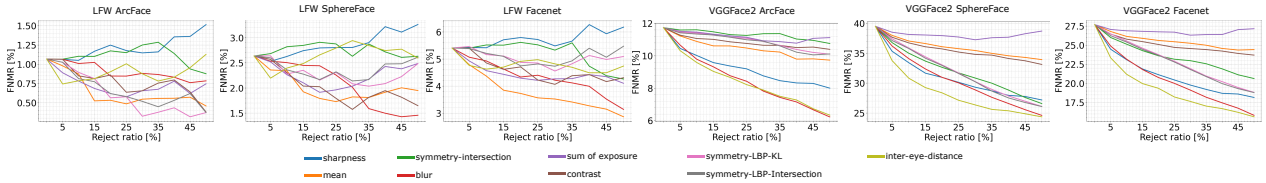


Figure 1. Error vs. reject characteristics at FMR1000 for handcrafted image quality measures and a face-related handcrafted measure like inter-eye distance. The rows reveal the ERC results for different face embeddings on LFW and VGGFace2. Inter-eye distance performed well on VGGFace2 using original images, while sharpness and blur are well performed for aligned images. However, individual feature contributes inconsistently across different settings, while a fusion could be more promising.

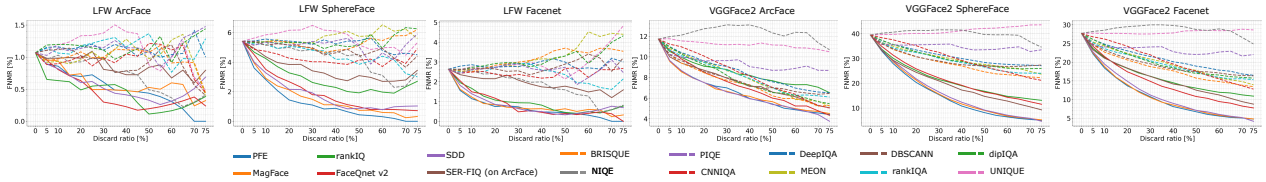


Figure 2. Error vs. reject characteristics at FMR1000 for IQA (dashed lines) and FIQA (solid lines) methods. The rows show the results for diverse face embeddings on LFW and VGGFace2. DL-based FIQA methods outperform IQA methods, while for VGGFace2 a clear decreasing trend can be observed for IQA methods as well.

the most other metrics according to the verification results. Nevertheless, in most settings, the learned IQA methods can also compete for the top-3 rank. As on VGGFace2, they all show a strong correlation with a clear trend towards decreasing error rates, with no clear winner in this category. The same result is confirmed in Figure 4. Besides the intra-category correlation, we also observed high overlapping ratios between the IQA metrics and the DL-based FIQA metrics, unlike handcrafted features and DL-based FIQAs. Especially IQA methods, such as PIQE, DeepIQA and DBCNN showed large overlap to MagFace, SDD-FIQA, and PFE, where the highest overlap is observed for DBCNN to PFE (44.20%). These three methods are representative of the model-based, CNN-based, and multitask-based IQA approaches, while the ranking-based approaches showed less correlation to DL-based FIQAs.

Looking at the score distributions in Figure 3, we observed that VGGFace2 database has a wider score distribution compared to LFW and Biosecure. This diversity can be the reason behind effectiveness of IQA on VGGFace2 database compared to the other databases.

To conclude, there is a clear correlation between IQA and face image utility. The drop in the ERC curve shows the effectiveness of the IQA metrics with respect to select the face image with high utility. This drop is relatively stronger for model-based and multitask-based IQA approaches.

Q3: How does DL-based FIQA metrics correlate to face image utility? Considering the verification results in Table 2, in general, one of the DL-based FIQA methods always dominates the top verification results independent of the database. Inspecting the results for VGGFace2 at FMR1000, SDD-FIQA and MagFace are the methods that have occupied the top rankings most of the times. One explanation for the good outcomes in terms of error performance is that FIQA methods are trained with face images for specific FR systems as in the proposed methods [40, 36].

Typical FIQA methods like SDD-FIQA, MagFace, SER-FIQ, and PFE, already learned discriminative face embeddings in the training phase using appropriate loss functions. Other FIQA methods like rankIQ and FaceQnet learned the utility directly from relative FR comparison scores. Moreover, face image utility is not fully represented by the perceptual quality and is not easily quantifiable. Therefore, opposite to general IQA metrics, FIQA inherently learn the face utility. In Figure 4, most DL-based FIQA methods show a high intra-category correlation. Especially the unsupervised methods, such as SER-FIQ, MagFace, SDD-FIQA, and PFE revealed a strong intra-category correlation, while FaceQnet and rankIQ possess low correlation to other DL-based FIQA metrics. The highest overlap is observed for SDD-FIQA to PFE with an overlap of 67.40% and MagFace to PFE with 66.67%.

To summarize, FIQA metrics are designed to relate a face image to face utility. Even trained for certain face embeddings, most FIQA methods work efficiently across different FR solutions. Unsupervised methods, such as SDD-FIQA, PFE, and MagFace outperform other supervised methods, like FaceQnet.

6. Visualization of selected model decisions

In addition, to evaluate the effectiveness of the different investigated methods in measuring face image utility, we go further by exploring the image parts that contribute the most to their performance, especially those from IQA and

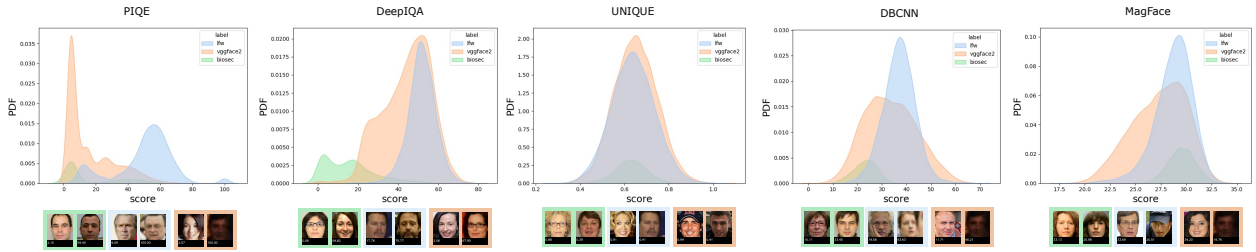


Figure 3. Score distribution for the three face database using one representative method from 4 IQA sub-categories and 1 FIQA method for reference. For each method and database, one high utility sample (left) and one low utility sample (right) are displayed. VGGFace2 has a wider score distribution compared to LFW and Biosecure in most cases, which might indicate that the IQA methods are more effective on this database.

LFW at FMR 1000												
	ArcFace				Sphereface				FaceNet			
	20%		40%		20%		40%		20%		40%	
DL-based FIQA	**rankIQ**	**0.488**	**FaceQnet**	**0.226**	**PFE**	**1.444**	**PFE**	**0.877**	**PFE**	**0.747**	**FaceQnet**	**0.453**
	SDD	**0.602**	**SDD**	**0.412**	**MagFace**	**2.125**	**SDD**	**1.072**	**MagFace**	**0.840**	**PFE**	**0.478**
	FaceQnet	0.632	**PFE**	**0.478**	**SDD**	**2.159**	**MagFace**	**1.118**	**SDD**	**0.853**	**SDD**	**0.495**
Image Quality	NIQE	0.925	NIQE	0.780	CNNIQA	4.263	CNNIQA	3.930	CNNIQA	2.385	NIQE	1.994
	MEON	0.959	DBCNN	0.786	NIQE	4.989	rankIQA	4.188	NIQE	2.417	CNNIQA	2.183
	CNNIQA	1.015	CNNIQA	1.048	rankIQA	5.071	DBCNN	4.628	DBCNN	2.651	DBCNN	2.532
Feature FIQA	inter eye dist	.9027	inter eye dist	.8319	inter eye dist	4.9147	inter eye dist	4.4925	inter eye dist	2.6579	inter eye dist	2.7454
	mean	**.5326**	mean	.5586	mean	3.7288	mean	3.2721	mean	1.7917	mean	1.9154
	sharpness	1.2487	sharpness	1.3559	sharpness	5.7942	sharpness	6.2711	sharpness	2.7972	sharpness	3.2203

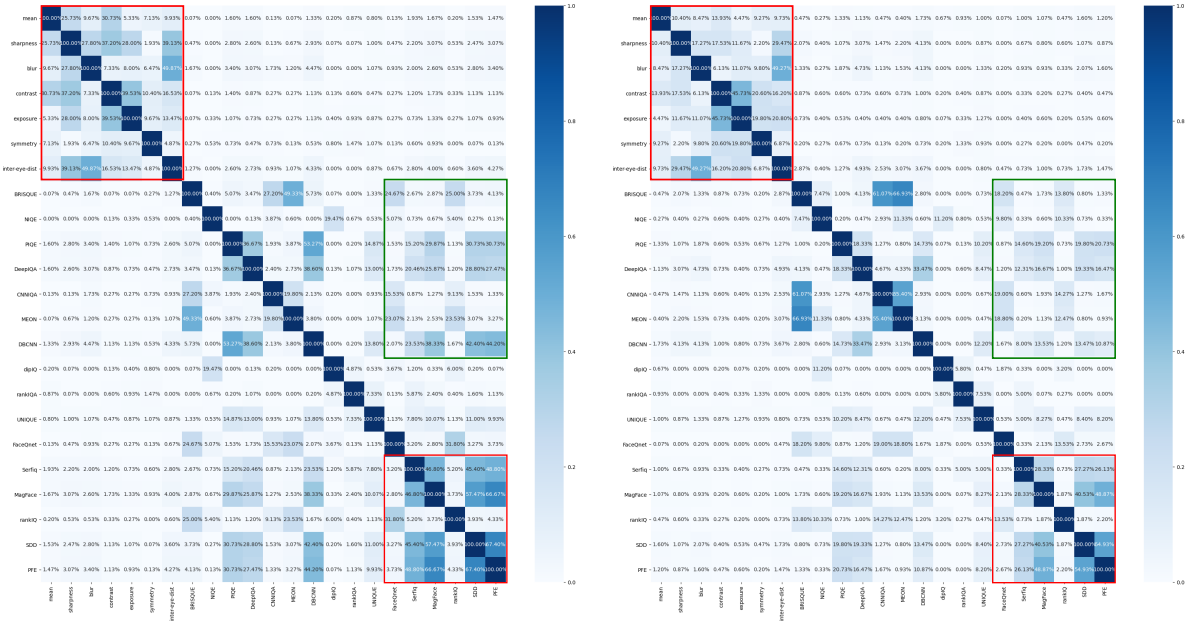
VGGFace2 at FMR 1000												
	ArcFace				Sphereface				FaceNet			
	20%		40%		20%		40%		20%		40%	
DL-based FIQA	**PFE**	**7.505**	**SDD**	**5.931**	**PFE**	**20.394**	**PFE**	**10.836**	**PFE**	**13.694**	**PFE**	**8.127**
	SDD	**7.508**	**PFE**	**5.955**	**MagFace**	**21.329**	**MagFace**	**11.650**	**MagFace**	**13.993**	**MagFace**	**8.540**
	MagFace	**7.520**	**MagFace**	**6.171**	**SDD**	**21.836**	**SDD**	**12.371**	**SDD**	**14.905**	**SDD**	**8.987**
Image Quality	BRISQUE	8.468	dipIQ	7.234	BRISQUE	30.706	BRISQUE	27.002	BRISQUE	20.593	BRISQUE	17.683
	DBCNN	8.565	DBCNN	7.246	DBCNN	30.813	DBCNN	27.502	DBCNN	21.012	DBCNN	18.144
	dipIQ	9.248	BRISQUE	7.292	rankIQA	32.459	rankIQA	28.355	rankIQA	22.066	dipIQ	18.519
Feature FIQA	inter eye dist	8.669	inter eye dist	7.2749	inter eye dist	28.378	inter eye dist	25.373	inter eye dist	19.364	inter eye dist	16.701
	mean	10.611	mean	9.7969	mean	36.088	mean	34.573	mean	25.628	mean	24.578
	sharpness	9.373	sharpness	8.3267	sharpness	30.967	sharpness	27.952	sharpness	21.145	sharpness	18.704

Table 2. Comparison between the top-3 performing DL-based FIQA, feature-based metrics, and IQA methods on two facial image DBs evaluated for the FNMR at two reject ratios (20 % and 40 %) at FMR1000 based on three FR models (ArcFace, SphereFace, Facenet). In bold are the best performing three methods across all categories.

FIQA categories. Figure 5 visualizes three selected IQA methods (left) and three FIQA methods (right). We use the Score-CAM [51] to display the attention map of the network's decision. Score-CAM is designed to display visual explanations for CNNs and works more efficient compared to other gradient-based visualization methods. UNIQUE, FaceQnet, MagFace, and SDD-FIQA are visualized using their ResNet-X base architecture. This visualization is unfortunately not applicable to other methods, such as patch-based approaches, or rankIQ due to the fusion of selected features in a two-stage process, and the SER-FIQ due to multiple dropout runs to modify network architectures.

Visualization for IQA methods: For patch-based IQA, such as CNNIQA and DeepIQA, pixel-wise network decision is visualized in Figure 5. It is observed that these IQA methods also paid attention to the background (see first and third images in Figure 5(a)) for CNNIQA. Such behavior might be the reason for the lower performance. Similar results are seen for DeepIQA (Figure 5(b)). Score-CAM is used for the ranking-based IQA method UNIQUE. Consistent finding with the other two IQA methods are observed for the third and sixth image in Figure 5(c). Opposite to DL-based FIQA methods, the networks attention for IQA methods focuses not fully on the center face area.

Visualization for FIQA methods: Looking at Figure 5 (d) to (f), the activation area for FIQA methods cover mostly the central part of a face. From Figure 5(a) to (c), it is already noticed that the IQA methods have also relatively higher contributions on non-facial areas of the image. In contrast, the FIQA methods mainly focus on the facial areas and neglect the background. This may be one of the main reasons behind the superior performance of such methods. Furthermore, it is to note that SDD-FIQA and MagFace have more locally refined attention compared to FaceQnet.



Bottom 10% Lowest Quality

Top 10% Highest Quality

Figure 4. The confusion matrix shows the ratio of overlapped samples between the samples with the lowest/highest 10% qualities (lowest on the left matrix and highest on the right matrix) as measured by two quality estimation methods (on the X and Y axes). A high intra-category overlap is seen in the handcrafted-based FIQA, as well as, the DL-based FIQA, both in red squares. A relatively high inter-category correlation is noticed between IQA and DL-based FIQA methods (in green box). This matrix is build on VGGFace2. Additional results for other two databases are provided in the supplementary materials.

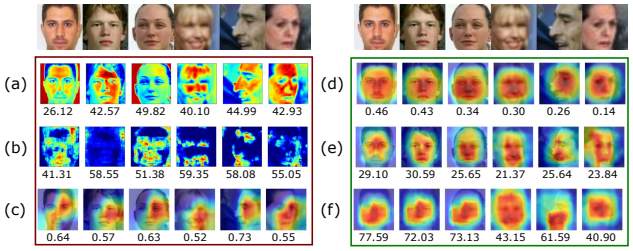


Figure 5. visualizes images of diverse network's attention and face quality score. Left are the three IQA methods: (a) CNNIQA, (b) DeepIQA, and (c) UNIQUE, and right are the three FIQA methods: (d) FaceQnet, (e) MagFace, and (f) SDD-FIQA. The IQA methods focus their decision additionally on the background, while FIQA methods focus mainly on center face area.

7. Conclusions

To address research questions concerning the face image utility correlation with general image quality and face-specific quality metrics, we thoroughly investigated a total of 25 quality metrics. We divided these into four different categories in 1) general image quality measures, 2) handcrafted image quality measures, 3) face-related handcrafted measures, and 4) learned face utility measures. These quality metrics are evaluated on three databases and three FR solutions and provided a deep discussion on the relationship between these families of quality estimation solutions.

Our evaluation showed little influence of single handcrafted feature as face image utility. Although these features revealed a clear intra-category correlation, they show less inter-category correlation to other learned IQA or FIQA metrics. Therefore, they demonstrate no clear indications to be useful as a generalized metric to assess face image utility. DL-based FIQA methods, as a face image utility predictor, are optimized for face images and show superior performance over IQA methods across different setups. Nevertheless, IQA methods show a clear correlation to face image utility, even though they do not outperform DL-based FIQA methods. Visualization of the IQA methods output revealed a focus on areas in the background, rather than solely on the face as in the learned FIQA methods. Accompanied with the advantage of FR model-independent training of such IQA methods, combining IQA metric with DL-based FIQA metric could lead to a more generalized measure across different FR systems and application scenarios.

Acknowledgements: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Lacey Best-Rowden and Anil K Jain. Learning face image quality from human assessments. *IEEE Transactions on Information Forensics and Security*, 13(12):3064–3077, 2018.

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE TIP*, 27(1):206–219, 2017.

[3] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. *CoRR*, abs/2109.09416, 2021.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Int. Conf. on Automatic Face and Gesture Recognition*, 2018.

[5] Rich Caruana. Algorithms and applications for multitask learning. In *ICML*, pages 87–95. Morgan Kaufmann, 1996.

[6] Jiansheng Chen, Yu Deng, Gaocheng Bai, and Guangda Su. Face image quality assessment based on learning to rank. *IEEE SPL*, 22(1):90–94, 2015.

[7] Naser Damer, Timotheos Samartzidis, and Alexander Nouak. Personalized face reference from video: Key-face selection and feature-level fusion. In *Face and Facial Expression Recognition from Real World Videos – Int. Workshop, FFER@ICPR 2014, Stockholm, Sweden, August 24, 2014, Revised Selected Papers*, volume 8912 of *Lecture Notes in Computer Science*, pages 85–98. Springer, 2014.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF CVPR*, pages 4690–4699, 2019.

[10] Abhishek Dutta, Raymond Veldhuis, and Luuk Spreeuwers. A bayesian model for predicting face recognition performance using image quality. In *IEEE Int. Joint Conf. on Biometrics*, pages 1–8. IEEE, 2014.

[11] Abhishek Dutta, Raymond Veldhuis, and Luuk Spreeuwers. Predicting face recognition performance using image quality. *arXiv preprint arXiv:1510.07119*, 2015.

[12] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. Face image conformance to ISO/ICAO standards in machine readable travel documents. *IEEE Trans. Inf. Forensics Secur.*, 7(4):1204–1213, 2012.

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

[14] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. The relative contributions of facial parts qualities to the face image utility. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2021.

[15] Biying Fu, Noémie Spiller, Cong Chen, and Naser 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.

[16] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, Apr. 2007.

[17] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conf. on Computer Vision*, pages 87–102. Springer, 2016.

[18] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *CVPR*, pages 6307–6315. IEEE Computer Society, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[20] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *CoRR*, abs/2006.03298, 2020.

[21] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. FaceQnet: Quality assessment for face recognition based on deep learning. In *Int. Conf. on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[22] Gary B Huang and Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep*, 14(003), 2014.

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

[24] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.

[25] Portrait quality (reference facial images for MRTD). ICAO Technical Report, 2018.

[26] Information technology – Biometric sample quality – Part 1: Framework. Int. Standard ISO/IEC 29794-1, 2016.

[27] Information technology – Biometric sample quality – Part 5: Face image data. Technical Report ISO/IEC TR 29794-5, 2010.

[28] J Janssen. Google Pay startet in Deutschland: Bezahlen mit dem Android-Handy. https://www.heise.de, 2018.

[29] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proc. of the IEEE Conf. on CVPR*, pages 1733–1740, 2014.

[30] Qiang Li and Zhou Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE journal of selected topics in signal processing*, 3(2):202–211, 2009.

[31] Stephanie Q Liu and Anna S Mattila. Apple Pay: Coolness and embarrassment in the service encounter. *Int. Journal of Hospitality Management*, 78:268–275, 2019.

[32] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[33] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *Proc. of the IEEE Int. Conf. on Computer Vision*, pages 1040–1049, 2017.

[34] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans. on Image Processing*, 26(8):3951–3964, 2017.

[35] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE TIP*, 27(3):1202–1213, 2017.

[36] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2021.

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

[38] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2012.

[39] Javier Ortega-Garcia, Julian Fiérrez, Fernando Alonso-Fernandez, Javier Galbally, Manuel R. Freire, Joaquin Gonzalez-Rodriguez, Carmen García-Mateo, José Luis Alba-Castro, Elisardo González-Agulla, Enrique Otero Muras, Sonia Garcia-Salicetti, Lorène Allano, Van-Bao Ly, Bernadette Dorizzi, Josef Kittler, Thirimachos Bourlai, Norman Poh, Farzin Deravi, Ming W. R. Ng, Michael C. Fairhurst, Jean Hennebert, Andreas Humm, Massimo Tistarelli, Linda Brodo, Jonas Richiardi, Andrzej Drygajlo, Harald Ganster, Federico Sukno, Sri-Kaushik Pavani, Alejandro F. Frangi, Lale Akarun, and Arman Savran. The multiscenario multienvironment biosecure multimodal database (BMDB). *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(6):1097–1111, 2010.

[40] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. *CoRR*, abs/2103.05977, 2021.

[41] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[42] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti. TID2008 – a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009.

[43] Abdul Rehman and Zhou Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE transactions on image processing*, 21(8):3378–3389, 2012.

[44] Daniel L Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, 1994.

[45] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[46] HR Sheikh. LIVE image quality assessment database release 2. `http://live.ece.utexas.edu/research/quality`, 2005.

[47] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *ICCV*, pages 6901–6910. IEEE, 2019.

[48] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE Int. Joint Conf. on Biometrics (IJCB)*, pages 1–11. IEEE, 2020.

[49] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 5651–5660, 2020.

[50] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. Blind image quality evaluation using perception based features. In *21st National Conf. on Communications (NCC)*, pages 1–6. IEEE, 2015.

[51] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[52] Z Wang, E P Simoncelli, and A C Bovik. Multiscale structural similarity for image quality assessment. In *Proc. 37th Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1398–1402, Pacific Grove, CA, Nov. 2003. IEEE Computer Society.

[53] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.

[54] Jinjian Wu, Weisi Lin, Guangming Shi, and Anmin Liu. Reduced-reference image quality assessment with visual information fidelity. *IEEE Transactions on Multimedia*, 15(7):1700–1705, 2013.

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

[56] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[57] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.*, 30(1):36–47, 2020.

[58] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Trans. Image Process.*, 30:3474–3486, 2021.