Face Image Quality Assessment: A Literature Survey

Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch

***Abstract*—The performance of face analysis and recognition systems depends on the quality of the acquired face data, which is influenced by numerous factors. Automatically assessing the quality of face data in terms of biometric utility can thus be useful to filter out low quality data. This survey provides an overview of the face quality assessment literature in the framework of face biometrics, with a focus on face recognition based on visible wavelength face images as opposed to e.g. depth or infrared quality assessment. A trend towards deep learning based methods is observed, including notable conceptual differences among the recent approaches. Besides image selection, face image quality assessment can also be used in a variety of other application scenarios, which are discussed herein. Open issues and challenges are pointed out, i.a. highlighting the importance of comparability for algorithm evaluations, and the challenge for future work to create deep learning approaches that are interpretable in addition to providing accurate utility predictions.**

***Index Terms*—Biometrics, biometric sample quality, face quality assessment, face recognition.**

I. Introduction

Face Quality Assessment (FQA) refers to the process of taking face data as input to produce some form of "quality" estimate as output, as illustrated in Figure 1. An FQA algorithm (FQAA [57]) is an automated FQA approach. FQA can consist of general Image Quality Assessment (IQA), but it is typically specialized to faces (e.g. by utilizing the position of the eyes), and thus unlikely to be applicable as general IQA.

This survey focuses on face images in the visible spectrum as input to the face processing pipeline, which represents the most common input to face recognition (FR) systems, as opposed to face images beyond the visible spectrum [58][59]. Also, only single-image input FQA approaches are considered, meaning that methods utilizing additional subject-dependent input, such as reference [60] images [61], are outside the survey's scope. Thus, unless otherwise specified, FQA(A) will refer to single-image Face image Quality Assessment Algorithms in the visible spectrum, with a Quality Score (QS [57]) output that can be represented by a single scalar value. A vector of quality values measuring different aspects such as sharpness or illumination can be generated as well. See Figure 2 for some example images that a FQAA would likely consider to be of lower quality for different reasons.

T. Schlett, C. Rathgeb and C. Busch are with the da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Germany, {torsten.schlett, christian.rathgeb, christoph.busch}@h-da.de

O. Henniger is with the Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany, olaf.henniger@igd.fraunhofer.de

J. Galbally is with the European Commission, Joint Research Center, Ispra, Italy, javier.galbally@ec.europa.eu

J. Fierrez is with the Universidad Autonoma de Madrid, Madrid, Spain, julian.fierrez@uam.es

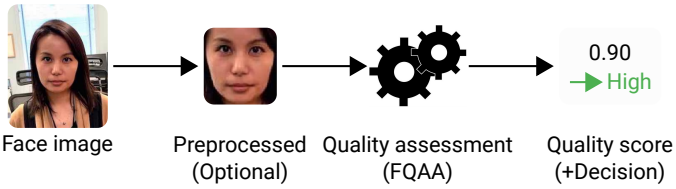

Fig. 1. Typical FQA process: A face image is preprocessed and an FQAA is applied, resulting in a scalar quality score output, based on which a decision can be made. Face image taken from [62].

The term "quality" could refer to various definitions, with ISO/IEC 29794-1 [63] differentiating between three aspects called character, fidelity, and utility. In the context of face biometrics these can be described as follows [64]:

- **Character**: Attributes associated with a biometric characteristic (e.g. the face topography or skin texture) that cannot be controlled during the biometric acquisition process (e.g. scars) [60].
- **Fidelity**: For a biometric sample [60], e.g. a face image, fidelity reflects the degree of similarity to its source biometric characteristic [63]. Thus a blurred image of a face omits detail and has low fidelity [62].
- **Utility**: The fitness of a sample to accomplish or fulfill the biometric function, which is influenced i.a. by the character and fidelity [60]. Thus, the term utility is used to indicate the value of an image to a receiving algorithm [62].

This survey considers the "utility" as the primary definition of what a quality score should convey, which is in accordance to i.a. the quality score definition of ISO/IEC 2382-37 [60] and the definition in the ongoing Face Recognition Vendor Test (FRVT) for face image quality assessment [62]. Thus, a Quality Score (QS) should be indicative of the Face Recognition (FR) performance. Note that this entails that the output of a specific FQAA may be more accurate for a specific FR system, so the FQA utility prediction effectivity ultimately depends on the combination of both. To facilitate interoperability, it is however desirable that the FQAA is predictive of recognition performance in general for a range of relevant systems, instead of being dependent on a single FR technology.

In short, under this survey's definitions, an FQAA is typically meant to output a scalar quality score to predict the FR performance from a single face input image. Being able to predict FR performance without necessarily running an FR algorithm makes FQA useful for a variety of scenarios, which are described further in subsection II-A. Note that the focus on FQA as a predictor for FR performance in the present survey

Fig. 2. Face images of a single subject with various qualities. Face quality degrade from left to right as quality degradation factors such as facial expression, pose, and illumination are introduced. Face images taken from [62].

is because this aspect has attracted the predominant interest from researchers so far. Other approaches to FQA such as the prediction of utility in face biometrics tasks like emotion analysis [65], attention level estimation [66], gender or other soft biometrics recognition [67], etc. may open interesting research lines in the future and can take advantage of current developments that employ FQA as FR performance predictor.

This survey revisits more than 50 FQAA publications from 2004 to 2020. These describe varying numbers of prior work themselves, with Hernandez-Ortega *et al.* [42] being a recent example that contains a summary for some prior publications ranging from 2006 to 2020. A fingerprint/iris/face quality assessment survey by Bharadwaj *et al.* [68] considered less than ten FQAA publications from 2005 to 2011. The European JRC-34751 [69] report listed some FQAAs, including general IQA and image-set FQAA [70] (i.e. multiple images as input for the FQAA) that are not considered here.

The remainder of this survey is organized as follows: Section II describes the FR context and a variety of scenarios for which FQA can be beneficial, as well as methods to evaluate the predictive performance of the FQA approaches themselves. The FQAA literature is surveyed in section III and section IV highlights associated open issues as well as challenges for future work. Section V provides a summary to conclude the survey. The survey does not have to be read in order and the chronologically sorted literature tables can be referred to for a quick overview. Abbreviations are defined on their first occurrence.

II. Quality Assessment in Face Recognition

During enrolment, a classical face recognition system acquires a reference face image from an individual, proceeds to detect and pre-process it, and finally extracts a set of features which are stored as reference template. At the time of authentication a probe face image is captured and processed in the same way and compared against a reference template of a claimed identity (verification) or up to all stored reference templates (identification). Refer to ISO/IEC 2382-37 [60] for the standardized vocabulary definitions of terms such as enrolment, templates or references.

Regarding the face image acquisition [60] one can differentiate between two acquisition scenarios for face images [69]:

- **Controlled**: In a controlled scenario, the biometric capture subject is cooperative [60], so that e.g. the head pose (see Figure 3) is adjusted to frontally face the camera with a neutral expression, and the environmental conditions such as lighting can be controlled. This is typically the case when face images are acquired for government-issued ID documents.
- **Unconstrained**: Here the capture subject is not cooperative, i.e. the subject is either indifferent [60] or intentionally uncooperative [60], and there is no control over the environmental conditions. Surveillance video FR is an example for this scenario [72].

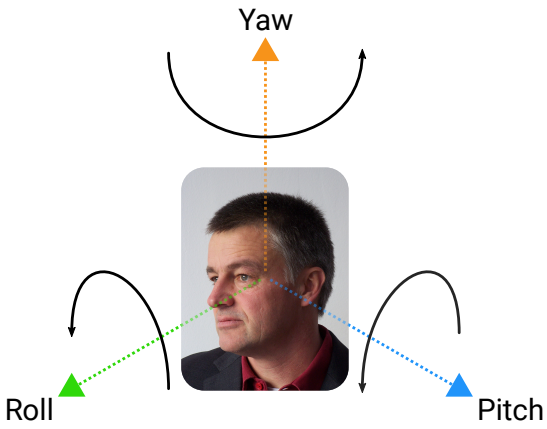

Fig. 3. Facial pose is usually defined by the pitch, yaw, and roll angles defined by ISO/IEC 39794-5 [71]. A frontal face has 0° for all three angles.

There are other scenarios in between those two extremes, e.g. smartphone FR with a cooperative capture subject but incomplete control over the environment [69], and the literature usually refers to close-to-optimal capture conditions as "controlled", with anything else falling under the "unconstrained" category [69]. FQA can be used during controlled acquisition to ensure a certain level of quality by providing immediate feedback. For unconstrained acquisition e.g. via video cameras, FQA can be used to filter out images below a certain quality level. While the same FQAA type and configuration could be used for both, stricter requirements that are desirable for a controlled government ID image acquisition scenario may be too strict for unconstrained scenarios. To facilitate helpful feedback, FQA for the controlled scenario preferably should also be able to provide an explanation in terms of multiple separate human-understandable factors, such as the pose angles (see Figure 3) or the illumination direction. In contrast, FQA for the fully unconstrained scenario by definition cannot benefit from explainability during the acquisition process since there is no control, e.g. when automatically deciding whether a video frame is processed further or not. Nevertheless, explainable FQA can still be useful to analyse collected images.

Furthermore, for quality assessment a distinction can be drawn between approaches that require a "reference" version of the input and those that do not [68][51][42] (not to be confused with biometric references [60] e.g. in a FR database):

- **Full-reference**: FQA that compares the input image against a known reference version thereof, i.e. a version that is known to be of higher or equal quality. Conversely, the input image can be seen as a potentially degraded (e.g.

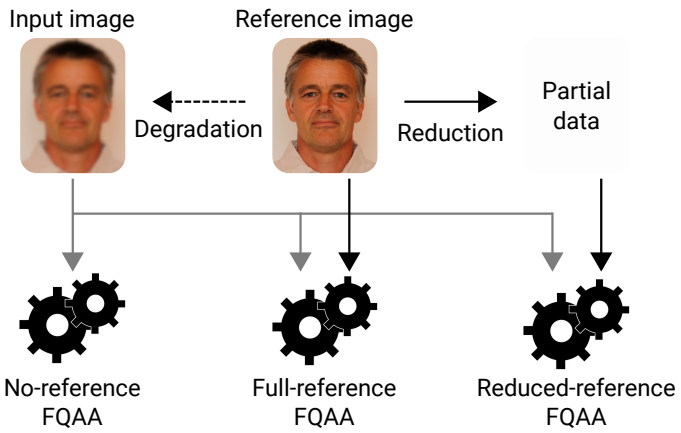

Fig. 4. Full-reference, reduced/partial-reference, and no-reference FQA approaches differ in the used input data, as described in section II.

blurred) version of the reference image.

- **Reduced-reference/Partial-reference**: Similar to full-reference FQA, a reference version of the input image has to exist first, but only incomplete information of the reference is known and used for the FQA.
- **No-reference**: No reference version of the input image is required for the FQA. Note that a FQAA can still use other kinds of images or other data for reference: A FQAA could e.g. utilize some fixed set of images unrelated to the input image and still be categorized as no-reference FQA. Likewise, machine learning FQA models are not automatically classified as reduced-reference FQA just because they incorporate information from training images.

See Figure 4 for an illustration of the three concepts. As stated in section I, this survey focuses on single-image FQA, so only no-reference FQA literature is listed herein. Single-image FQA input additionally implies that no further data specific to the corresponding person (or biometric capture subject [60]) is required to facilitate the FQA.

A. Application areas of FQA

FQA can be applied to a variety of application areas. Apart from further possible application areas, the most relevant ones, which have partly been investigated in the scientific literature, are:

- **Acquisition process threshold**: Face images that result in a quality score below a set threshold can be rejected during the acquisition process [60]. Besides assessing image data stemming directly from cameras, FQA could also be applied to measure the impact of printing and scanning, but among the surveyed literature this was only evaluated indirectly in one work by Liao *et al.* [8].
- **Acquisition process feedback**: One or multiple FQAAs may not only be used for image rejection, but also to provide feedback to assist the FR system operator. E.g. individual requirements from ISO/IEC 39794-5 [71], ICAO [73][74], or ISO/IEC 19794-5 [75] can be checked and reported automatically when an image is acquired for FR system enrolment [60], or for passports and other government-issued ID documents. See subsection III-B for corresponding literature. Capture subjects [60] themselves can also receive immediate feedback for possibly less rigid requirements, e.g. during ABC (Automatic Border Control) at airports. Quality can also be monitored in relation to acquisition circumstances, such as the used capture devices [60] or locations [63], e.g. to identify comparatively underperforming capture devices over time.
- **Video frame selection**: Images in a video sequence can be ranked and selected by their assigned quality scores. Literature that specifically considers FQA for this purpose is listed in subsection III-C. This can be used e.g. to improve both computational performance and recognition performance for identification via video-surveillance.
- **Conditional enhancement**: Optional image enhancement could be applied to images within a certain quality range: Images with sufficiently high quality may not require enhancement, images with very low quality may not be salvageable by enhancement, and images within a quality range may be adequate for enhancement. In addition, multiple enhancement passes could be applied depending on the change in quality after each application, and different enhancement configurations may be selected for different quality aspects. While image enhancement could be applied to every image unconditionally, this could technically degrade/falsify otherwise high quality images, and involve a significant computational overhead that could make additional hardware necessary (e.g. GPUs). The former drawback was shown e.g. for illumination FQA by Rizo-Rodriguez *et al.* [11]. And the FQA application list of Hernandez-Ortega *et al.* [42] noted [76] and [77] as examples for the latter drawback, with [76] listing multiple methods taking seconds to minutes, while [77] states a requirement of 30ms per single image using a GPU. Furthermore, multiple images can be selected by quality as a collective basis to construct an improved image - this was done e.g. in an enhancement approach stage of the video-focused method [37] by Nasrollahi and Moeslund. Lastly, it is also possible to enhance image regions individually depending on region-specific quality scores, which was done e.g. in one approach of Sellahewa and Jassim [14].
- **Compression control**: The change in quality can be measured when an image is compressed in a lossy fashion. Analogous to conditional enhancement, this measurement can further be used to control the compression, e.g. by iteratively adjusting the overall compression factor. Besides the FQAA literature listed in this survey, it is also possible to employ full/reduced-reference FQA/IQA for this use case, since a reference is available in form of the compression input image.
- **Database maintenance**: Existing images in a database can be ranked and filtered by quality. This means that the image with the highest quality can be selected per subject, and that a FR system operator can be notified automatically if a subject has no image of sufficient quality. In systems that do not store images to preserve privacy or

storage space, any image-based FQAA of course needs to be applied beforehand to obtain a QS. Furthermore, images or templates [60] in the database can be updated in a controlled manner, by comparing the associated QS to the QS of a new image/template. This could be done automatically e.g. after a successful verification. Hernandez-Ortega *et al.* [42] noted that such updates may also consist out of incremental improvements [78][79], instead of replacements. Besides subject-specific incremental improvements, new quality-controlled data can also be employed to improve biometric models via online learning [80][68].

- **Context switching** [68][42]: A recognition system can adapt to different quality aspects by switching between multiple recognition algorithm configurations (or modes [60]), using quality assessment for the switch activation [81]. This does not have to be a pure FR system, i.e. this can apply to a multi-modal [60] biometric system.
- **Quality-weighted fusion** [68][42]: Similar to full context switching, a potentially multi-modal system can fuse scores or decisions in a weighted fashion based on quality assessments [82][83]. Quality-based feature-level fusion for face video frames is considered e.g. in the surveyed literature [33] and [49]. Quality scores themselves can naturally also be fused, e.g. to produce a single scalar QS result from multiple different FQAAs.
- **Comparison improvement**: Quality can be used directly as part of FR comparisons [60]. For example, Shi and Jain [49] computed quality in terms of uncertainty for each FR feature dimension and incorporated it in their comparison algorithm.
- **Face detection filter**: In more general terms than video frame selection, FQA could inherently be used to increase the robustness of face detection by ignoring candidate areas in an image with especially low quality. This kind of application is however only indirectly examined through the video frame selection works among the surveyed literature. Conversely, the confidence of face detectors themselves can be utilized as a kind of FQA, which was done e.g. by Damer *et al.* [33].
- **Partial presentation attack avoidance**: Although the surveyed literature doesn't focus on this application, rejecting or weighing images based on their assessed quality can inherently also reduce the opportunities for presentation attacks [60][84], since accepting images for enrolment or as probe irrespective of their quality could be an attack vector. FQA or IQA can also be employed specifically for the purpose of PAD (Presentation Attack Detection) [85]. Pure FQA is however not meant for comprehensive PAD, because such attacks can consist of data with high utility [63] too.
- **Progressive identification**: An identification search process could progress from templates with the highest associated quality to templates with the lowest quality in the database. Assuming that these templates vary noticeably in quality and that the search requires an extensive amount of time, this may help by showing results with higher confidence (due to higher qualities) early on in the search process. This could also be used to stop a search early, i.e. once an amount of results with acceptable certainty has been found. However, a sufficiently fast identification over the entire database makes such considerations irrelevant, and this approach is presumably not as useful as general computational workload reductions surveyed by Drozdowski *et al.* [86], since it relies on the existence of exploitable quality variety in the database. While the listed FQA literature does not explore this approach, it does consider FQA-based computational workload reduction in terms of video frame selection. Instead of progressing from highest to lowest quality, Hernandez-Ortega *et al.* [42] noted that the system could use the quality of the probe image to start with comparisons to templates of similar quality, which may also imply similar acquisition conditions, and thus could improve the accuracy.

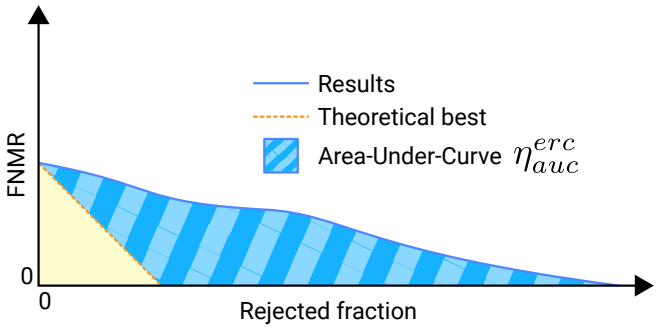

Fig. 5. Error-versus-Reject-Curve (ERC) example. Here FNMR (False Non-Match Rate) is plotted as error, but other types could be plotted instead, e.g. the FMR (False Match Rate). This is a conceptual example without real values or scale. Plots for real FQAAs usually start with a low FNMR and show a zoomed-in plot instead of the full 0 to 1 range.

B. Evaluating FQAA performance

An Error-versus-Reject-Curve (ERC) can be plotted to evaluate the predictive performance of quality assessment algorithms, as proposed by Grother and Tabassi [87]. In the context of FQA, a FR system and a face dataset with subject labels is required in addition to the FQAA to compute the ERC. The FR system compares face image pairs with a fixed comparison threshold [60] to decide between match [60] and non-match [60] for each pair. Quality scores produced by the FQAA per image are combined for the image pairs (e.g. by taking the minimum). A progressively increasing quality threshold is applied to these image pair quality scores, and a FR error measure is calculated for the resulting comparison subsets. In [87], it is suggested that the FNMR (False Non-Match Rate) [60] error measure should be used as the primary performance indicator. In this case the FR threshold can be derived for a fixed FMR (False Match Rate) [60] on the unfiltered image pairs - and vice versa if the FMR were plotted as the error measure. An abstract ERC example is shown in Figure 5. The error is typically plotted on the vertical axis. The rejected fraction, plotted on the horizontal axis, denotes the relative amount of comparisons (0 to 100%) rejected due

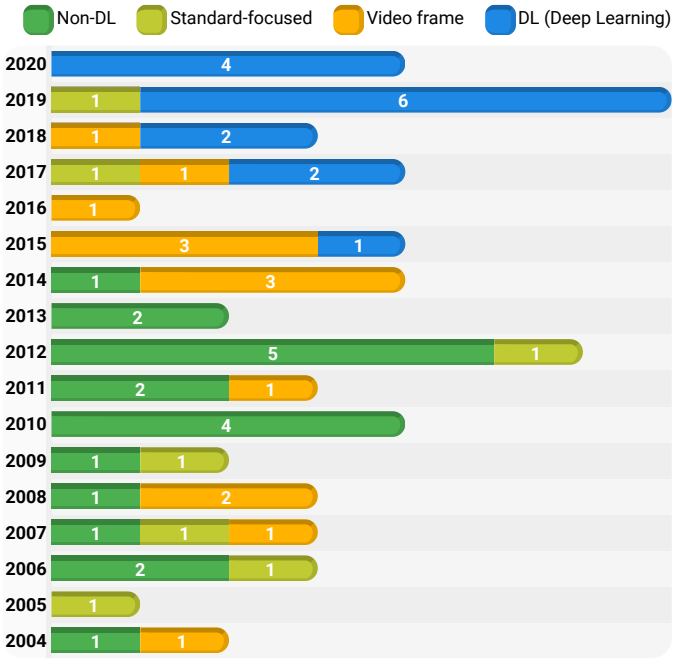

Fig. 6. Timeline of the surveyed FQA literature with categories.

to the quality score. Plotting this fraction instead of e.g. an increasing QS threshold normalizes the axis independently of the given FQAA.

Olsen *et al.* [88] further proposed to compute the scalar Area-Under-Curve (AUC) for some rejection fraction range of an ERC:

$$\int_a^b ERC - \text{area under theoretical best}$$

More precisely, [88] proposed to compute η_{auc}^{erc} for the full $[0, 1]$ range, and η_{pauc20}^{erc} to focus only on the $[0, 20\%]$ range. The "area under theoretical best" term refers to the best case where the error value decrease equals the rejected fraction percentage.

However, the FQAA literature listed in this survey does not necessarily provide AUC or ERC evaluation results. For example, some literature evaluates the FQAA in terms of quality label prediction performance, and doesn't evaluate the FQAA in terms of FR performance improvements. Even if all of the literature had utilized a common evaluation result format, e.g. ERC plots with the same error measures, there would still be differences in the used FR systems and datasets. This issue makes a precise performance comparison based solely on reported results impossible. Refer to section IV for further discussions regarding this and other issues. Nevertheless, various works include multiple FQAAs in their evaluations, and clear conceptual FQA differences can be observed in the literature. Furthermore, there is the ongoing Face Recognition Vendor Test (FRVT) for face image quality assessment [62], which evaluates FQAAs combined with a number of FR algorithms and dataset types, showing results i.a. in the form of ERC plots.

III. Face Quality Assessment Algorithms

The surveyed FQA literature has been subdivided into four categories to provide a clearer overview:

- **Non-DL** (subsection III-A): Literature that does not use deep learning (DL) to implement FQAAs. This encompasses the majority of the literature.
- **Standard-focused** (subsection III-B): Literature that considers FQA for standard-defined quality factors. In comparison to the other categories, most of this literature incorporates especially large amounts of separate human-understandable factors. There are no deep learning approaches within this specialized category.
- **Video frame** (subsection III-C): Literature that considers FQA for the purpose of selecting video frames. There are only two deep learning approaches within this specialized category.
- **DL** (subsection III-D): Literature that uses deep learning (DL) to implement FQAAs. This represents the most recent literature.

Almost all of the literature in the two specialized categories (standard-focused and video frame literature) is non-DL literature, with two of the video frame works counting as DL literature, but the tables and subsections do not contain any literature overlap for the sake of clarity. Note that non-DL literature does encompass FQA approaches based on other kinds of machine learning (including shallow artificial neural networks), in addition to purely hand-crafted methods. The deep learning FQA literature emerged more recently, and the trend towards DL-based FQA research can be observed in Figure 6.

Some of the works share at least one author with a different surveyed FQA work, and the groups formed via such author relationships include up to four works each. Most of the literature however names a unique set of authors, indicating that the FQA research is driven by a large variety of research groups. Independently of author relationships, various FQA works are clearly based on prior work, which is noted both in the text and the overview tables of the following subsections.

Table V lists the datasets used to develop and evaluate the FQA approaches of the surveyed literature. The implications of the dataset variety are discussed in subsection IV-A.

A. Non-DL FQA literature

The clear majority of the surveyed FQA literature consists of non-DL (non-Deep-Learning) FQAA literature, i.e. literature that uses hand-crafted algorithms or other machine learning types for FQA (Table I). The standard-focused (Table II) and video frame FQA (Table III) literature listings also mostly consist of non-DL approaches. In the present subsection we survey non-DL FQAA literature in chronological order, see Table I from bottom to top.

Luo [20] considered general IQA related to brightness, blur, and noise in the context of face images. 10 features were extracted from a grayscale image and passed to a RBF (Radial Basis Function) ANN (Artificial Neural Network) to produce the final quality score. As an alternative to the ANN, a GMM (Gaussian Mixture Model) was used as well, but reportedly

TABLE I
MOST RELEVANT WORKS ON NON-DL FQA.

Reference	Year	Method(s)	Datasets	Miscellaneous
Abaza *et al.* [1]	2014	ANN on 5 factors/7 measures equivalent to [7] vs. logistic regression, SVR, and 10 normalization/fusion combinations	CAS-PEAL, Yale, GBU, FERET, MBGC, Q-FIRE	Continuation of [7].
Phillips *et al.* [2]	2013	9 FQAA, i.a. Illumination (Direction), SEMC [12], Edge density [13], ..., and SVM vs. GPO oracle	Unknown, GBU, PaSC	Continuation of [12].
Bharadwaj *et al.* [3]	2013	4-class SVM on Gist[89] or HOG	SCface, CAS-PEAL	–
Qu *et al.* [4]	2012	Illumination (Gaussian low-pass filter vs. fixed 38-image-average reference)	Extended Yale	–
Klare and Jain [5]	2012	Impostor-based Uniqueness Measure	In-house (Police)	–
Hua *et al.* [6]	2012	Blur (MTF vs.: ED [90], LoG, SG, DCT)	Q-FIRE	–
Abaza *et al.* [7]	2012	12 measures: Sharpness $\times 4$, Contrast $\times 2$, Illumination $\times 2$, Focus $\times 2$, Brightness $\times 2$	CAS-PEAL, Yale, GBU, FERET, MBGC	Proposes combined FQAA with 7 factors.
Liao *et al.* [8]	2012	5-class cascade SVM with Gabor magnitude features	In-house	–
Wong *et al.* [9]	2011	Per block low-frequency 2D DCT components compared to "ideal" frontal face	FERET, CMU-PIE, ChokePoint	–
De Marsico *et al.* [10]	2011	Landmark-based: Pose (Yaw/pitch/roll), Illumination (Histogram mass center variance), Symmetry (Lines)	FERET, LFW, SCface	–
Rizo-Rodriguez *et al.* [11]	2010	Illumination of triangle mesh regions (Mean, ANN-weighted, Combined)	Extended Yale, XM2VTS	Mentions INCITS FR Format standard.
Beveridge *et al.* [12]	2010	Illumination (Direction), SEMC (Strong Edge and Motion Compensated focus measure), Edge density	FRVT 2006, CMU-PIE	Continuation of [13].
Beveridge *et al.* [13]	2010	Region density, Edge density, Eye distance	FRVT 2006	Continuation of [16].
Sellahewa and Jassim [14]	2010	Illumination (Luminance distortion from [91])	Extended Yale, AT&T	–
Zhang and Wang [15]	2009	Symmetry (3 variations based on SIFT [92])	CAS-PEAL	–
Beveridge *et al.* [16]	2008	Edge density, Eye distance	FRVT 2006	–
Abdel-Mottaleb and Mahoor [17]	2007	4 measures: Blur (Frequency kurtosis), Illumination (Weighted sum), Pose (Yaw), Expression (GMM)	FERET, WVU, Cohn-Kanade	–
Kryszczuk and Drygajlo [18]	2006	Same as [19], plus another score-level measure	BANCA	Continuation of [19].
Kryszczuk and Drygajlo [19]	2006	Average face image correlation, Blur, Classification score sum of log-likelihoods	BANCA	–
Luo [20]	2004	RBF-ANN on: Brightness, Spectrum $\times 7$, Noise $\times 2$	Unspecified (850 images)	–

resulted in worse performance. The IQA was trained with and compared against the quality estimates of a single human on an unspecified dataset. The 10 features consist of 1 measure for average pixel brightness, 7 values derived from the sub-bands of two-level wavelet decomposition, and 2 different noise measures (one based on a square window with minimum grayscale pixel value standard deviation, and one combining the standard deviation of square windows in binarized versions of the high-frequency sub-bands).

Two image-based ("signal-level") and one classification-score-based ("score-level") FQAA were used in the approach of Kryszczuk and Drygajlo [19], all of which were combined into a binary decision by means of two GMMs (for "correct" and "erroneous" classifier decisions) with 12 Gaussian components each. The authors also added another score-level measure to the approach in [18]. But the inclusion of a classification-score-based FQAA means that the combined FQAA can only be used after a FR comparison has taken place, so this component would have to be excluded to allow isolated single-image FQA using the remaining two image-based FQAAs. Of these, one measures sharpness as the mean of horizontal/vertical pixel intensity differences (corresponding to high-frequency features), and the other computes Pearson's cross-correlation coefficient between the face image and an average face image (corresponding to low-frequency features). This average face image is the average of the first eight PCA (Principal Component Analysis) eigenfaces for a given training image set.

Abdel-Mottaleb and Mahoor [17] proposed FQAAs to assess blur, lighting, pose, and facial expression. Blur is measured as the kurtosis in the frequency domain. The lighting QS is formed by a weighted sum of the mean intensity values for 16 weight-defined regions (used to focus more on the center of the image). Pose is estimated as the yaw angle (see Figure 3), derived by comparing the amount of skin tone pixels between the left/right-side triangle, which are defined by the three center points of the eyes and the mouth. Fisher Discriminant Analysis (FDA) is employed to differentiate skin pixels from other regions. To assess whether the expression is good or bad in terms of quality, a GMM is trained based on the correct/incorrect decisions of an FR algorithm for a labeled facial expression dataset.

Beveridge *et al.* examined the impact of a number of factors on FR verification performance in [16] and [13] using GLMMs (Generalized Linear Mixed Models). Taking preexisting FQA-input labels such as age or gender out of consideration, three described measurements are considered for automatic image-only FQA, one of which is the image resolution/eye distance. Two more complex measurements remain, with [16] introducing an edge density metric consisting of the averaged Sobel filter pixel magnitude, and [13] adding a region density metric that segments the face and counts the distinct regions. Both of these metrics were applied on grayscale images, with the face area being masked by an ellipse to reduce the metrics' sensitivity to environmental factors in the rest of the image. The authors continued in [12] by comparing their edge density

metric to two newly introduced FQAAs. One is the Strong Edge and Motion Compensated focus measure (SEMC), a successor to the edge density metric that is computed based on the strongest edges in the face region (instead of all), which was intended to correlate more clearly to focus/blur in images (instead of also being affected by other factors such as illumination). The second new FQAA estimates to which degree a face is lit from the front (positive number output) or the side (negative number). Experiments in [12] used GLMMs and FRVT 2006 test data/FR algorithms similar to [16] and [13], and found that the illumination measure subsumes both the edge density and the SEMC measure regarding FR performance prediction. These measures were studied further in [2], as described below.

Zhang and Wang [15] proposed three symmetry measure variations based on SIFT [92] (Scale Invariant Feature Transform). The first variation counts the number of SIFT points in the left and right half of the image, and divides the minimum of the two numbers by their maximum to obtain the QS. Using the fixed left/right image halves entails that this measure is intended for frontal face images. The second QS variation is formed by the amount of SIFT points that have a mated point in the other half based on their location. And the third variation further adds a Euclidean distance comparison of the SIFT feature vectors to define corresponding points, using a horizontally flipped version of the image to establish target points with directly comparable SIFT features. As part of the evaluation, the first and simplest variant was shown to have the highest correlation with Eigenface- and LBP-based FR comparison scores.

An IQA approach for frontal illumination was presented by Sellahewa and Jassim [14], using the luminance distortion component from the "universal image quality index" [91] to compare a face input image against a fixed average reference image generated from a training set. This is done by sliding a window (with 8×8 pixel size in [14]) simultaneously over the input and reference image, computing $2L_{\text{input}}L_{\text{reference}}/(L^2_{\text{input}} + L^2_{\text{reference}})$ therein with L being the mean luminance, and using the mean of all window results as the final $[0, 1]$ QS.

Rizo-Rodriguez *et al.* [11] presented a frontal illumination assessment method more specific to faces than e.g. the luminance distortion in [14]. First, a triangular mesh is fitted to the face in the input image. Then the mean luminance is computed for each of the triangle regions, forming a histogram of mean luminance values per face, which was observed to approximate a normal distribution in face images with homogeneous frontal illumination. This was used to derive a binary QS using an experimentally obtained threshold. To additionally account for differences in importance between the regions, a three layer perceptron was trained for important regions only - i.e. input neurons for 24 triangles in the vicinity of the nose. A binary QS was obtained from this ANN as well, and both of the QS decisions were optionally combined.

De Marsico *et al.* [10] proposed landmark-based measures for pose, illumination, and symmetry. For pose, the yaw/pitch/roll angles are assessed using landmarks for the eye centers, the tip and root of the nose, and the chin. A weighted sum of the three $[0, 1]$ angle QSs forms the pose QS, whereby the weights were derived experimentally as yaw 0.6, pitch 0.3, and roll 0.1. Illumination is measured by applying a sigmoid function to the variance of the mass centers for 8 gray level histograms, which are computed for areas around 8 landmarks (3 on the nasal ridge, 2 on each cheek, 1 on the chin). Symmetry is measured by comparing the grayscale values of point pairs sampled along 8 lines defined by landmark pairs on each side of the face. All three measures result in a $[0, 1]$ scalar QS. They are not fused, but it was noted that the symmetry measure inherently takes both pose and illumination into account. The evaluations demonstrated i.a. that the FR performance improvement capabilities of the measures differ depending on the used FR algorithm.

Wong *et al.* [9] presented a FQAA for frontal face images that is more general than e.g. the frontal pose illumination FQAAs of [11] or [14]. Low-frequency 2D DCT (Discrete Cosine Transform) components are extracted for overlapping blocks of a normalized grayscale face image. Per block, these are compared against Gaussian distributions derived from a set of training images with frontal illumination, and a final QS is formed by fusing the resulting probabilities.

Liao *et al.* [8] trained an SVM (Support Vector Machine) cascade to predict subjective QS labels using Gabor filter magnitude values as features. The SVM cascade has four stages, each being a binary classifier, so that the approach predicts integer QS levels from 1 to 5 (e.g. the first SVM decides whether the QS is 1, or whether it might be higher). In addition, two of such SVM cascades were used for two different image crop sizes, and their output QSs were fused by taking the mean. Training and evaluation used partitions of a dataset with 22,720 grayscale images, all with subjective ground truth QS labels (1 to 5; 1 being the best quality). The evaluation showed that the fusion approach provided the best predictive performance overall.

Multiple IQA methods were examined for FQA by Abaza *et al.* in [7], and later [1], i.a. incorporating synthetic image degradations regarding contrast, brightness, and blurriness for the evaluations. Of the 12 tested individual measures in [7], 7 were retained to represent 5 input factors for a combined single-image FQAA, using Gaussian models for normalization and the geometric mean for fusion. Contrast was measured as the RMS (Root Mean Square) of image intensity, brightness as the average HSB (i.e. HSV, Hue Saturation Value/Brightness) color space brightness (computable as the maximum of the normalized red/green/blue channel value per pixel [1]), focus as the mean of the image gradient's L_1-norm and the Laplacian energy [93], sharpness as the mean of the two average gradient measures [18] and [25], and illumination using the weighted sum technique proposed by [17]. The 5 measures that weren't used for the combined FQAA comprise the Michelson contrast measure [94], the brightness measure from [95], the Tenengrad sharpness measure plus an adaptive variant from [96], and the luminance distortion [91] measure previously seen in [14]. Note that according to both [7] and [1] the selected brightness measurement was chosen due to its reduced computational workload in comparison to the other tested method (which achieved better predictive performance), but this drawback

might not be relevant anymore due to processor hardware improvements since 2012/2014. Continuing with [1], the same 5 factors based on the chosen 7 (of 12) measures were presented as in [7], but now an ANN was trained to combine the 5 factors without any prior normalization to produce a binary QS classification. A single-layer ANN with six neurons was found to provide the best classification results among 10 different ANNs with either 1 or 2 layers (and 4 to 20 neurons per layer), logistic regression, SVR (Support Vector Regression), as well as 10 combination approaches formed from a normalization ($\times 2$, linear or Gaussian model) and a fusion ($\times 5$) part, including the previous method from [7]. However, the tested methods/ANNs' 5-factor input vector apparently was the per-element minimum of the vectors for both a probe and a gallery image, so here the probe image isn't used in isolation.

To measure blur in face images, Hua *et al.* [6] proposed using the Modulation Transfer Function (MTF), and evaluated this approach together with various other blur related measures: A measure based on the radial spatial frequencies of 2D DCT coefficients, a Squared Gradient (SG in Table I) metric that consists of the gradient image (edge) magnitudes, and a Laplacian of Gaussian (LoG) method. There also was an Edge Density (ED) measure, which was formed by subtracting the 3×3 mean filtered image from the original, then taking the average of the result's absolute pixel values [90]. This measure also occurs in [39] under subsection III-C (not to be confused with the previously mentioned Sobel filter edge density from [16] and [13]). The correlation of these measures (applied to a face image) to a ground truth MTF applied to an optical chart is assessed, with the face image MTF showing the highest, and edge density the lowest average correlation, the other mentioned measures having high correlation closer to the MTF result.

An approach that is inherently adaptive to any used FR system was presented by Klare and Jain [5], which computes a QS called impostor-based uniqueness measure (IUM) for a face image by comparing it against a given set of "impostor" face images/feature vectors via the FR system itself. Based on experiments, [5] proposed to use 1,000 feature vectors from different subjects to form this set. Note that the paper appears to only utilize frontal face images (from an operational police dataset).

Qu *et al.* [4] proposed an illumination FQAA based on Gaussian blur. The Gaussian blur is applied to the input image, which is then compared against a reference image formed by the average of 38 training images. This comparison is calculated as the normalized correlation. The paper evaluated a range of sizes for the Gaussian blur. FR performance wasn't evaluated, but an evaluation can be found as part of the illumination methods considered in [30].

Bharadwaj *et al.* [3] trained a one-vs-all SVM for 4 quality bins using either sparsely pooled Histogram of Oriented Gradient (HOG) or Gist [89] input features. The quality bin training labels were obtained using two COTS (Commercial Off-The-Shelf) FR systems on training images that have a single designated good/studio quality image in addition to several probe images per subject.

Phillips *et al.* [2] examined 13 quality measures, including the edge density metric from [16] and [13], plus the SEMC measure from [12] (all four of these papers share authors). There also is an "illumination direction" measure that might correspond to [12] too, but this wasn't clarified. Similar to the two prior papers [16] and [13], the 13 quality measures in [2] contain preexisting labels from EXIF (Exchangeable Image File Format) metadata, e.g. exposure time, leaving 9 measures that can clearly consist of FQA approaches which use the actual image (pixel) data: Edge density [16], SEMC [12], illumination direction (possibly [12]), left-right side illumination histogram comparison, eye distance, face saturation (the number of face pixels holding the maximum intensity value), pixel standard deviation, mean ratio (mean pixel value of the face region compared to the entire image), and pose (yaw angle, 0 being frontal). The 13th quality measure is an SVM that summarizes the other 12 measures. Pruning via the 13 measures was compared against a Greedy Pruned Order (GPO) oracle that discards images in an approximately optimal fashion to improve FR performance, thus representing an upper bound for FR performance improvements enabled by some FQAA. Experimental results indicated a substantial gap between the oracle and the 13 quality measures, with various measures such as the illumination direction additionally leading to worse FAR (False Acceptance Rate) results. Furthermore, another FQAA using PCA followed by LDA (Linear Discriminant Analysis) was trained, but it was observed to generalize poorly to the test set.

B. Standard-focused FQA literature

The works listed in Table II specifically considered FQA in relation to the image requirements given by the standards ISO/IEC 19794-5 [75], ICAO 9303 [74], and ISO/IEC TR 29794-5 [57], which is under development as an International Standard. There also is the more recently established ISO/IEC 39794-5 [71] face image data standard, which is however only cited by Khodabakhsh *et al.* [21] among the literature listed in Table II, since the rest predates this standard. Further standards relevant to FQA are currently in development, namely ISO/IEC 24357 ("Performance evaluation of face image quality algorithms"), ISO/IEC 24358 ("Face-aware capture subsystem specifications"), and the next edition of ISO/IEC TR 29794-5 [57].

As mentioned in section II, face images for government-issued ID documents are usually acquired under controlled conditions with comparatively strict quality requirements for a number of human-understandable factors. ISO/IEC 19794-5 [75] for example defines recommended ranges for the pitch/yaw/roll head pose angles (see Figure 3), for the width to height ratio of the image, or for the face alignment within the image. In the rest of the present subsection we thematically navigate Table II. Refer to this table for an overview of the factors that are considered in the literature.

Subasic *et al.* [27] and Ferrara *et al.* [23] both described FQAA for comparatively many of the standard requirements, whereby [23] is especially noteworthy due to the introduction of the "BioLab-ICAO" framework, which was used among the

TABLE II
MOST RELEVANT WORKS ON STANDARD-FOCUSED FQA.

Reference	Year	Method(s)	Datasets	Miscellaneous
Khodabakhsh *et al.* [21]	2019	8 factors compared to mean scores from 26 humans	In-house (Smartphone)	Continuation of [22].
Wasnik *et al.* [22]	2017	9 factors, combined via random forest classifier: Lighting Symmetry, Pose Symmetry, Image Brightness, Image Contrast, Global Contrast Factor, Exposure, Blur, Sharpness, Edge Density	In-house (Smartphone)	–
Ferrara *et al.* [23]	2012	30 factors: Eye/Face Location Accuracy, Eye Distance, Vertical/Horizontal Position, Head Image Width/Height Ratio, Blurred, Looking Away, Ink Marked/Creased, Unnatural Skin Tone, Too Dark/Light, Washed Out, Pixelation, Hair Across Eyes, Eyes Closed, Varied Background, Roll/Pitch/Yaw Greater Than $8°$, Flash Reflection on Skin, Red Eyes, Shadows Behind Head, Shadows Across Face, Dark Tinted Lenses, Flash Reflection on Lenses, Frames Too Heavy, Frames Covering Eyes, Hat/Cap, Veil Over Face, Mouth Open, Objects Close to Face	BioLab-ICAO	Presents the "BioLab-ICAO" framework.
Sang *et al.* [24]	2009	2 factors: Symmetry (Gabor wavelet), Blur ((I)DCT)	FERET, CMU-PIE	–
Gao *et al.* [25]	2007	6 factors: Lighting + Pose symmetry (LBP), Eye distance, Illumination Strength (Histogram), Contrast (Standard deviation), Blur (Gradient)	Yale	Incorporated into ISO/IEC TR 29794-5:2010 [57].
Hsu *et al.* [26]	2006	27 factors listed, no details; 5-FAR/FRR-point-based QS normalization; $3\times$ QS fusion, i.a. ANN-based	In-house (Passport database), FRGC	–
Subasic *et al.* [27]	2005	17 factors: Image resolution/AR, Blur, Illumination, Color balance, Background uniformity/tone, Shadows, Hot spots, Eyes tilt/position/red/looking away, Head width/height/rotation	Unspecified (189 images)	–

other FQAA literature as part of the training data preparation for FaceQnet v0 [50] & v1 [42].

Similarly, Hsu *et al.* [26] listed a large number of mostly standard-derived face quality measures as well, but only gave a very brief explanation regarding their implementation, with the normalization and fusion thereof being described instead.

The other works [25], [24], [22], [21] each consider fewer quality factors than [27][23][26]. The methods proposed by Gao *et al.* [25] for symmetry, eye distance, illumination, contrast, and blur FQAA have been incorporated into the ISO/IEC JTC 1/SC 37 technical report ISO/IEC TR 29794-5:2010 [57]. Sang *et al.* [24] compared a Gabor-based (illumination/pose-related) symmetry FQAA to [25], in addition to proposing an FQA approach to measure blur via DCT+IDCT (Discrete Cosine Transform + Inverse DCT).

More recent works by Wasnik *et al.* [22] and Khodabakhsh *et al.* [21] are of special interest for smartphone camera images, whereof [21] can be considered as a continuation of [22] that also examines the FQAA in comparison to subjective quality assessments made by 26 human participants, concluding i.a. that the human FQA highly correlates with FR performance, but not with the tested FQAAs, indicating that the tested FQAAs have their limitations.

C. Video frame FQA literature

Table III lists works that specifically considered FQA for the purpose of extracting relevant video frames by employing single-image FQA (with some noted partial exceptions), meaning that these FQA approaches should also be applicable to other scenarios. Most of these FQAAs also belong to the non-DL category, with the only DL-based literature here being [31] and [28]. In the following we navigate Table III in chronological order from bottom to top.

The approach of Yang *et al.* [41] estimated only the left-right/up-down pose angle, without producing any kind of normalized QS other than the binary decision between frontal and non-frontal pose; faces being declared "frontal" when both pose angles have absolute values not higher than $10°$. While pure pose estimation literature is outside the scope of this survey, this paper demonstrated that pose estimation can easily be used in isolation as a kind of FQAA.

The rest of the non-DL video frame papers all utilize multiple factors for their FQA: Pose, blur, illumination, and resolution factors are predominant, whereby the pose estimation approaches differ especially.

Fourney and Laganiere [40] defined a pose QS as linearly degrading from $0°$ to $45°$, anything above $45°$ resulting in a score of 0, a clear contrast to the binary decision in [41]. The pose estimation in [40] also works in a different manner, namely by locating the eye positions in a gradient image, which was noted to be ineffective for faces with glasses or non-upright orientation. Based on this pose estimation data, illumination symmetry FQA was also conducted by comparing normalized histograms of the left/right side of the face, which was done in addition to an assessment of the overall utilization of the available (e.g. 8-bit grayscale) illumination range within the face image. The remaining factors in [40] were unrelated to the pose estimation: A normalized blur/sharpness QS is derived from the frequency domain; the face image resolution/ pixel count is transformed into a normalized QS, with anything at or above 60×60 pixels corresponding to the maximum (a QS of 1); and a "skin content" measure detects whether human skin appears to be present in the image, which is

TABLE III
MOST RELEVANT WORKS ON VIDEO FRAME FQA.

Reference	Year	Method(s)	Datasets	Miscellaneous
Qi *et al.* [28]	2018	CNN with inception module, trained using gallery FR comparison score minima for detected faces	In-house, PaSC, ChokePoint, CMU-FIA	–
Wang [29]	2017	Subjective QS random forest, 7 hand-crafted features	LFW, Honda/UCSD	–
Hu *et al.* [30]	2016	Kernel Partial Least Squares Regression using mean luminance and Laplacian of 10×10 image sub-blocks	CAS-PEAL, FERET, MIT, FEI, AT&T	–
Vignesh *et al.* [31]	2015	CNN, PCA whitened input, QS labels via MSM	ChokePoint	–
Kim *et al.* [32]	2015	AdaBoost on 3 "objective" measures [34] + optional training-set-"relative" measures	FRGC	Continuation of [34].
Damer *et al.* [33]	2015	Entropy, Viola-Jones [97] face detection confidence	YTF	–
Kim *et al.* [34]	2014	Pose/Alignment (Reconstruction), Blur, Illumination	FRGC	–
Raghavendra *et al.* [35]	2014	Two stages: 1. Pose (yaw/roll via eye & nose position), 2. 12 GLCM features [98] fed into a GMM	In-house (ABC), FRGC, AR, AT&T	–
Nikitin *et al.* [36]	2014	Facial symmetry, Illumination, Blur, Resolution	YTF	–
Nasrollahi and Moeslund [37]	2011	Pose (Linear Auto-associative Neural Networks), Illumination, Blur, Resolution	In-house (100 videos), Pointing'04, IIT-NRC	QS relative to face image sequence. Continuation of [39].
Rúa *et al.* [38]	2008	Blur (Sobel & Laplacian), Symmetry (Per-pixel)	BANCA	–
Nasrollahi and Moeslund [39]	2008	Pose (Center of mass distance), Illumination, Blur, Resolution	FRI-CVL, HERMES project	QS relative to face image sequence.
Fourney and Laganiere [40]	2007	Pose (Eye positions via gradient image), Illumination range & symmetry, Blur, Resolution, Skin content	Unspecified (7 videos)	–
Yang *et al.* [41]	2004	Pose (Haar features learned via SquareLev.R)	Unspecified (300 faces)	–

done by determining the percentage of pixels with a hue of $[-30°, +30°]$ and saturation of $[5\%, 95\%]$. The final combined QS of the six factors consisted of a number of satisfied per-factor thresholds, plus a weighted sum of the factor scores to break ties between frames.

Rúa *et al.* [38] proposed three IQA methods in the context of face video frame selection. One method measures symmetry by comparing the image against a horizontally flipped version of itself, calculating the per-pixel difference, meaning that this measure assumes a centered frontal pose. The other two are general blur IQA methods that compute the average value for either the Sobel or the Laplace operator over the entire input image.

For the works [39] and [37] from Nasrollahi and Moeslund, it is important to note that both derived a QS for each of their factors, except resolution, relative to minimum or maximum values for a sequence of face images - so the described approaches aren't directly usable for single-image FQA. We can remedy this obstacle using simple tricks, for example by choosing constant minima/maxima, hence why these works are nevertheless listed in Table III. The first of the two papers, [39], i.a. cited [40] and directly adapted the face image resolution factor, but presented different approaches to measure the other shared factors: The FQAA starts with information gathered as part of the face detection stage, which determines potential facial regions per-pixel by skin tone, applying a cascading classifier thereon to obtain the face image(s) for further steps. Skin tone pixel count percentages were however not used directly for a QS, in contrast to [40] and [23]. Instead, i.a. the facial center of mass is derived from this per-pixel segmentation. The paper noted that estimating the pose cannot be reliable when using facial features (such as the eyes in [40]), since they may not be visible for sufficiently large angles of rotation, or can be occluded by e.g. glasses. Therefore the difference between the facial center of mass and the center of the face image was used, a method diverging from the previously mentioned approaches that estimate specific angles. Illumination was measured as the average pixel brightness over the face image (against the maximum value for a face image sequence; but here a simple normalization could be applied instead for single-image FQA). Sharpness/blur was assessed using the approach presented in [90], i.e. by first subtracting a low-pass (3×3 mean filtered) version of the face image from the original per-pixel, then averaging the absolute values of all these pixel differences. The FQAA that is part of [37] can be seen as a continuation of [39], with the sharpness, brightness, and resolution measures being almost identical. Brightness is now more clearly defined as the Y component of the YCbCr color space and the resolution QS is now unbounded (i.e. completely relative to an image sequence). What does change is the pose estimation, stating that the prior center of mass approach in [39] tends to be sensitive to environmental conditions. The new approach estimated actual angles and is adapted from [99], using one auto-associative memory (an ANN without hidden layers) per detectable pose.

In the single-image FQAA of Nikitin *et al.* [36] the resolution and illumination measurement, as well as the fusion to combine the factor-QSs, doesn't differ much from what has been mentioned so far (resolution QS relative to constants, illumination dynamic range usage QS, fusion via weighted sum). But here facial landmarks were detected to measure symmetry by comparing the left/right landmark-local gradient histograms, and to measure sharpness via averaged Laplace operator values only within the landmark-defined facial area.

A two stage approach was proposed for - and evaluated with - an ABC (Automatic Border Control) system by Raghavendra *et al.* [35], with the first stage consisting of a yaw/roll angle pose estimation based on the eye and nose position. The final QS was represented by three bins, poor/fair/good, and if the pose isn't detected as frontal, the overall FQAA stops,

assigning the image to the poor QS bin. If the pose is detected as frontal, the second stage decides between the fair/good QS bin assignment. It consists of computing 12 GLCM (Gray Level Co-occurrence Matrix) features [98], which are further processed by a GMM (Gaussian Mixture Model) trained on public non-ABC datasets, and the output thereof was used to obtain the final binary QS bin decision via a threshold.

Of the two works [34] and [32] from Kim *et al.*, [32] is included here in Table III (video frame) instead of Table I (non-DL) because it is a close continuation of [34], even though [32] by itself is arguably less focused on video frame selection. The approach of [34] begins by employing (frontal) face reconstruction to assess pose/alignment quality as the difference between the original and the reconstructed face image; then in stage two blur is measured as the kurtosis of the CDF (Cumulative Distribution Function) of the DFT (Discrete Fourier Transform) magnitude; and the last stage assesses brightness by comparing the histogram for the face image against a given reference histogram, whereby the latter is simply chosen to be the uniform histogram. Each of these three stages ends by comparing the error value result against a predefined threshold, aborting the overall FQAA if the threshold is exceeded - so there are three binary decisions, not a fused QS. This cascaded approach in [34] is primarily meant to reduce the computational complexity for video processing. In the follow-up paper [32] the same three measures were utilized, but without the cascaded approach. Instead, the output of the three so called "objective" measures forms a QS vector. An additional "relative" quality measurement is conducted to assess the dissimilarity of the input image (e.g. from the test dataset) to the training dataset images. This is done via a multivariate Gaussian distribution for a 6-dimensional vector, consisting of the averaged red/green/blue color channel values, and the three aforementioned "objective" measure values. To finally predict a binary QS label, an unspecified number of weak classifiers were learned via AdaBoost to form a combined FQAA, the input thereof being a 9-dimensional vector made up of the 3-dimensional "objective" and 6-dimensional "relative" measure output. Note that the "relative" measure is entirely optional, but does improve the quality assessment according to the evaluation in [32]. In these evaluations both variants of the proposed FQAA appeared superior to the also tested RQS [56], which seemed to actually degrade FR performance.

The work [33] by Damer *et al.* included three face frame selection methods, of which two could be considered for single-image FQA. One method measures the entropy of the color channels, higher entropy being preferred. The other method calculates the confidence for a Viola-Jones [97] face detector as the sub-image classifier detection count, which can correspond i.a. to pose and illumination.

In [31] by Vignesh *et al.* a CNN is utilized to directly output a final FR-performance-focused QS for a 64×64 face image input. The network has 4 convolutional layers and the face image input is preprocessed using PCA whitening. Training this approach requires a ground truth QS corresponding to each training image, which the paper notably computed by comparing each given probe frame against a sequence of gallery frames via the MSM (Mutual Subspace Method) based on either LBP or HOG features. Since the CNN itself only uses single-image input, this ground truth QS generation could naturally be replaced by some single-image approach as well.

Hu *et al.* [30] proposed to train a KPLSR (Kernel Partial Least Squares Regression) model for FQA. Two features are derived for 10×10 sub-blocks of an image, forming a 200-dimensional feature vector as input for the KPLSR model. These features are the mean luminance and Laplacian per sub-block. The training ground truth QSs are LBP-based FR comparison scores, whereby each image pair consists of one image with "standard" (i.e. presumably good and unaltered) illumination, and one image variant with reduced luminance/ contrast. A strong correlation between the FQAA and the FR performance was demonstrated in the evaluation.

Wang [29] presented a hybrid approach to estimate subjective QSs using features consisting of 7 factor-specific scores. The factors comprise brightness, dynamic range, illuminance uniformity, sharpness, pose (yaw/pitch angles), as well as the landmark-based similarity to a "typical" face formed from the average of various training images. A random forest regressor was trained using these factors to estimate subjective ground truth QSs from 1 to 5. The single-image part of the evaluation compares the predictive performance of this approach against the cascaded SVM method of [8], with the results favoring the proposed approach for QSs 2 to 3.

Finally, Qi *et al.* [28] used a CNN architecture with an inception module for FQA. Ground truth QS labels were established in form of gallery DL FR comparison dissimilarity score (i.e. cosine distance) minima for detected faces in training video data. In other words, each training probe image was compared to all training gallery images, and the best score was selected as the ground truth QS to train the FQA network. A pretrained VGG-16 [100] and Inception-v3 [101] network was used for the FR part. The video frame FR performance improvement evaluation i.a. compares against the CNN approach of Vignesh *et al.* [31] and the learning to rank approach of Chen *et al.* [56], with the proposed CNN showing the best results.

D. DL FQA literature

The deep learning FQA approaches that do not fit the categorizations of the prior subsections are listed in Table IV (i.e. all except two in the video frame FQA subsection III-C). Most of these DL FQA papers were published in 2019 and 2020. In the present subsection we navigate Table IV in chronological order from bottom to top.

The learning to rank approach of Chen *et al.* [56] works in two stages. In stage one a number of preexisting feature extractors are used on the input image, and for each feature output vector thereof a RQS (Rank based Quality Score) is derived as the features' weighted sum. Stage two applies a polynomial kernel to the RQS output vector of stage one, and again uses the weighted sum of the resulting vector elements to obtain the final scalar RQS (normalized to $[0, 100]$). "Learning to rank" refers to learning the various weights for said weighted sums so that each RQS differentiates between images from a number

TABLE IV
MOST RELVANT WORKS ON DL FQA.

Reference	Year	Method(s)	Datasets	Miscellaneous
Hernandez-Ortega *et al.* [42]	2020	Same as [50], but with dropout before the first fully connected layer, and multiple FR feature extractors to obtain the ground truth QSs	VGGFace2, LFW, CyberExtruder, BioSecure	Continuation of [50]. Open source. Benchmarked in NIST FRVT QA [62].
Chang *et al.* [43]	2020	Two methods to learn both uncertainty and FR features: 1. Kullback-Leibler divergence loss to train an entire network, 2. Extension of fixed FR network with loss relative to subject feature centers	MS-Celeb-1M, LFW, MegaFace, CFP, YTF, IJB-C	Builds upon the uncertainty vector concept of [49]. Needs no QS labels.
Terhörst *et al.* [44]	2020	QS based on comparing embeddings from 100 random subnetworks; Works on FR networks trained with dropout, or by adding a network on top	MS-Celeb-1M, FERET, Adience, LFW	Usable without extra training. Open source.
Rose and Bourlai [45]	2020	3 binary attributes (Eyes open? Glasses? Fontal?); DL: Pretrained AlexNet [102], GoogLeNet [103]; Non-DL: 23 models, i.a. SVMs	In-house, MEDS-II	SVM + DL score-level fusion led to the best results.
Zhao *et al.* [46]	2019	CNN trained on binary labels derived automatically based on fewer manual labels and non-DL methods	In-house, CASIA-WebFace	Predicts scalar QSs after binary training.
Lijun *et al.* [47]	2019	Multi-branch CNN trained for 4 factors: Alignment, Occlusion, Pose, Blur (+ fused overall QS); QS ground truths manually annotated for 3000 images	IJB-A, MS-Celeb-1M, CASIA-WebFace, LFW	–
Rose and Bourlai [48]	2019	Same as [45], but i.a. with smartphone images	In-house, MEDS-II	Continuation of [45].
Shi and Jain [49]	2019	Based on a pretrained FR network, trains separate two-layer perceptron network to measure per-feature-dimension uncertainty, compares via MLS (Mutual Likelihood Score)	CASIA-WebFace, MS-Celeb-1M, LFW, YTF, MegaFace, CFP, IJB-A, IJB-C, IJB-S	Output is an uncertainty vector. No need for QS labels. Open source.
Hernandez-Ortega *et al.* [50]	2019	Frozen FR-pretrained ResNet-50 [104], training two new final layer replacements on QSs derived from FR features vs. BioLab-ICAO[23]-selected references	VGGFace2, BioSecure	Open source. Benchmarked in NIST FRVT QA [62].
Yang *et al.* [51]	2019	"DFQA", a SqueezeNet[105]-based two-branch CNN; training with SVR loss; ground truth QSs generated by another CNN, in turn trained using 3000 rule-guided human QS labels	ImageNet, IJB-A, MS-Celeb-1M, CASIA-WebFace, VGGFace2, LFW	Direct SqueezeNet successor: SqueezeNext [106].
Wasnik *et al.* [52]	2018	14 methods: 2 FQA CNNs, 5 non-FQA CNNs, 3 non-FQA mobile CNNs, 3 Hand-crafted, 1 COTS; Binary training labels (good/bad)	In-house, CAS-PEAL, Extended Yale, AR, FRGC, NCKU face, ChokePoint, SCface	Considers FQA for smartphone FR, but not exclusively.
Yu *et al.* [53]	2018	CNN with MFM[107] & NIN[108] layers, trained using 15 synthetic degradation classes (5 types $\times$ 3 settings)	CASIA-WebFace, LFW, YTF	–
Best-Rowden and Jain [54]	2017	5 methods, using either human or FR-based labels, DL or L2R [56] features, and SVR or L2R models	LFW, IJB-A, CASIA-WebFace	Another paper version is [109].
Zhang *et al.* [55]	2017	ResNet-50 trained on subjective illumination QSs	FIIQD	Open source.
Chen *et al.* [56]	2015	2-stage learning to rank for five feature extractors: CNN (Landmarks), HOG, Gist [89], Gabor, LBP (per-feature-vector QS formed by weighted sum)	In-house/Unknown, FERET, FRGC, LFW, AFLW, SCface	Can be considered non-DL by removing the CNN extractor. Open source.

of training datasets with a given assumed quality ordering (e.g. some training dataset A is defined to be of higher quality than dataset B, which in turn is defined to be of higher quality than dataset C). Conceptually, this approach doesn't have to use any deep learning, but the evaluated FQAA implementation incorporates a CNN for facial landmark detection as one of five feature extractors, thus [56] has been categorized as part of this subsection. The other four (non-DL) feature extractors comprise Gist [89], HOG, Gabor, and LBP.

Zhang *et al.* [55] created FIIQD, a "Face Image Illumination Quality Database" with subjective illumination quality scores for $224,733$ images with 200 different illumination patterns (established patterns were transferred to images from various other databases, together with their associated ground truth QS labels). Then a model based on ResNet-50 [104] was trained with that data to estimate the illumination quality. A strong correlation was shown between the predicted illumination QSs and the labels, but the impact on FR performance was not evaluated.

In [54] and [109], Best-Rowden and Jain presented multiple FQAA variants partially based on DL. Five FQAAs were evaluated, including the RQS approach of [56]. Of the four newly proposed FQAAs, three use training ground truth QSs derived from pairwise relative human assessments, and one derives the ground truth QSs from FR-method-dependent comparison scores with manually selected gallery images. Two of the methods use the 320-dimensional feature vector of a FR CNN [110] to train a SVR model for the QS prediction, one method targeting the FR scores (Matcher Quality Values, "MQV"), the other targeting the human assessment ground truth (Human Quality Values, "HQV-0"). The CNN features are also used in another one of the human ground truth methods, which replaces the SVR with the L2R (learning to rank) approach of [56] ("HQV-1"). The fourth method trains the L2R approach of [56] with the features described therein, but for the human ground truth instead of the RQS dataset

constraints [56] ("HQV-2"). In the evaluation, the CNN of [110] was also used as one of the FR algorithms, in addition to two unnamed COTS. The methods HQV-2 and MQV showed the lowest improvements regarding FR performance. The best FR improvements were achieved using HQV-1 for the CNN [110], and RQS [56] for one of the COTS.

Yu *et al.* [53] proposed using a CNN architecture with MFM [107] (Max-Feature-Map) and NIN [108] (Network In Network) layers for FQA. There are 16 classes used for training, one being the the original unmodified training images, while the other 15 represent 5 types of synthetic degradation thereof with 3 configurations of increasing severity each. These 5 degradation types comprise nearest-neighbor downscaling, Gaussian blur, AWGN (Additive White Gaussian Noise), salt-and-pepper noise, and Poisson noise. This is sufficient to train a network to classify these degradations. To also estimate a scalar QS, a FR accuracy score is precomputed for each of the 16 classes, and the sum of the multiplication of those scores with the 16 classification probabilities forms the combined QS. The proposed CNN architecture was also used for the FR part (as a separately trained model), using Cosine distance as the dissimilarity measure. Three variants of the network were evaluated for FQA: One trained from scratch for FQA, one first trained for FR before training for FQA, and one that uses ReLU instead of MFM layers. The evaluation i.a. compared the variants regarding their degradation classification performance, showing superior accuracy for the two MFM variants in contrast to the ReLU architecture, whereby the best overall results stemmed from the FR transfer learning variant. Regarding the 5 degradation types, the FR performance appeared to be predominantly affected by AWGN as well as salt-and-pepper noise, while the other types were less impactful even for their more severe configurations.

Wasnik *et al.* [52] compared 14 methods for FQA using 7 publicly available datasets (plus in-house datasets) in the context of smartphone FR. Of the 14 methods, 10 are CNNs, 3 are hand-crafted, and 1 is a COTS (VeriLook 5.4 [111]). Among the 3 hand-crafted methods, 2 are general IQAAs (BLIINDS-II [112], BRISQUE [113]), and 1 is Wasnik *et al.* [22] from subsection III-B. Among the 10 pretrained CNNs, 2 are meant specifically for FQA (the illumination-focused FQA [55], and the general FQA [28], both in subsection III-C), 3 are mobile networks (MobileNetV2 [114], DenseNet-169 [115], NASNet [116]), and the other 5 are AlexNet [102], VGG-16/VGG-19 [100], Inception [103], and Xception [117]. Of the 2 FQA-specific CNNs, for [55] a pretrained network provided by the authors was used, and for [28] the network described therein was recreated while using the training dataset of [52]. To adapt the non-FQA CNNs for the FQA task, the last three layers were replaced by fully connected layers of size 1024, 512 and 2, 2 being the number of training data classes. So training images were either labeled good or bad regarding quality, with the latter referring to presumed flaws for e.g. illumination or pose. Note that this means that the training didn't directly target some ground truth QS produced via e.g. an FR system. Nevertheless, the best FR performance improvements in the evaluation were achieved by the two larger FQA-adapted CNNs AlexNet and Inception. This evaluation used 5 separate datasets, and the VeriLook SDK 5.4 [111] for FR comparisons.

Yang *et al.* [51] presented "DFQA", a FQA CNN based on SqueezeNet [105], which itself is notably meant to provide performance comparable to AlexNet with 50× fewer parameters (also note that by this point in time a direct successor exists, namely SqueezeNext [106]). However, it isn't proven whether this performance equivalence is true for the biometric FQA task here, since [51] doesn't compare against any AlexNet-based FQA variant, e.g. one analogous to their SqueezeNet-based approach, or the one used in [52]. Most of the SqueezeNet architecture parts in the DFQA [51] network are represented in two functionally identical weight-sharing branches (also called "streams" in [51]), each of which is followed by a (no longer weight-sharing) 1×1 kernel convolutional layer with 9×9 output. Then the mean of the two outputs is fed to an average pooling layer, resulting in the output feature vector. The paper compares both Euclidean and SVR loss, showing better results for the latter. Different branch counts, 1 to 4, were evaluated as well. For training, 3,000 images were first manually annotated with ground truth QS values, using a defined set of rules to increase the level QS objectivity/subject-independence. These images were used to train another CNN, based on a pretrained SqueezeNet, to predict ground truth QSs for the MS-Celeb-1M [118] dataset, which were then used to train the actual DFQA.

Hernandez-Ortega *et al.* created the open source FQAA "FaceQnet" v0 [50] and v1 [42]. As part of the training data preparation for both FaceQnet versions, the BioLab-ICAO framework from [23] is employed to select suitable high-quality images per subject, which are used to compute the ground truth QSs for the subjects' remaining training images. This ground truth QS computation consists of the normalized Euclidean distances of embeddings produced by a number of FR feature extractors (three for v1; and only one, FaceNet [119], for v0). Both FaceQnet versions are based on a ResNet-50 [104] model pretrained for FR using the VGGFace2 [120] dataset, replacing the final output layer with two fully connected layers. Only these two new layers are trained, the rest of the network weights are frozen. FaceQnet v1 extends the training architecture by adding dropout before the first fully connected layer. I.e. the architecture of FaceQnet v1 and v0 after training are identical, but FaceQnet v1 was trained with dropout and using ground truth QSs derived from multiple feature extractors. Both versions used a 300-subject subset of the VGGFace2 [120] for training. FaceQnet is one of the very few (if not the only) FQAA surveyed here that has been benchmarked by an independent evaluator, with FaceQnet v0 being the only surveyed work that is included in the current report of the new NIST FRVT Quality Assessment campaign [62].

Shi and Jain [49], and based thereon Chang *et al.* [43], proposed to compute an uncertainty vector that directly corresponds to the FR feature vector for a single face image. In other words, the two output vectors are representing the Gaussian variance and mean, respectively. Both of the papers focus on using the uncertainty as part of the FR comparison, so producing a single scalar QS is not the primary goal.

Consequently, the papers primarily compare against other FR methods, not against FQAAs. Both papers nevertheless also note that the uncertainty could be used for FQA purposes. E.g. in [49] an evaluation showed that filtering images by the inverse harmonic mean of the uncertainty vector elements can be more effective to improve FR performance than filtering using face detection scores. So the uncertainty can certainly be considered as a kind of QS, and a scalar QS can be derived from such a vector. The implementation of [49] uses a fixed pretrained FR network as basis to compute the FR feature vector (i.e. Gaussian mean), and trains an additional uncertainty module for the uncertainty vector (i.e. variance), on the same training dataset used for the FR network. Said uncertainty module is a two-layer perceptron network, using the same input as the FR layer that outputs the original feature vector. To incorporate the uncertainty vector in the FR comparison, a MLS (Mutual Likelihood Score) is proposed by [49], which weighs and penalizes feature dimensions depending on the uncertainty. The uncertainty module training attempts to maximize this MLS for all genuine image pairs. In addition, [49] explains how the uncertainty can be used to fuse embeddings for multiple images. Extending the concept of [49], [43] proposed two methods to learn both uncertainty (variance) and feature (mean) at the same time, without a separate uncertainty module. This means that the uncertainty can improve the overall training by reducing the influence of low quality images, which implies that the FR performance may improve even if the uncertainty isn't used after training, although it is noted that this kind of quality attention can reduce performance when only low quality cases are considered after training. By omitting a separate uncertainty vector for comparisons, the MLS of [49] does not have to be used, thus avoiding increased computational complexity as evaluated in [43]. One of the two methods in [43] is "classification-based" and learns an entire FR network with both regular feature and uncertainty output, together forming a sampling representation for training, using the reparameterization trick [121] to enable backpropagation. Instead of using the MLS, the cost function consists of a softmax classification loss, plus a regularization term to control the uncertainty aspect. The latter is the Kullback-Leibler divergence scaled by a scalar hyper-parameter, comparing the mean and variance output relative to a normal distribution. The other learning method of [43] is "regression-based" and more akin to the separate uncertainty module training concept of [49]: Similar to [49] it begins by using a FR feature network trained in isolation, then the weights are frozen and uncertainty output is added. But in contrast to [49] the FR features (mean) aren't frozen with the rest of the pretrained layers, and the method continues training them simultaneously with the uncertainty, using loss based on the per-subject feature vector center derived from the isolated FR network stage. As part of the evaluations on multiple FR base models in [43], the two methods of [43] (using cosine similarity for comparisons) and the method of [49] (using MLS for comparisons, including fusion where applicable) were compared. The "classification-based" method [43] is found to mostly result in better performance increases than the predecessor method of [49], while the "regression-based" method [43] appears either worse or better depending on the scenario (and is considered for further examination due to some observed performance regression with respect to the FR baseline).

Rose and Bourlai evaluated DL and non-DL methods to determine three binary facial attributes in [45] and [48] (which is a continuation of [45] despite the publication date order): Whether the eyes are open or closed, whether there are glasses or not, and whether the face pose is mostly frontal or not. The two DL methods in both papers consisted of AlexNet [102] and GoogLeNet [103] (an incarnation of the Inception architecture), pretrained on ImageNet [122] data. Their architectures were modified to classify 2 labels per attribute (i.e. 6 classes). And there were 23 non-DL models tested in [45], including SVMs, K-Nearest Neighbors, Decision Trees, and Ensemble classifiers. LBP and HOG features were evaluated for these non-DL methods, and HOG was found to consistently outperform LBP. A score-level fusion of a SVM and either AlexNet or GoogLeNet led to the best results in [45]. The evaluations in [48] employ a smartphone (iPhone 5S) dataset in addition to the non-smartphone data used in [45], the latter of which is only used for training. Of the non-DL methods, result values in [48] are only shown for the cubic kernel SVM approach, because the other methods performed worse. Whether the performance of the SVM or one of the two DL methods is better varies between the experiments of [48], which proposed to use the SVM trained on a combination of all used datasets (score-level fusion of SVM and one of the DL networks is not tested).

Lijun *et al.* [47] proposed a multi-branch FQA network (called MFQA) consisting of a feature extraction and a quality score part. The former is a CNN to derive image features. The latter feeds these features into four fully connected branches for different quality properties, and fuses the output thereof into a final QS via another fully connected layer. These four branches correspond to scores for alignment, visibility (i.e. occlusion), pose, and clarity (i.e. blur). Multi-task learning is employed, and 3,000 images were manually annotated with ground truth labels for the four factor scores and the overall QS.

Zhao *et al.* [46] trained a CNN for FQA in a semi-supervised fashion. First, binary labels (good/bad) are manually assigned to a number of images to train a preliminary version of the DL model. This preliminary network then predicts labels for a different (larger) dataset in the second stage. The third stage updates these labels utilizing various additional binary constraints derived from the inter-eye distance, the pitch and yaw rotation, the contrast, and further factors not listed in [46] due to paper length limitations. For all "good" labels predicted by the preliminary network, the label will be changed to "bad" if any of these binary constraints are "bad". I.e. "bad" label predictions are not altered. This newly labeled dataset is then used in the fourth and final stage to fine-tune the model. Hinge loss is used during training for the binary classification task, but after training the network is modified to output a [0, 1] scalar QS prediction instead. The paper notes that the CNN has better computational performance than the CNN proposed by [53].

Terhörst *et al.* [44] proposed the open source "SER-FIQ" method in two variants, measuring FR-model-specific quality by comparing the output embeddings of a number of randomly chosen subnetworks, i.e. without requiring any ground truth QS training labels. A QS is computed as the sigmoid of the negative mean of the Euclidean distances between all random subnetwork embeddings, meaning that the computational complexity grows quadratically with respect to the number of subnetworks (100 are used in [44]). The "same model" variant of SER-FIQ can be used on FR networks trained using dropout, without additional training. For this variant's implementation in [44], the random subnetwork passes use the last two FR layers. The other variant is the "on-top model", meaning that a small additional network is trained with dropout on top of the FR model to transform its FR embeddings. Five layers with dropout are used in the implementation, which includes the identity classification layer for training. Removing that, the first and last layer of the network have the same dimensions as the FR embedding. Evaluations used FaceNet [119] and ArcFace [123] for FR, and selected images using QSs from both SER-FIQ variants, FaceQnet v0 [50], an approach proposed by Best-Rowden in [109], 3 general IQAAs (BRISQUE [113], NIQE [124], PIQE [125]), as well as a COTS (Neurotec Biometric SDK 11.1 [111]). The SER-FIQ "on-top model" was noted to mostly outperform all baseline approaches, and to always deliver close to top performance. The "same model" approach mostly outperformed the baseline methods by a larger margin, showing especially strong FNMR (False Non-Match Rate) performance improvements for a fixed FMR (False Match Rate) of 0.001.

IV. Open Issues and Challenges

An obvious challenge consists of the further improvement of FQA methods in terms of predictive and computational performance. For deep learning FQA approaches, finding better network architectures and training methods is interwoven with general deep learning research progress, for example in the field of automated machine learning [166]. Naturally, FQA with the goal of predicting quality scores that represent FR utility [63] also depends on FR research.

The following subsections describe further issues and challenges, as well as potential avenues for future work, and the summary section V highlights the identified key challenges.

A. Comparisons and Reproducibility

As previously noted in subsection II-B, it would be challenging to comprehensively compare the performance of the surveyed FQA approaches, since the evaluations presented in the literature differ in multiple aspects that would need to be aligned to facilitate direct comparisons:

- **Datasets**: As shown by Table V, a variety of datasets are used for the evaluations among the literature. Besides these named datasets, some of the literature only utilized private or unspecified data for evaluation. In addition, some literature used only a subset of a dataset (see e.g. [42] or [51] regarding the VGGFace2 [120] dataset), or modified the data e.g. by synthetically degrading images via increased blur or contrast (see e.g. [1]). Where training data is required for the FQAA, the chosen subdivision of the datasets into training and test data also influences the evaluation results. And various works assign ground truth quality scores or labels to the dataset for FQAA training and/or for the evaluation. When FQA is evaluated in terms of FR performance improvements, the selection of image pairs that are considered initially for FR comparisons [60] (i.e. before filtering them via FQA decisions) alters the results as well.
- **Evaluation methods**: Different evaluation methods and result presentations are used among the literature. Some FQA approaches are only tested by comparing predicted quality scores or labels against a given ground truth (e.g. assigned by humans), so not all of the literature evaluates FQA in terms of FR utility [63][64] in the first place. Instead of evaluating the FQA on its own, there also is literature that includes image enhancement steps in the evaluation. For FR performance improvement evaluations via an ERC as described in subsection II-B, the FR comparison score threshold [60] and the error type configuration can differ between evaluations, which also applies to ERC-derived AUC results. And some of the works evaluated FR performance exclusively by other means than an ERC - for example, performance was evaluated for 4 FQA-derived quality bins in [3].
- **FR algorithms**: Evaluating FQA in terms of FR performance improvements is desirable to examine how well quality scores of a FQAA reflect FR utility [63], but this also introduces the FR algorithm choice for feature extraction [60] and comparison [60] as another evaluation factor. Furthermore, there are FQA approaches among the literature which are conceptually based on FR models to begin with (see e.g. [44]), and FR algorithms are used by various works to establish ground truth quality scores/ labels (see e.g. [42] for scores, or [3] for labels in the form of 4 quality bins). Lastly, some literature exclusively used COTS FR systems (see e.g. [3]).

Due to the amount of existing and possible FQA evaluation configurations, the comparison of FQAAs can be considered as a key challenge. If desired, this open issue could be limited in scope e.g. by only considering FQA approaches that can conceptually adapt to deep learning FR systems (instead of relying on hand-crafted algorithms, settings, or ground truth quality scores). One solution for future work is to provide the presented FQAAs to an ongoing comparison project, such as the previously mentioned NIST FRVT Quality Assessment evaluation [62].

Another solution is to publicly provide the FQAA implementations, allowing other researchers to integrate them in different evaluation environments without re-implementation. Besides being redundant effort, a re-implementation can diverge from the original implementation to some degree even without introducing errors, since e.g. deep learning model weight initialization can be random (which however might only be a minor issue). Since evaluations of machine learning FQA in particular depend on the used training data, publishing

TABLE V
DATASETS THAT WERE USED IN THE LITERATURE TO TRAIN OR EVALUATE FQA APPROACHES. IN-HOUSE DATASETS OR DATASETS USED ONLY FOR OTHER PURPOSES (SUCH AS FR MODEL TRAINING) ARE NOT LISTED. THE LEFT TABLE LISTS DATASETS THAT WERE USED ONCE, AND THE RIGHT TABLE LISTS DATASETS USED IN MULTIPLE WORKS.

Dataset	Year	Used in
Adience [126]	2020	[44]
CyberExtruder [127]	2020	[42]
IJB-S [128]	2019	[49]
ImageNet [122]	2019	[51]
CMU-FIA [129]	2018	[28]
NCKU face [130]	2018	[52]
FIIQD [55]	2017	[55]
Honda/UCSD [131]	2017	[29]
FEI [132]	2016	[30]
MIT [133]	2016	[30]
AFLW [134]	2015	[56]
BioLab-ICAO [23]	2012	[23]
IIT-NRC [135]	2011	[37]
Pointing'04 [136]	2011	[37]
XM2VTS [137]	2010	[11]
FRI-CVL [138]	2008	[39]
HERMES project [139]	2008	[39]
Cohn-Kanade [140]	2007	[17]
WVU [141]	2007	[17]

Dataset	Usage timespan	Used in
LFW [142]	2011 to 2020	11: [10][29][42][43][44][47][49][51][53][54][56]
FERET [143]	2007 to 2020	9: [1][7][9][10][17][24][30][44][56]
CAS-PEAL [144]	2009 to 2018	6: [1][3][7][15][30][52]
CASIA-WebFace [145]	2017 to 2019	6: [46][47][49][51][53][54]
FRGC [146]	2006 to 2018	6: [26][32][34][35][52][56]
MS-Celeb-1M [118]	2019 to 2020	5: [43][44][47][49][51]
YTF [147]	2014 to 2020	5: [33][36][43][49][53]
ChokePoint [9]	2011 to 2018	4: [9][28][31][52]
Extended Yale [148]	2010 to 2018	4: [4][11][14][52]
IJB-A [149]	2017 to 2019	4: [47][49][51][54]
SCface [150]	2011 to 2018	4: [3][10][52][56]
AT&T [151]	2010 to 2016	3: [14][30][35]
BANCA [152]	2006 to 2008	3: [18][19][38]
CMU-PIE [153]	2009 to 2011	3: [9][12][24]
FRVT 2006 [154]	2008 to 2010	3: [12][13][16]
GBU [155]	2012 to 2014	3: [1][2][7]
VGGFace2 [120]	2019 to 2020	3: [42][50][51]
Yale [156]	2007 to 2014	3: [1][7][25]
AR [157]	2014 to 2018	2: [35][52]
BioSecure [158]	2019 to 2020	2: [42][50]
CFP [159]	2019 to 2020	2: [43][49]
IJB-C [160]	2019 to 2020	2: [43][49]
MBGC [161]	2012 to 2014	2: [1][7]
MEDS-II [162]	2019 to 2020	2: [45][48]
MegaFace [163]	2019 to 2020	2: [43][49]
PaSC [164]	2013 to 2018	2: [2][28]
Q-FIRE [165]	2012 to 2014	2: [1][6]

source code is preferable to pure black box implementations. So for the sake of both comparability and reproducibility, future work should provide source code and trained models where applicable. This may also serve as a basis for new FQA approaches in later work by other researchers. Effective reuse of prior work implementations can i.a. be observed in the surveyed literature by the utilization of pretrained FR models. Providing source code is not necessarily important for approaches that can easily be described in complete detail within a paper, e.g. simpler hand-crafted methods without any machine learning and few parameters, but approaches in the recent literature tend to be more complex. While most of the older surveyed literature didn't appear to publish accompanying source code (irrespective of the implementation complexity), more recent deep learning FQA works tend to do so, with code being publicly available for e.g. FaceQnet [50][42], PFE (Probabilistic Face Embeddings) [49], and SER-FIQ [44].

Likewise, public datasets should preferably be used, and precise evaluation configurations could be published alongside the implementation. It may also be helpful to publish the raw evaluation result as supplementary data, e.g. the computed comparison scores and quality scores, although this may be unnecessary if the results are reproducible already. This result data could e.g. be used to directly create new visualizations that combine results from multiple works.

Outside of evaluating the predictive performance of FQAAs, evaluating the computational performance may be of relevance as well. This is only sparsely considered in the surveyed FQA literature. Computational performance tests can usually focus on measuring the duration required to process input images with a certain format (e.g. grayscale) and resolution, since they are typically not influenced by other factors that are unavoidable in predictive performance evaluations. Other factors do however become relevant, namely the computational optimization of the FQAA, as well as the used hardware and the robustness of the time measurements.

B. Robustness and Capabilities

The surveyed machine learning FQA literature does not study adversarial attacks, i.e. attacks that specifically modify the input (physical [167] or digital after being captured and processed [168]) to confuse the FQA model. An investigation into this topic could be of interest in the context of FQA robustness.

Also, while the more recent deep learning FQA approaches are trained specifically to output quality scores in terms of FR utility [60][64], they currently aren't as interpretable/ explainable as e.g. hand-crafted approaches that estimate specific human-understandable factors such as blur. This can be considered as another key challenge. Optimally, FQA models should be able to predict FR utility [60] while also providing useful feedback regarding quality-degrading causes. Future work could thus attempt to improve upon this area, perhaps by adding generational capabilities to visually represent a disentangled latent space that corresponds to different kinds of quality degradations. In this line of explainable Artificial Intelligence (AI) and, in particular, in fairness and bias control in AI systems [169][170], we expect growing interest in analyzing the behavior of FQA methods for different population groups and the development of FQA methods more transparent [171] and agnostic to selected covariates [172].

For FQA in general, preferably large amounts of realistic data including different quality levels with different quality-degrading causes should be used for evaluation (and training where applicable), so that the robustness can be verified for various cases with a high certainty. Existing images can also be degraded synthetically - this was done by some of the literature (e.g. [1]), but wasn't common. So both known techniques from prior work, such as Gaussian blurring, and more sophisticated techniques, such as deep learning style transfer, could be harnessed in the future. It is also possible to generate fully synthetic face images (see e.g. StyleALAE [173]), which hasn't been used in the surveyed FQA literature. While fully synthetic data might be less realistic, it could allow for larger datasets with better control (in terms of training/ evaluation sample bias) than what e.g. filtering a real dataset might provide. As a side effect, using fully synthetic data may potentially also alleviate licensing or privacy concerns (see e.g. the controversy surrounding MS-Celeb-1M [118], which has been used by some of the FQA literature as well). This latter point is however not entirely clear, since deep learning face synthesis itself is typically trained on real face images.

Examining and improving interoperability in terms of FQA FR utility [60] prediction generality could be another goal for future work. While this may partially stand in conflict with the goal of maximising FR-system-specific utility prediction performance, interoperability can be relevant to avoid vendor lock-in and may coincide with increased robustness. An example in the literature is the FaceQnet approach, which went from using only one FR system as part of the training process in v0 [50] to using three in v1 [42].

As described in subsection II-A, there are further application areas that are barely or not at all examined in the surveyed literature. For example, lossy compression control isn't considered at all, although compression artifacts are mentioned as a quality degrading factor by various works.

V. Summary

Face image quality assessment is an active research area, and can be used for a variety of face recognition related application scenarios, including gender or other soft biometrics recognition [67], attention level estimation [66], emotion analysis [65], etc. The literature surveyed in this work predominantly focused on evaluating their proposed FQA approaches either in terms of predictive performance with respect to given ground truth quality score labels, or in terms of utility [60][64] for the purpose of aiding face recognition by discarding images based on the assessed quality or some kind of quality-based processing or fusion [82]. A subset of the literature also covered the more specific topic of using FQA for video frame selection (see subsection III-C). Automatic face quality assessment is especially relevant for FR as part of large-scale systems, e.g. the European SIS (Schengen Information System) [69], due to the amount of data and the multitude of different acquisition locations/devices.

A progression over time towards deep learning was observed in the FQA literature (see Figure 6). The three most recent listed FQA works contain deep learning FQA approaches (see Table IV), and have clear conceptual distinctions: The FaceQnet v1 [42] approach trains the model with ground truth quality scores derived by means of three FR algorithms, the data uncertainty learning [43] approach extends FR models to predict uncertainty corresponding to each FR feature vector dimension, and the SER-FIQ [44] approach measures quality by comparing the output embeddings of multiple randomly chosen subnetworks within an unmodified FR model or a small model trained on top of a FR model.

One key challenge is to facilitate comparability of the FQA evaluations, since many differing evaluation configurations were employed in the literature (see subsection IV-A). Thus, future work should preferably provide the implementations of the proposed FQAAs publicly, especially in the form of source code, enabling evaluations in later works to more easily include these FQA approaches. There also is the ongoing NIST FRVT Quality Assessment evaluation [62], to which FQAAs can be submitted. Besides evaluating the predictive capabilities of FQAAs, more attention could be paid to computational performance evaluations in the future.

Another key challenge is to improve the interpretability of deep learning based FQA, which so far didn't focus on providing extensive feedback for human operators to adjust acquisition conditions for increased biometric utility [60] (see subsection IV-B).

And of course there is the obvious key challenge of further improving performance in terms of both utility [60] and computational workload (e.g. with new deep learning network architectures), as well as improving robustness/decreasing bias [169][170] (e.g. via the selection or synthetic extension of datasets for different quality degradation cases), since the FQA approaches are unlikely to be optimal already.

Besides these key challenges, various other application scenarios can be explored further, e.g. FQA-guided image enhancement or compression (see subsection II-A).

Acknowledgment

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, project BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), and project TRESPASS-ETN (H2020-MSCA-ITN-2019-860813).

References

[1] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross, "Design and evaluation of photometric image quality measures for effective face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 314–324, Dec. 2014, ISSN: 2047-4938, 2047-4946.

[2] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, *et al.*, "On the existence of face quality measures," in *Proc. 6th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, Sep. 2013, pp. 1–8, ISBN: 978-1-4799-0527-0.

[3] S. Bharadwaj, M. Vatsa, and R. Singh, "Can holistic representations be used for face biometric quality assessment?" In *Proc. 20th Intl. Conf. on Image Processing (ICIP)*, IEEE, Sep. 2013, pp. 2792–2796, ISBN: 978-1-4799-2341-0.

[4] F. Qu, D. Ren, X. Liu, Z. Jing, and L. Yan, "A face image illumination quality evaluation method based on Gaussian low-pass filter," in *2nd Intl. Conf. on Cloud Computing and Intelligence Systems (CCIS)*, IEEE, Oct. 2012, pp. 176–180.

[5] B. F. Klare and A. K. Jain, "Face recognition: Impostor-based measures of uniqueness and quality," in *Proc. 5th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, Arlington, VA, USA: IEEE, 2012, pp. 237–244.

[6] F. Hua, P. Johnson, N. Sazonova, P. Lopez-Meyer, and S. Schuckers, "Impact of out-of-focus blur on face recognition performance based on modular transfer function," in *Proc. 5th IAPR Intl. Conf. on Biometrics (ICB)*, IEEE, Mar. 2012, pp. 85–90, ISBN: 978-1-4673-0397-2 978-1-4673-0396-5 978-1-4673-0395-8.

[7] A. Abaza, M. A. Harrison, and T. Bourlai, "Quality metrics for practical face recognition," *Proc. 21st Intl. Conf. on Pattern Recognition (ICPR)*, p. 5, 2012.

[8] P. Liao, H. Lin, P. Zeng, S. Bai, H. Ma, *et al.*, "Facial Image Quality Assessment Based on Support Vector Machines," in *Intl. Conf. on Biomedical Engineering and Biotechnology (ICBEB)*, May 2012, pp. 810–813.

[9] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2011, pp. 74–81, ISBN: 978-1-4577-0529-8.

[10] M. D. Marsico, M. Nappi, and D. Riccio, "Measuring measures for face sample quality," in *Proc. of the 3rd Intl. ACM Workshop on Multimedia in Forensics and Intelligence (MiFor)*, ACM Press, 2011, p. 7, ISBN: 978-1-4503-0987-5.

[11] D. Rizo-Rodriguez, H. Méndez-Vázquez, and E. Garcia-Reyes, "An Illumination Quality Measure for Face Recognition," in *Proc. 20th Intl. Conf. on Pattern Recognition (ICPR)*, IEEE, Aug. 2010, pp. 1477–1480, ISBN: 978-1-4244-7542-1.

[12] J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, *et al.*, "Quantifying how lighting and focus affect face recognition performance," in *Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2010, pp. 74–81, ISBN: 978-1-4244-7029-7.

[13] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, *et al.*, "FRVT 2006: Quo Vadis face quality," *Image and Vision Computing*, vol. 28, no. 5, pp. 732–743, May 2010, ISSN: 02628856.

[14] H. Sellahewa and S. A. Jassim, "Image-Quality-Based Adaptive Face Recognition," *IEEE Trans. on Instrumentation and Measurement (TIM)*, vol. 59, no. 4, pp. 805–813, Apr. 2010, ISSN: 0018-9456, 1557-9662.

[15] G. Zhang and Y. Wang, "Asymmetry-Based Quality Assessment of Face Images," in *Advances in Visual Computing: 5th Intl. Symposium (ISVC)*, vol. 5876, Springer Berlin / Heidelberg, 2009, pp. 499–508.

[16] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui, "Focus on quality, predicting FRVT 2006 performance," in *Proc. Intl. Conf. on Automatic Face and Gesture Recognition*, IEEE, Sep. 2008, pp. 1–8, ISBN: 978-1-4244-2153-4.

[17] M. Abdel-Mottaleb and M. H. Mahoor, "Algorithms for Assessing the Quality of Facial Images," *IEEE Computational Intelligence Magazine (CIM)*, vol. 2, no. 2, pp. 10–17, May 2007, ISSN: 1556-6048.

[18] K. Kryszczuk and A. Drygajlo, "On combining evidence for reliability estimation in face verification," *14th European Signal Processing Conf. (EUSIPCO)*, Sep. 2006.

[19] ——, "On Face Image Quality Measures," *Proc. 2nd Workshop on Multimodal User Authentication (MMUA)*, p. 8, 2006.

[20] H. Luo, "A training-based no-reference image quality assessment algorithm," in *Proc. Intl. Conf. on Image Processing (ICIP)*, vol. 5, IEEE, 2004, pp. 2973–2976, ISBN: 978-0-7803-8554-2.

[21] A. Khodabakhsh, M. Pedersen, and C. Busch, "Subjective Versus Objective Face Image Quality Evaluation For Face Recognition," in *Proc. 3rd Intl. Conf. on Biometric Engineering and Applications (ICBEA)*, ACM Press, 2019, pp. 36–42.

[22] P. Wasnik, K. B. Raja, R. Raghavendra, and C. Busch, "Assessing face image quality for smartphone based face recognition system," in *5th Intl. Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2017, pp. 1–6.

[23] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face Image Conformance to ISO/ICAO Standards in Machine Readable Travel Documents," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 4, pp. 1204–1213, Aug. 2012, ISSN: 1556-6013, 1556-6021.

[24] J. Sang, Z. Lei, and S. Z. Li, "Face Image Quality Evaluation for ISO/IEC Standards 19794-5 and 29794-5," in *Proc. 3rd IAPR Intl. Conf. on Biometrics (ICB)*, vol. 5558, Springer Berlin Heidelberg, 2009, pp. 229–238, ISBN: 978-3-642-01792-6 978-3-642-01793-3.

[25] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of Face Image Sample Quality," *Proc. 2nd IAPR Intl. Conf. on Biometrics (ICB)*, p. 10, 2007.

[26] R. L. V. Hsu, J. Shah, and B. Martin, "Quality Assessment of Facial Images," in *Biometrics Symposium: Special Session on Research at the Biometric Consortium Conf. (BCC)*, IEEE, Sep. 2006, pp. 1–6, ISBN: 978-1-4244-0486-5 978-1-4244-0487-2.

[27] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec, "Face image validation system," in *Proc. 4th Intl. Symposium on Image and Signal Processing and Analysis (ISPA)*, IEEE, 2005, pp. 30–33, ISBN: 978-953-184-089-7.

[28] X. Qi, C. Liu, and S. Schuckers, "Boosting Face in Video Recognition via CNN Based Key Frame Extraction," in *Intl. Conf. on Biometrics (ICB)*, IEEE, Feb. 2018, pp. 132–139, ISBN: 978-1-5386-4285-6.

[29] C. Wang, "A learning-based human facial image quality evaluation method in video-based face recognition systems," in *3rd Intl. Conf. on Computer and Communications (ICCC)*, IEEE, Dec. 2017, pp. 1632–1636, ISBN: 978-1-5090-6352-9.

[30] X. Hu, L. Zhuo, J. Zhang, and X. Li, "Face image illumination quality assessment for surveillance video using KPLSR," in *Intl. Conf. on Progress in Informatics and Computing (PIC)*, IEEE, Dec. 2016, pp. 330–335, ISBN: 978-1-5090-3484-0.

[31] S. Vignesh, K. Manasa Priya, and S. S. Channappayya, "Face image quality assessment for face selection in surveillance video using convolutional neural networks," in *Proc. 3rd IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, IEEE, Dec. 2015, pp. 577–581, ISBN: 978-1-4799-7591-4.

[32] H. I. Kim, S. H. Lee, and Y. M. Ro, "Face image assessment learned with objective and relative face image qualities for improved face recognition," in *Proc. Intl. Conf. on Image Processing (ICIP)*, IEEE, Sep. 2015, pp. 4027–4031, ISBN: 978-1-4799-8339-1.

[33] N. Damer, T. Samartzidis, and A. Nouak, "Personalized Face Reference from Video: Key-Face Selection and Feature-Level Fusion," in *Face and Facial Expression Recognition from Real World Videos (FFER)*, vol. 8912, Springer International Publishing, 2015, pp. 85–98, ISBN: 978-3-319-13736-0 978-3-319-13737-7.

[34] H. I. Kim, S. H. Lee, and Y. M. Ro, "Investigating Cascaded Face Quality Assessment for Practical Face Recognition System," in *Intl. Symposium on Multimedia (ISM)*, IEEE, Dec. 2014, pp. 399–400, ISBN: 978-1-4799-4311-1 978-1-4799-4312-8.

[35] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch, "Automatic Face Quality Assessment from Video Using Gray Level Co-occurrence Matrix: An Empirical Study on Automatic Border Control System," in *Proc. 22nd Intl. Conf. on Pattern Recognition (ICPR)*, IEEE, Aug. 2014, pp. 438–443, ISBN: 978-1-4799-5209-0.

[36] M. Nikitin, V. Konushin, and A. Konushin, "Face Quality Assessment for Face Verification in Video," *Proc. 24th Intl. Conf. on Computer Graphics and Vision (GraphiCon)*, p. 4, 2014.

[37] K. Nasrollahi and T. B. Moeslund, "Extracting a Good Quality Frontal Face Image From a Low-Resolution Video Sequence," *IEEE Trans. on Circuits and Systems for Video Technology (TCSVT)*, vol. 21, no. 10, pp. 1353–1362, Oct. 2011, ISSN: 1051-8215, 1558-2205.

[38] E. A. Rúa, J. E. L. A. Castro, and C. G. Mateo, "Quality-Based Score Normalization and Frame Selection for Video-Based Person Authentication," in *Biometrics and Identity Management (BioID)*, vol. 5372, Springer Berlin / Heidelberg, 2008, pp. 1–9, ISBN: 978-3-540-89990-7 978-3-540-89991-4.

[39] K. Nasrollahi and T. B. Moeslund, "Face Quality Assessment System in Video Sequences," in *Biometrics and Identity Management (BioID)*, vol. 5372, Springer Berlin Heidelberg, 2008, pp. 10–18, ISBN: 978-3-540-89990-7 978-3-540-89991-4.

[40] A. Fourney and R. Laganiere, "Constructing Face Image Logs that are Both Complete and Concise," in *Proc. 4th Canadian Conf. on Computer and Robot Vision (CRV)*, IEEE, May 2007, pp. 488–494, ISBN: 978-0-7695-2786-4.

[41] Z. Yang, H. Ai, B. Wu, S. Lao, and L. Cai, "Face pose estimation and its application in video shot selection," in *Proc. 17th Intl. Conf. on Pattern Recognition (ICPR)*, IEEE, 2004, 322–325 Vol.1, ISBN: 978-0-7695-2128-2.

[42] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, "Biometric Quality: Review and Application to Face Recognition with FaceQnet," Jun. 2020. arXiv: 2006.03298.

[43] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data Uncertainty Learning in Face Recognition," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2020. arXiv: 2003.11339.

[44] 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," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, Mar. 2020. arXiv: 2003.09373.

[45] J. Rose and T. Bourlai, "On Designing a Forensic Toolkit for Rapid Detection of Factors that Impact Face Recognition Performance When Processing Large Scale Face Datasets," in *Securing Social Identity in Mobile Platforms: Technologies for Security, Privacy and Identity Management*, ser. Advanced Sciences and Technologies for Security Applications, Springer International Publishing, 2020, pp. 61–76, ISBN: 978-3-030-39489-9.

[46] X. Zhao, Y. Li, and S. Wang, "Face Quality Assessment via Semi-supervised Learning," in *Proc. 8th Intl. Conf. on Computing and Pattern Recognition (ICCPR)*, ACM, Oct. 2019, pp. 288–293, ISBN: 978-1-4503-7657-0.

[47] Z. Lijun, S. Xiaohu, Y. Fei, D. Pingling, Z. Xiangdong, *et al.*, "Multi-branch Face Quality Assessment for Face Recognition," in *Proc. 19th Intl. Conf. on Communication Technology (ICCT)*, IEEE, Oct. 2019, pp. 1659–1664, ISBN: 978-1-72810-535-2.

[48] J. Rose and T. Bourlai, "Deep learning based estimation of facial attributes on challenging mobile phone face datasets," in *Proc. Intl. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*, ACM, Aug. 2019, pp. 1120–1127, ISBN: 978-1-4503-6868-1.

[49] Y. Shi and A. K. Jain, "Probabilistic Face Embeddings," *Proc. Intl. Conf. on Computer Vision (ICCV)*, Aug. 2019. arXiv: 1904.09658.

[50] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "FaceQnet: Quality Assessment for Face Recognition based on Deep Learning," *Proc. 12th IAPR Intl. Conf. on Biometrics (ICB)*, Apr. 2019. arXiv: 1904.01740.

[51] F. Yang, X. Shao, L. Zhang, P. Deng, X. Zhou, *et al.*, "DFQA: Deep Face Image Quality Assessment," in *Proc. 10th Intl. Conf. on Image and Graphics (ICIG)*, vol. 11902, Springer International Publishing, 2019, pp. 655–667, ISBN: 978-3-030-34109-1 978-3-030-34110-7.

[52] P. Wasnik, R. Raghavendra, K. Raja, and C. Busch, "An Empirical Evaluation of Deep Architectures on Generalization of Smartphone-based Face Image Quality Assessment," in *Proc. 9th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, Oct. 2018.

[53] J. Yu, K. Sun, F. Gao, and S. Zhu, "Face biometric quality assessment via light CNN," *Pattern Recognition Letters*, vol. 107, pp. 25–32, May 2018, ISSN: 01678655.

[54] L. Best-Rowden and A. K. Jain, "Automatic Face Image Quality Prediction," Jun. 2017. arXiv: 1706.09887 [cs].

[55] L. Zhang, L. Zhang, and L. Li, "Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model," in *Intl. Conf. on Neural Information Processing (ICONIP)*, vol. 10636, Springer International Publishing, 2017, pp. 583–593, ISBN: 978-3-319-70089-2 978-3-319-70090-8.

[56] J. Chen, Y. Deng, G. Bai, and G. Su, "Face Image Quality Assessment Based on Learning to Rank," *IEEE Signal Processing Letters (SPL)*, vol. 22, no. 1, pp. 90–94, Jan. 2015, ISSN: 1070-9908, 1558-2361.

[57] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC TR 29794-5:2010 Information technology - Biometric sample quality - Part 5: Face image data*, International Organization for Standardization, 2010.

[58] M. Moreno-Moreno, J. Fierrez, and J. Ortega-Garcia, "Biometrics beyond the visible spectrum: Imaging technologies and applications," in *Proc. of BioID-MultiComm*, ser. LNCS, vol. 5707, Springer, Sep. 2009, pp. 154–161.

[59] J. Long and S. Li, "Near Infrared Face Image Quality Assessment System of Video Sequences," in *Proc. 6th Intl. Conf. on Image and Graphics (ICIG)*, IEEE, Aug. 2011, pp. 275–279, ISBN: 978-1-4577-1560-0.

[60] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 2382-37:2017 Information technology - Vocabulary - Part 37: Biometrics*, International Organization for Standardization, 2017.

[61] A. Dutta, R. Veldhuis, and L. Spreeuwers, "Predicting Face Recognition Performance Using Image Quality," Oct. 2015. arXiv: 1510.07119 [cs].

[62] P. Grother, A. Hom, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Assessment (3rd Draft)," National Institute of Standards and Technology, Tech. Rep., Jul. 2020, p. 33.

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

[64] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality measures in biometric systems," *IEEE Security & Privacy*, vol. 10, no. 6, pp. 52–62, Dec. 2012, ISSN: 1540-7993.

[65] A. Pena, J. Fierrez, A. Lapedriza, and A. Morales, "Learning emotional-blinded face representations," in *IAPR Intl. Conf. on Pattern Recognition (ICPR 2020)*, Jan. 2021.

[66] R. Daza, A. Morales, J. Fierrez, and R. Tolosana, "mEBAL: A Multimodal Database for Eye Blink Detection and Attention Level Estimation," Jun. 2020. arXiv: 2006.05327 [cs].

[67] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition

in the Wild: Recent Works, Annotation, and COTS Evaluation," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, Aug. 2018, ISSN: 1556-6021.

[68] S. Bharadwaj, M. Vatsa, and R. Singh, "Biometric quality: A review of fingerprint, iris, and face," *EURASIP Journal on Image and Video Processing (JIVP)*, p. 28, Jul. 2014, ISSN: 1687-5281.

[69] J. Galbally, P. Ferrara, R. Haraksim, A. Psyllos, and L. Beslay, *JRC-34751 - Study on Face Identification Technology for its Implementation in the Schengen Information System*. Publication Office of the European Union, 2019, ISBN: 978-92-76-08843-1.

[70] Y. Liu, J. Yan, and W. Ouyang, "Quality Aware Network for Set to Set Recognition," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4694–4703, 2017, ISSN: 1063-6919. arXiv: 1704.03373.

[71] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 39794-5:2019 Information technology - Extensible biometric data interchange formats - Part 5: Face image data*, International Organization for Standardization, 2019.

[72] H. Proença, M. Nixon, M. Nappi, E. Ghaleb, G. Özbulak, *et al.*, "Trends and Controversies," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 41–67, May 2018, ISSN: 1541-1672, 1941-1294.

[73] ISO/IEC JTC1 SC17 WG3, "Portrait Quality - Reference Facial Images for MRTD," International Civil Aviation Organization, Tech. Rep., Apr. 2018, p. 85.

[74] International Civil Aviation Organization, *Machine Readable Passports – Part 9 – Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs*, http://www.icao.int/publications/Documents/9303_p9_cons_en.pdf, 2015.

[75] ISO/IEC JTC1 SC37 Biometrics, *ISO/IEC 19794-5:2011 Information technology - Biometric data interchange formats - Part 5: Face image data*, International Organization for Standardization, 2011.

[76] Y. Song, J. Zhang, L. Gong, S. He, L. Bao, *et al.*, "Joint Face Hallucination and Deblurring via Structure Generation and Detail Enhancement," *Intl. Journal of Computer Vision (IJCV)*, vol. 127, no. 6-7, pp. 785–800, Jun. 2019, ISSN: 0920-5691, 1573-1405.

[77] K. Grm, W. J. Scheirer, and V. Štruc, "Face Hallucination Using Cascaded Super-Resolution and Identity Priors," *IEEE Trans. on Image Processing (TIP)*, vol. 29, pp. 2150–2165, 2020, ISSN: 1941-0042.

[78] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental Face Alignment in the Wild," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2014, pp. 1859–1866, ISBN: 978-1-4799-5118-5.

[79] L. Didaci, G. L. Marcialis, and F. Roli, "Analysis of unsupervised template update in biometric recognition systems," *Pattern Recognition Letters*, vol. 37, pp. 151–160, Feb. 2014, ISSN: 01678655.

[80] H. S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, A. Noore, *et al.*, "On co-training online biometric classifiers," in *Intl. Joint Conf. on Biometrics (IJCB)*, Oct. 2011, pp. 1–7.

[81] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez, "Quality-Based Conditional Processing in Multi-Biometrics: Application to Sensor Interoperability," *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 6, pp. 1168–1179, Nov. 2010, ISSN: 1558-2426.

[82] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple classifiers in biometrics. Part 2: Trends and challenges," *Information Fusion*, vol. 44, pp. 103–112, Nov. 2018, ISSN: 1566-2535.

[83] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187–205, Dec. 2019, ISSN: 15662535.

[84] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics Systems Under Spoofing Attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, Sep. 2015, ISSN: 1053-5888.

[85] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Trans. on Image Processing*, vol. 23, no. 2, pp. 710–724, 2014.

[86] P. Drozdowski, C. Rathgeb, and C. Busch, "Computational workload in biometric identification systems: An overview," *IET Biometrics*, vol. 8, no. 6, pp. 351–368, Nov. 2019, ISSN: 2047-4938.

[87] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, Apr. 2007.

[88] M. Olsen, V. Šmida, and C. Busch, "Finger image quality assessment features - definitions and evaluation," *IET Biometrics*, vol. 5, no. 2, pp. 47–64, Jun. 2016.

[89] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *Intl. Journal of Computer Vision (IJCV)*, vol. 42, no. 3, pp. 145–175, 2001, ISSN: 0920-5691.

[90] F. Weber, "Some quality measures for face images and their relationship to recognition performance," *NIST Biometric Quality Workshop*, 2006.

[91] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, Mar. 2002, ISSN: 1070-9908, 1558-2361.

[92] G. D. Lowe., "Distinctive image features from scale-invariant keypoints," *Intl. Journal of Computer Vision*, vol. 60, pp. 91–110, 2 2004, ISSN: 0920-5691.

[93] P. T. Yap and P. Raveendran, "Image focus measure based on Chebyshev moments," *IEE Proc. - Vision, Image, and Signal Processing*, vol. 151, no. 2, pp. 128–136, 2004, ISSN: 1350-245X.

[94] P. J. Bex and W. Makous, "Spatial frequency, phase, and the contrast of natural images," *JOSA A*, vol. 19, no. 6, pp. 1096–1106, Jun. 2002, ISSN: 1520-8532.

[95] S. Bezryadin, P. Bourov, and D. Ilinih, "Brightness Calculation in Digital Image Processing," *Intl. Symposium on Technologies for Digital Photo Fulfillment*, vol. 2007, no. 1, pp. 10–15, Jan. 2007, ISSN: 21694664, 21694672.

[96] Y. Yao, B. R. Abidi, N. D. Kalka, N. A. Schmid, and M. A. Abidi, "Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement," *Computer Vision and Image Understanding*, vol. 111, no. 2, pp. 111–125, Aug. 2008, ISSN: 10773142.

[97] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. 511–5181.

[98] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural Features for Image Classification," vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, ISSN: 0018-9472, 2168-2909.

[99] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley, "Head Pose Estimation on Low Resolution Images," in *Multimodal Technologies for Perception of Humans*, vol. 4122, Springer Berlin Heidelberg, 2007, pp. 270–280, ISBN: 978-3-540-69567-7 978-3-540-69568-4.

[100] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015. arXiv: 1409.1556 [cs].

[101] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 2818–2826, ISBN: 978-1-4673-8851-1.

[102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks,"

Communications of the ACM, vol. 60, no. 6, pp. 84–90, May 2017, ISSN: 0001-0782, 1557-7317.

[103] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, *et al.*, "Going Deeper with Convolutions," Sep. 2014. arXiv: 1409.4842 [cs].

[104] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015. arXiv: 1512.03385 [cs].

[105] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, *et al.*, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Nov. 2016. arXiv: 1602.07360 [cs].

[106] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, *et al.*, "SqueezeNext: Hardware-Aware Neural Network Design," Aug. 2018. arXiv: 1803.10615 [cs].

[107] X. Wu, R. He, Z. Sun, and T. Tan, "A Light CNN for Deep Face Representation with Noisy Labels," Aug. 2018. arXiv: 1511.02683 [cs].

[108] M. Lin, Q. Chen, and S. Yan, "Network In Network," Mar. 2014. arXiv: 1312.4400 [cs].

[109] L. Best-Rowden and A. K. Jain, "Learning Face Image Quality From Human Assessments," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 12, pp. 3064–3077, Dec. 2018, ISSN: 1556-6013, 1556-6021.

[110] D. Wang, C. Otto, and A. K. Jain, "Face Search at Scale: 80 Million Gallery," Jul. 2015. arXiv: 1507.07242 [cs].

[111] NEUROtechnology. (2020). Face biometrics, [Online]. Available: http://www.neurotechnology.com/face-biometrics.html.

[112] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain," *IEEE Trans. on Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012, ISSN: 1057-7149, 1941-0042.

[113] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Trans. on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012, ISSN: 1057-7149, 1941-0042.

[114] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Mar. 2019. arXiv: 1801.04381 [cs].

[115] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Jan. 2018. arXiv: 1608.06993 [cs].

[116] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning Transferable Architectures for Scalable Image Recognition," Apr. 2018. arXiv: 1707.07012 [cs, stat].

[117] F. C. C. O. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Apr. 2017. arXiv: 1610.02357 [cs].

[118] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," *Proc. 14th European Conf. on Computer Vision*, Jul. 2016. arXiv: 1607.08221.

[119] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, Jun. 2015. arXiv: 1503.03832.

[120] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," *Intl. Conf. on Automatic Face and Gesture Recognition*, May 2018. arXiv: 1710.08092.

[121] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," May 2014. arXiv: 1312.6114 [cs, stat].

[122] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, *et al.*, "ImageNet: A Large-Scale Hierarchical Image Database," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[123] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[124] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, Mar. 2013, ISSN: 1070-9908, 1558-2361.

[125] V. N, P. D, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *2015 Twenty First National Conf. on Communications (NCC)*, Feb. 2015, pp. 1–6.

[126] E. Eidinger, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," *IEEE Trans. on Information Forensics and Security*, vol. 9, no. 12, pp. 2170–2179, Dec. 2014, ISSN: 1556-6021.

[127] CyberExtruder. (2020). Ultimate face matching data set, [Online]. Available: https : / / cyberextruder . com / face - matching-data-set-download/.

[128] N. D. Kalka, B. Maze, J. A. Duncan, K. OrConnor, S. Elliott, *et al.*, "IJB-S: IARPA Janus Surveillance Video Benchmark," in *9th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, Oct. 2018, pp. 1–9, ISBN: 978-1-5386-7180-1.

[129] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action (FIA) Database," *2nd Intl. Workshop on Analysis and Modelling of Faces and Gestures (AMFG)*, pp. 255–263, 2005.

[130] National Cheng Kung University. (2020). NCKU face database, [Online]. Available: http://robotics.csie.ncku.edu.tw/Databases/FaceDetect_PoseEstimate.htm.

[131] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding (CVIU)*, vol. 99, no. 3, pp. 303–331, Sep. 2005, ISSN: 10773142.

[132] C. E. Thomaz and G. A. Giraldi, "A new ranking method for principal components analysis and its application to face image analysis," *Image and Vision Computing*, vol. 28, no. 6, pp. 902–913, Jun. 2010, ISSN: 02628856.

[133] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[134] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization," in *Proc. Intl. Conf. on Computer Vision Workshops (ICCVW)*, IEEE, Nov. 2011, pp. 2144–2151.

[135] D. O. Gorodnichy, "Video-based framework for face recognition in video," in *Proc. 2nd Canadian Conf. on Computer and Robot Vision (CRV)*, May 2005, pp. 330–338.

[136] N. Gourier and J. Letessier, "The Pointing'04 Data Sets," in *Proc. Pointing 2004 Intl. Workshop on Visual Observation of Deictic Gestures*, 2004.

[137] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," *Proc. 2nd Intl. Conf. on Audio and Video-based Biometric Person Authentication (AVBPA)*, p. 6, 1999.

[138] F. Solina, P. Peer, B. Batagelj, S. Juvan, and J. Kovac, "Color-based face detection in the "15 seconds of fame" art installation," *Proc. Intl. Conf. Mirage 2003*, p. 10, 2003.

[139] P. T. Duizer, D. M. Hansen, and T. B. Moeslund, "Data set: Head direction," HERMES project (FP6 IST-027110), Tech. Rep. CVMT-07-02. [Online]. Available: http://www.cvmt.dk/projects/Hermes/head-data.html.

[140] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th Intl. Conf. on Automatic Face and Gesture Recognition*, IEEE, 2000, pp. 46–53, ISBN: 978-0-7695-0580-0.

[141] S. Mandala, "The effect of lighting direction on face recognition performance," Master Thesis, West Virginia University, 2005.

[142] 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., Oct. 2007.

[143] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000, ISSN: 01628828.

[144] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, *et al.*, "The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations," *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans (TSMCA)*, vol. 38, no. 1, pp. 149–161, Jan. 2008, ISSN: 1558-2426.

[145] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning Face Representation from Scratch," Nov. 2014. arXiv: 1411.7923 [cs].

[146] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, *et al.*, "Overview of the Face Recognition Grand Challenge," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, IEEE, Jun. 2005, pp. 947–954.

[147] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2011, pp. 529–534, ISBN: 978-1-4577-0394-2.

[148] K. C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, May 2005, ISSN: 1939-3539.

[149] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 1931–1939, ISBN: 978-1-4673-6964-0.

[150] M. Grgic, K. Delac, and S. Grgic, "SCface - surveillance cameras face database," *Multimedia Tools and Applications*, vol. 51, no. 3, pp. 863–879, 2011, ISSN: 1380-7501, 1573-7721.

[151] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Workshop on Applications of Computer Vision (ACV)*, IEEE, 1994, pp. 138–142, ISBN: 978-0-8186-6410-6.

[152] E. Bailly-Bailliére, S. Bengio, F. E. D. E. R. Bimbot, M. Hamouz, J. Kittler, *et al.*, "The BANCA Database and Evaluation Protocol," in *Audio- and Video-Based Biometric Person Authentication (AVBPA)*, vol. 2688, Springer Berlin / Heidelberg, 2003, pp. 625–638, ISBN: 978-3-540-40302-9 978-3-540-44887-7.

[153] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression Database," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003, ISSN: 0162-8828.

[154] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, *et al.*, "FRVT 2006 and ICE 2006 large-scale results," National Institute of Standards and Technology, Tech. Rep. NIST IR 7408, 2007.

[155] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, *et al.*, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *Intl. Conf. on Automatic Face and Gesture Recognition*, IEEE, Mar. 2011, pp. 346–353, ISBN: 978-1-4244-9140-7.

[156] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, Jun. 2001, ISSN: 01628828.

[157] A. Martinez and R. Benavente, "The AR Face Database," *CVC Technical Report #24*, Jun. 1998.

[158] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. R. Freire, *et al.*, "The multiscenario multienvironment BioSecure multimodal database (BMDB)," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1097–1111, 2010, ISSN: 0162-8828.

[159] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, *et al.*, "Frontal to profile face verification in the wild," in *Winter Conf. on Applications of Computer Vision (WACV)*, IEEE, Mar. 2016, pp. 1–9, ISBN: 978-1-5090-0641-0.

[160] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, *et al.*, "IARPA Janus Benchmark - C: Face Dataset and Protocol," in *Proc. 11th IAPR Intl. Conf. on Biometrics (ICB)*, IEEE, Feb. 2018, pp. 158–165, ISBN: 978-1-5386-4285-6.

[161] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, *et al.*, "Overview of the Multiple Biometrics Grand Challenge," in *Proc. 3rd Intl. Conf. on Biometrics (ICB)*, ser. Lecture Notes in Computer Science, Springer, 2009, pp. 705–714, ISBN: 978-3-642-01793-3.

[162] A. P. Founds, N. Orlans, G. Whiddon, and C. Watson, "NIST Special Databse 32 - Multiple Encounter Dataset II (MEDS-II)," National Institute of Standards and Technology, Tech. Rep. NIST IR 7807, 2011.

[163] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv: 1512.00596.

[164] J. R. Beveridge, K. W. Bowyer, P. J. Flynn, S. Cheng, P. J. Phillips, *et al.*, "The challenge of face recognition from digital point-and-shoot cameras," in *6th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, Sep. 2013, pp. 1–8, ISBN: 978-1-4799-0527-0.

[165] P. A. Johnson, P. Lopez-Meyer, N. Sazonova, F. Hua, and S. Schuckers, "Quality in face and iris research ensemble (Q-FIRE)," in *Proc. 4th Intl. Conf. on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, Sep. 2010, pp. 1–6, ISBN: 978-1-4244-7581-0.

[166] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated machine learning: Methods, systems, challenges*. Springer, 2018, In press, available at http://automl.org/book.

[167] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, Dec. 2014, ISSN: 2169-3536.

[168] J. Galbally, C. McCool, J. Fierrez, S. Marcel, and J. Ortega-Garcia, "On the vulnerability of face verification systems to hill-climbing attacks," *Pattern Recognition*, vol. 43, no. 3, pp. 1027–1038, Mar. 2010, ISSN: 00313203.

[169] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, *et al.*, "Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics," in *AAAI Workshop on Artificial Intelligence Safety (SafeAI)*, Feb. 2020.

[170] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *Trans. on Technology and Society (TTS)*, vol. 1, no. 2, Jun. 2020, ISSN: 2637-6415.

[171] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, ISSN: 15662535.

[172] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: Learning Agnostic Representations with Application to Face Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.

[173] S. Pidhorskyi, D. Adjeroh, and G. Doretto, "Adversarial Latent Autoencoders," Apr. 2020. arXiv: 2004.04467 [cs].