Face Image Quality Assessment: A Literature Survey

TORSTEN SCHLETT and CHRISTIAN RATHGEB, da/sec - Biometrics and Internet Security
Research Group, Hochschule Darmstadt
OLAF HENNIGER, Fraunhofer Institute for Computer Graphics Research IGD
JAVIER GALBALLY, European Commission, Joint Research Center
JULIAN FIERREZ, Universidad Autonoma de Madrid
CHRISTOPH BUSCH, da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt

210

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 detect low-quality data and make decisions accordingly. This survey provides an overview of the face image quality assessment literature, which predominantly focuses on visible wavelength face image input. A trend towards deep learning-based methods is observed, including notable conceptual differences among the recent approaches, such as the integration of quality assessment into face recognition models. 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.

CCS Concepts: • **Security and privacy → Biometrics;**

Additional Key Words and Phrases: Biometric sample quality, face image quality assessment, face recognition

1 INTRODUCTION

Face Image Quality Assessment (FIQA) refers to the process of taking a face image as input to produce some form of "quality" estimate as output, as illustrated in Figure 1. A **FIQA algorithm (FIQAA)** is an automated FIQA approach. See Figure 2 for some example images with varying quality. While FIQA and general **Image Quality Assessment (IQA)** are overlapping research areas, there are important distinctions, which we discuss in Section 2.2. Most of the published FIQA literature focuses on single face image input in the visible spectrum. Therefore, unless otherwise specified in this survey, FIQA(A) refers to single-image Face Image Quality Assessment (Algorithms) in the visible spectrum, with a **Quality Score (QS)** [66] output that can be represented by: (A) a single scalar value or (B) a vector of quality values measuring different quality-related features. For a discussion of (F)IQA that instead compares two image variants, i.e., full/reduced-reference methods, see Section 2.3. Regarding FIQA outside the visible spectrum, see Section 6.7.

The term "quality" is an intrinsically subjective concept that can be defined in different ways, with ISO/IEC 29794-1 [68] differentiating between three aspects referred to as character, fidelity, and utility. In the context of facial biometrics these can be described as follows [69]:

- **Character:** Attributes inherent to the source biometric characteristic being acquired (e.g., the face topography or skin texture) that cannot be controlled during the biometric acquisition process (e.g., scars) [70].
- **Fidelity:** For a biometric sample [70], e.g., a face image, fidelity reflects the degree of similarity to its source biometric characteristic [68]. For instance, a blurred image of a face omits detail and has low fidelity [67].
- **Utility:** The fitness of a sample to accomplish or fulfill the biometric function (e.g., face recognition comparison), which is influenced, i.a., by the character and fidelity [70]. Thus, the term utility is used to indicate the value of an image to a receiving algorithm [67].

This survey considers "utility" as the primary definition of what a quality score should convey, which is in accordance to the quality score definition of ISO/IEC 2382-37 [70] and the definition in the ongoing **Face Recognition Vendor Test (FRVT)** for face image quality assessment [67]. Thus, a QS should be indicative of the **Face Recognition (FR)** performance. Note that this entails that the output of a specific FIQAA may be more accurate for a specific FR system, so the FIQA utility prediction effectivity ultimately depends on the combination of both, the FIQAA and the FR system. To facilitate interoperability, it is, however, desirable that the FIQAA 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, a FIQAA 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 FIQA useful for a variety of scenarios, which are described further in Section 2.5. FIQA as a predictor for FR performance has attracted the predominant interest of researchers so far and is thus the main focus in the present survey. FIQA for other tasks in the field of face biometrics, such as emotion analysis [71], attention-level estimation [72], gender or other soft biometrics recognition [73], and so on, may open interesting research lines in the future and can take advantage of current developments that employ FIQA for FR performance prediction.

The contributions of this survey are:

- An introduction to FIQA (Section 2), i.a., including the distinction against general IQA (Section 2.2), the conceptual problem with single-image utility assessment (Section 2.4), and an overview of both common and uncommon FIQA application areas (Section 2.5).
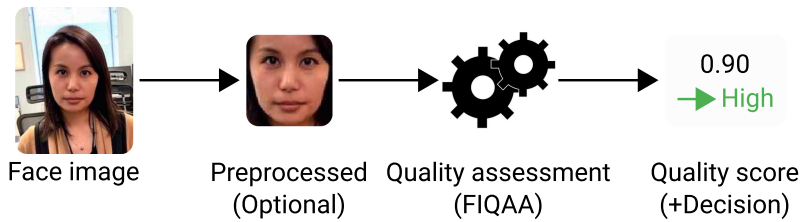
Fig. 1. Typical FIQA (Face Image Quality Assessment) process: A face image is preprocessed and a FIQAA (FIQA algorithm) is applied, resulting in a scalar quality score output, based on which a decision can be made. Face image taken from Reference [67].

Fig. 2. Face images of a single subject with various qualities. Face image quality degrades from left to right as quality degradation factors such as facial expression, pose, and illumination are introduced. Face images taken from Reference [67].

- A categorization of the surveyed FIQA approaches (Section 3) with a taxonomy that differentiates between factor-specific and monolithic approaches, in addition to various other aspects (Figure 6).
- A survey of more than 60 FIQAA publications from 2004 to 2021 (Section 4), including condensed overview tables for the publications (Tables 3 and 4) and their used datasets (Table 2). This part is meant for literature overview purposes and does not have to be read in sequence.
  Prior work listed varying publication numbers, with Hernandez-Ortega et al. [48] being a recent example that contained a summary for some prior publications ranging from 2006 to 2020. A fingerprint/iris/face quality assessment survey by Bharadwaj et al. [74] considered less than 10 FIQAA publications from 2005 to 2011. The European JRC-34751 report [75] also listed some FIQAAs from 2007 to 2018. To our knowledge, this FIQA survey is the most comprehensive one to date.
- An introduction for the **Error-versus-Reject-Characteristic (ERC)** evaluation methodology (Section 5.1), which is a standardization candidate in addition to being commonly used in recent FIQA literature, and a subsequent concrete evaluation that includes a variety FIQA approaches (Section 5.2). The ERC introduction mentions details not considered in recent FIQA literature, and the evaluation discusses its weaknesses to note opportunities and challenges for future work.
- A detailed discussion of various FIQA issues and challenges (Section 6), including avenues for future work.

Table 1 should allow readers with different intent and background knowledge to quickly identify the most relevant parts of this survey.

2 QUALITY ASSESSMENT IN FACE RECOGNITION

During enrollment, a classical face recognition system acquires a reference face image from an individual, proceeds to pre-process it, including the step of face detection, and finally extracts a

Table 1. Most Relevant Survey Parts for Readers with Different Intent and Knowledge Background

Intent of knowledge acquisition	Knowledge background	Relevant parts
Basics (definition, goal, etc.)	Non-expert	Section 1
Concepts and categorization (input data, training data, etc.)	Expert	Sections 2.1 to 2.4 and 3
Applications (use-cases in automated systems)	Non-expert	Section 2.5
Overview of published works (coarse)	Expert	Sections 4.1, 4.3, and 7; Tables 2, 3, and 4
Survey of published works (detailed)	Expert	Sections 4.2 and 4.4
Comparison and evaluation (selective comparison, metrics, etc.)	Expert	Section 5
Open issues and challenges (research directions, problems, etc.)	Non-expert	Sections 6 and 7
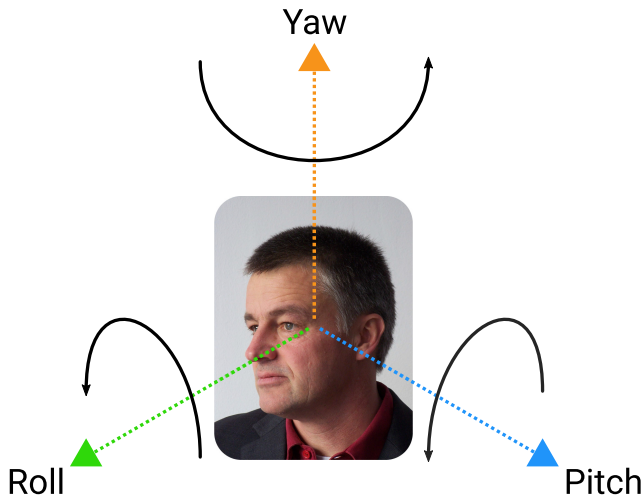
Fig. 3. Facial pose is usually represented by the pitch, yaw, and roll angles defined by ISO/IEC 39794-5 [76]. Pitch and yaw are also known as tilt and pan. A frontal face has 0° for all three angles.

set of features that 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 [70] for the standardized vocabulary definitions of terms such as enrollment, templates, or references.

2.1 Controlled and Unconstrained Acquisition

Regarding the face image acquisition [70], two different scenarios can be distinguished [75]:

- **Controlled:** In a controlled scenario, the biometric capture subject is cooperative [70], so, 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 [70] or intentionally uncooperative [70], and there is no control over the environmental conditions. Surveillance video FR is an example for this scenario [77].

There are other scenarios in between those two extremes, e.g., smartphone FR with a cooperative capture subject but incomplete control over the environment [75], and the literature usually refers

to close-to-optimal capture conditions as "controlled," with anything else falling under the "unconstrained" category [75]. FIQA 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, FIQA can be used to filter out images below a certain quality level. While the same FIQAA 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, FIQA 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, FIQA 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. However, explainable FIQA can also be beneficial when images are analyzed after the acquisition process is complete. Hence, using FIQA for actionable feedback during a controlled acquisition is just one important application scenario, while other use cases are independent of the acquisition type.

2.2 FIQA versus IQA

FIQA can be seen as a specific application within the wider field of **Image Quality Assessment (IQA)**, which is a very active research area of image processing. Even though related to IQA, FIQA has been mainly developed within the biometric context and focuses on distinctive face features. Consequentially, general **IQA algorithms (IQAA)** have shown poor performance when directly applied to FIQA, and, conversely, the very specific FIQA algorithms usually do not generalize to the broader application field of IQA.

General non-biometric IQA typically aims to assess images in terms of subjective (human) perceptual quality, meaning that technically objective quality scores generated by such IQAAs usually intent to predict or model subjective perceptual quality [78].

Biometric FIQA, however, is usually concerned with the assessment of the biometric utility for facial biometrics, which can be objectively defined in the context of specific FR systems. FIQA works may also test or train FIQAAs using ground truth data stemming from human quality assessments, but for biometric purposes the intent still differs from general perceptual quality assessment, insofar that the question is how well the images can be used for facial biometrics, versus how good/undistorted the images look overall for a human.

It can be expected that perceptual quality and biometric utility coincide to some degree, thus general IQA can be utilized for FIQA as well. The reverse is less likely, since FIQA algorithms may be specifically developed for face images, so results for non-face images are not expected to be useful. This also means that FIQA can perform better for the purpose of biometric utility prediction than a general IQA that has not been developed with facial biometrics in mind. Some of the surveyed FIQA literature tested known IQA algorithms together with specialized FIQA algorithms. For instance, Terhörst et al. [50] tested the general IQAAs BRISQUE [79], NIQE [80], and PIQE [81]) together with their fully FR-specialized SER-FIQ FIQAA and three other FIQAAs.

2.3 Full/Reduced/No-reference Quality Assessment

IQA literature draws a distinction between approaches that require a "reference" version of the input and those that do not [48, 54, 74] (not to be confused with biometric references [70], e.g., in an FR database):

- **Full-reference:** IQA 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
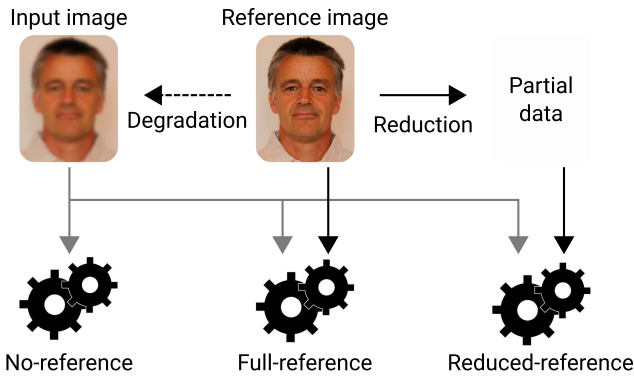
Fig. 4. Full-reference, reduced/partial-reference, and no-reference quality assessment approaches differ in the used input data, as described in Section 2.

input image can be seen as a potentially degraded (e.g., blurred) version of the reference image.

- **Reduced-reference/Partial-reference:** Similar to full-reference IQA, a reference version of the input image has to exist first, but only incomplete information of the reference is known and used for the IQA, e.g., some statistics of the image. The distinction between full-reference and reduced-reference approaches is not necessarily clear, since full-reference approaches may also "reduce" their input to a different representation, with information loss, before the comparison step.
- **No-reference:** No reference version of the input image is required for the IQA. Note that such an IQAA can still use other forms of internal data: An IQAA could, e.g., utilize some fixed set of images unrelated to the input image and still be categorized as no-reference IQA. Likewise, machine learning IQA models are not automatically classified as reduced-reference IQA just because they incorporate information from training images.

See Figure 4 for an illustration of the three concepts. Full- or reduced-reference approaches are more common and viable for IQA than for FIQA, since both an original and a degraded image exists, e.g., for an image or video compression scenario [78] (a use case neglected by FIQA literature so far). Almost all of the published FIQA literature more specifically considered single-image input FIQA approaches, which implies no-reference FIQA, and means that no other data specific to the corresponding person (or biometric capture subject [70]) is required to facilitate the FIQA. An outlier is the recent work from Dihin et al. [82], which does consider multiple full-reference IQAAs for face images, for both FIQA and for FR. Note that any FR comparison method can technically fall under the definition of full/reduced-reference (F)IQA if the comparison scores are repurposed as quality scores. Furthermore, any full/reduced-reference (F)IQA method can technically be used as a no-reference method if an image degradation function is added, such that the single input image serves as the unmodified "reference" as well as the degraded input. Obviously this has less potential for FIQA than specialized approaches. Nonetheless, this idea has in fact been applied to utilize full-reference IQA for single-image face **presentation attack detection (PAD)**. A prominent example for this is the work by Galbally and Marcel [83], which incorporated various full-reference IQAAs and applied Gaussian filtering as the degradation function, using the IQAA output to classify the input image as either genuine or as a presentation attack. Many of these PAD works that are utilizing full-reference IQA appear to use similar IQAA configurations, and neither FIQA nor FR is their primary concern, so we do not reference more herein.

2.4 The Quality Paradox

Usually FIQA algorithms are intended to predict biometric utility for a single biometric sample, meaning that a single quality score is produced for a single image. Predicting biometric utility in the context of face recognition implies that the quality score has to indicate the "accuracy" or "certainty" of comparison scores generated for a sample pair that includes the assessed sample. Thus, a FIQAA only receives a single sample S, which is also part of one or more comparisons with other samples unknown to the FIQAA during the assessment of sample S. This conceptual problem is referred to as the "quality paradox." How FIQA approaches are affected by this quality paradox differs with the concepts:

- FIQA approaches that only repurpose general IQA methods are already inherently not conceptually linked to FR utility, i.e., independently of the quality paradox.
- FIQA approaches trained on ground truth QSs do have to consider the quality paradox when the ground truth QSs are generated:
  - Relying on human-defined ground truth QSs will generally depend on the subjective assessments, again technically independent of the quality paradox, except for human quality assessments that are guided by some protocol (e.g., collective human FIQA via pairwise comparisons in Reference [57]).
  - For FR-derived ground truth QSs the quality paradox becomes fully relevant, since the FR comparison pairs have to be selected and the pairwise FR comparison scores have to be transformed into QSs per sample. Thus, the task of deriving the ground truth QSs itself becomes important to the FIQA design. Some recent examples of differing ground truth generation approaches are:
    * FaceQnet v0 [53]: Normalized comparison score between a target sample and a mated ICAO-compliant (i.e., assumed high quality) sample as the target sample ground truth QS.
    * FaceQnet v1 [48]: Extended the FaceQnet v0 [53] approach by score fusion for multiple FR systems.
    * PCNet [47]: FIQA model training with loss as the squared difference between the minimum of the predicted per-sample QS for a mated pair of samples and a corresponding FR comparison score.
    * SDD-FIQA [45]: Computed the ground truth QS per sample as the Wasserstein distance between FR comparison score sets for randomly selected mated and non-mated pairs (that each include the sample).
- There also exist FIQA approaches that directly use FR models during training/inference without ground truth QS generation, and approaches that unify FR/FIQA in one model. While these approaches still technically have to contend with the limits imposed by the quality paradox for single-sample FIQA, they can more directly estimate the quality (or "certainty") of the feature embeddings that the FR model generates.

The data aspect categorization described in Section 3.2 is especially relevant with respect to these considerations.

2.5 Application Areas of FIQA

There are various use cases for FIQA:

- **Acquisition process threshold:** Face images that result in a quality score below a set threshold can be rejected during the acquisition process [70]. Besides assessing image data stemming directly from cameras, FIQA 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. [20].

- **Acquisition process feedback:** One or multiple FIQAAs 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 [76], ICAO [84, 85], or ISO/IEC 19794-5 [86] can be checked and reported automatically when an image is acquired for FR system enrollment [70] or for passports and other government-issued ID documents. Capture subjects [70] themselves can also receive immediate feedback for possibly less rigid requirements, e.g., during **ABC (Automatic Border Control)** at airports.

- **Quality summarization [87]:** Quality can also be monitored by summarizing it over time, for different capture devices [70] or locations [68], or per user. This, for instance, enables the identification of defective or underperforming capture devices, problematic locations, times of day, or seasonal variations, as well as users that consistently yield low quality samples [87].

- **Video frame selection:** Images in a video sequence can be ranked and selected by their assigned quality scores. 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 of sufficiently high quality may not require enhancement, images with very low quality may not be salvageable by enhancement, and images within a medium quality range may be adequate for enhancement. In addition, multiple enhancement steps could be applied depending on the quality variation 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 introduce a significant computational overhead that could make additional hardware necessary (e.g., GPUs). The former drawback was shown, e.g., for illumination FIQA by Rizo-Rodriguez et al. [23]. Likewise, the FIQA application list of Hernandez-Ortega et al. [48] noted References [88] and [89] as examples for the latter drawback, with Reference [88] listing multiple methods taking seconds to minutes, while Reference [89] states a requirement of 30 ms 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 in an enhancement approach stage of the video-focused method by Nasrollahi and Moeslund [21]. Last, it is also possible to enhance image regions individually, depending on region-specific quality scores, which was done in one approach of Sellahewa and Jassim [65].

- **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 FIQAA literature listed in this survey, it is also possible to employ full/reduced-reference FIQA/IQA for this use case, since a reference is available in the 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 an 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 FIQAA of course needs to be applied beforehand to obtain a **quality score (QS)**. Furthermore, images or templates [70] 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. [48] noted that such updates may also consist of incremental improvements [90, 91], instead of replacements. Besides subject-specific incremental improvements, new quality-controlled data can also be employed to improve biometric models via online learning [74, 92]. Database maintenance, in conjunction with quality summarization/monitoring, is especially relevant in large systems with multiple contributors to a single central database, such as the European **Schengen Information System (SIS)**, the **VISA Information System (VIS)**, the **Entry Exit System (EES)**, or the US **ESTA (Electronic System for Travel Authorization)**.

- **Context switching [48, 74]:** A recognition system can adapt to different quality contexts by switching between multiple recognition algorithm configurations (or modes [70]), using quality assessment for the switch activation [93]. Such a strategy does not necessarily have to be applied to a pure FR system—it could also be devised for a multi-modal biometric system [70].
- **Quality-weighted fusion [48, 74]:** Similar to full context switching, a biometric system can fuse scores or decisions in a weighted fashion based on quality assessments [94, 95]. Quality-based feature-level fusion for face video frames is considered, e.g., in the surveyed literature [11] and [52].
- **Comparison improvement:** Quality can be used directly as part of FR comparisons [70]. For example, Shi and Jain [52] 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, FIQA 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 type of FIQA, which was used by Damer et al. [11].
- **Partial presentation attack avoidance:** Although the surveyed literature does not focus on this application, rejecting or weighing images based on their assessed quality can also reduce opportunities for presentation attacks [70, 96, 97], since accepting images for enrollment or as probe irrespective of their quality could be a potential vulnerability. FIQA or IQA can also be employed specifically for the purpose of **PAD (Presentation Attack Detection)** [98]. Pure FIQA is, however, not meant for comprehensive PAD, because such attacks can consist of data with high biometric utility, too.
- **Progressive identification:** An identification system could conduct searches going progressively from the highest-quality reference templates to the lowest-quality ones. Assuming that these templates vary noticeably in quality and that the search requires an extensive amount of time, such a strategy can 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 a number of matches 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 reduction strategies surveyed by Drozdowski et al. [99], since it relies on the existence of exploitable quality variation in the database. While the listed FIQA literature does not explore this approach, it does consider FIQA-based computational workload reduction in terms of video frame selection. Instead of progressing from highest to lowest quality, Hernandez-Ortega et al. [48] 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.
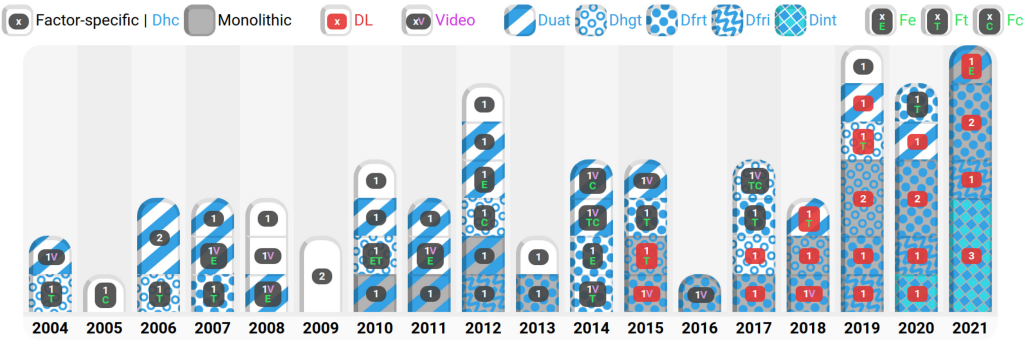
Fig. 5. Timeline of the surveyed FIQA literature with categories as depicted by Figure 6. Numbers in the bars denote literature counts.
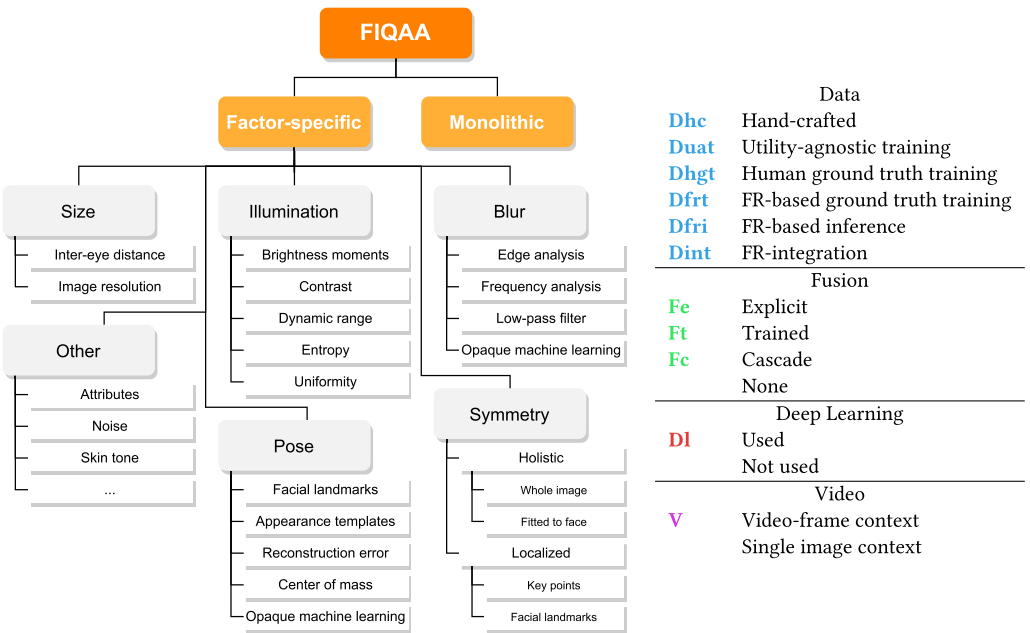
Fig. 6. A taxonomy of the FIQA approaches in the surveyed literature (left), with additional separate aspect-specific categories (right).

3 CATEGORIZATION

The surveyed works are categorized using a taxonomy and several additional aspects. At the highest level, our taxonomy differentiates between factor-specific FIQA approaches and mono-lithic FIQA approaches. The factor-specific taxonomy branch subdivides methods into categories for interpretable (and typically actionable) factors, such as blur, which could help an operator to avoid face image deficiencies in a re-capture attempt. The monolithic approaches produce compar-atively opaque assessments/quality scores, which cannot be immediately interpreted with respect to some concrete separable factor by themselves, but can indicate overall FR utility. As described in

Section 3.1, some of the factor-specific branches can be seen as predominantly capture-related or subject-related. The subsequent Sections 3.2, 3.3, 3.4, and 3.5 describe aspects that are assigned per literature in Tables 3 and 4. Figure 6 shows an overview of both the taxonomy and the per-literature aspect abbreviations. The primary approach commonalities are described together with the corresponding literature references in Sections 4.1 and 4.3.

Note that the taxonomy is meant to group common FIQA approaches in the surveyed literature, it is not meant to enumerate all feasible FIQA concepts. Also note that many of the surveyed works described multiple approaches that belong to different categories of the taxonomy. Some of the surveyed works considered certain quality measure types, but did not specify a concrete approach, and are consequently not present in the method-specific reference lists of the taxonomy-describing text passages (e.g., pose in References [34] or [16]).

3.1 Aspect: Capture- and Subject-related FIQA

ISO/IEC TR 29794-5:2010 [66] includes an informative facial quality classification scheme that distinguishes between static/dynamic subject characteristics/acquisition process properties. At the time of writing a standard ISO/IEC 29794-5 is under development, which will replace the former **Technical Report (TR)**, and it is intended to further categorize its included factor-specific measures as either capture-related or subject-related.

Capture-related FIQA is influenced by circumstances external to the capture subject, such as the used sensor (e.g., camera focus, resolution) or the illumination setup. Subject-related FIQA, conversely, is influenced by the subject, e.g., pose, expression, or movement. While some methods or factors can be predominantly seen as either capture- or subject-related, others are more obviously influenced by a mixture of capture- and subject-related properties. This can be mapped directly to the factor-specific categories used in this survey, instead of individual methods or papers:

- Size - Inter-eye distance: This is subject-related (distance to camera, facial structure). It is technically capture-related as well, since the camera/image resolution is involved, but that typically is a static acquisition property. i.e., it is usually assumed that the camera and its resolution cannot be improved during acquisition, meaning that only the distance to the subject can be adjusted in a re-capture attempt.
- Size - Image resolution: If the considered image was cropped to the face, then the measure is subject-related similar to inter-eye distance. Otherwise, if the camera's full image resolution is assessed, then this factor is fully capture-related.
- Illumination: Illumination is generally seen as a dynamic acquisition process property [66], i.e., capture-related. But measures may be influenced by subject-related properties, too— e.g., facial hair and skin tone (lighter/darker hair/skin), or possibly pose. Conversely, it is of course also possible that illumination conditions happen to be sufficiently extreme to disrupt any primarily subject-related measure.
- Pose: This is predominantly subject-related.
- Blur: Blur is both capture-related and subject-related, since it can be caused by subject/camera motion or improper camera configuration.
- Symmetry: Measures for symmetry depend on symmetric illumination, and most of the surveyed variants implicitly measured frontal pose deviation as well (landmark-based approaches being the exception, although they naturally still rely on a pose that allows landmark detection). Thus, these measures are both capture-related and subject-related.

Monolithic approaches can by definition generally be considered as both capture-related and subject-related.

3.2 Aspect: Data

The following data aspect categories are ordered to reflect the degree of FR(-data)-integration or -utilization, ranging from hand-crafted designs to full FR model integration:

(1) **Dhc** - Hand-crafted: Methods that do not require any training data, except for the optional tuning of parameters such as thresholds. All of the surveyed approaches belonging to this category are factor-specific, such as for example the symmetry and blur measures in Reference [26].

(2) **Duat** - Utility-agnostic training: Methods that require some kind of training data, but do not train to predict ground truth QSs. Pose angle estimation for FIQA is one example where training may be required, but where the training does not intend to directly predict utility. This category also includes approaches that compare the input image against information (e.g., some image statistics) derived from a training set, as long as this comparison does not use an FR system. In this category, a concrete example for a factor-specific approach is the landmark-based pose estimation in Reference [22], and a concrete example for a monolithic approach is Reference [62], which compares the input against a fixed averaged image.

(3) Ground truth QS training: Approaches that are trained using ground truth QSs to predict utility or subjective estimates thereof.

  (a) **Dhgt** - Human ground truth: Works using human assessments for training. The multi-branch deep learning model in Reference [3] is a factor-specific example in this category, and the deep learning model trained on human-derived binary quality labels in Reference [51] is a monolithic example.

  (b) **Dfrt** - FR-based ground truth: Ground truth QSs were derived either via one or multiple FR systems. A recent factor-specific example for this category is the random forest fusion in Reference [1], and a prominent monolithic example is "FaceQnet" [48, 53].

(4) **Dfri** - FR-based inference: Approaches that directly utilize FR models during FIQA model training or inference, without FIQA model training on ground truth QSs. This obviates a distinction between FR-derived and human-defined ground truth QSs, although, e.g., the subject identities of the FR training data may still be specified by humans. The used FR models themselves are not modified with respect to their FR feature inference. All surveyed approaches in this category are monolithic. Recent examples are "SER-FIQ" [50] and "ProbFace" [46].

(5) **Dint** - FR-integration: Hybrid FR/FIQA approaches that simultaneously trained FR and FIQA as part of a single integrated system/model, generating both FR features and quality assessment output during inference. The only surveyed approaches that fall into this category are the recent monolithic "data uncertainty learning" [49] and "MagFace" [44]. Most recently, the latter has also been included in pure evaluation literature [40, 41].

Many surveyed works considered multiple clearly separable approaches. Thus, to minimize clutter in the overview tables, each work is marked only with the highest applicable category as per the list order above, i.e., from **Dhc** to **Dint**.

3.3 Aspect: Fusion

Various works fused multiple separable FIQAAs. Note that only pure FIQAA fusion methods are marked, since some surveyed works included approaches that also incorporated non-FIQAA-derived information into the fusion, such as FR scores [36] or EXIF data [16]. While the output of fusion methods may be similarly opaque to the output of monolithic FIQAAs, their input FIQAAs can be (and often were) factor-specific.

- **Fe** - Explicit: These approaches derived a single QS from the output of the separable FIQAAs by computing weighted sums with manually determined weights [21, 30, 31], or via other hand-crafted fusion functions [12, 19, 23, 42].
- **Ft** - Trained: Trained fusion approaches did likewise include weighted sum computation, except with automatically derived weights [15, 57, 60], but more often relied on various types of machine learning models such as **ANNs (Artificial Neural Networks**, including deep learning) [3, 6, 23, 34, 39], **GMMs (Gaussian Mixture Models)** [14, 33, 39], AdaBoost [10], or random forests [1, 7, 8].
- **Fc** - Cascade: Cascaded approaches [7, 13, 14, 20, 37] combined FIQAAs in multiple stages. Since the cascade algorithm itself was hand-crafted in all surveyed cases, these approaches can be considered as a special kind of explicit fusion. The difference to the other explicit fusion methods is that these approaches can exit the cascade early in each stage if the quality is deemed to be too low. This design can help to reduce the computational workload of the entire quality assessment subsystem when many of the input images are of low quality, e.g., in a video frame selection scenario. While the FIQAAs within the stages are clearly separable, approaches may reuse common data to further improve computational efficiency, as done in Reference [37]. Also, while the per-stage FIQAAs are clearly separable in the sense that they could technically be used as individual FIQAAs, the cascaded **SVM (Support Vector Machine)** approach in References [7, 20] trained binary SVM classifiers specifically for the cascaded combination, which used the early exits to determine a discrete quality level per stage (1 to 5).

3.4 Aspect: Deep Learning

The surveyed FIQA literature can be broadly categorized into works that do not make use of deep learning for FIQA ("non-DL") and works that do ("DL"). Most of the surveyed works overall are non-DL literature, but the majority of the more recent works are DL literature. The trend towards DL-based FIQA research is illustrated by the timeline in Figure 5. In the taxonomy most of the non-DL works belong to the factor-specific branch, while most DL works can be found under the monolithic category. Note that non-DL literature does encompass FIQA approaches based on other kinds of machine learning (including shallow artificial neural networks), as well as purely hand-crafted methods. The DL literature is marked with "**Dl**" in the tables.

3.5 Aspect: Video

While face video quality assessment that used temporal inter-frame information is outside the scope of this face (single-)image quality assessment survey, we do include video-centric literature that used single-image methods to assess isolated video-frames. These works are marked with "**V**" in the tables to distinguish them from the "pure" FIQA works, but be aware that this does not indicate a technical difference of the FIQAAs themselves.

4 FACE IMAGE QUALITY ASSESSMENT ALGORITHMS

The following subsections and tables are divided into the factor-specific and monolithic categories introduced in Section 3. For each there is one subsection that highlights the overarching commonalities/differences (factor-specific Section 4.1, monolithic Section 4.3), followed by a corresponding subsection with introductions for all of the surveyed works in chronological order (factor-specific Section 4.2, monolithic Section 4.4). Table 3 (factor-specific) and Table 4 (monolithic) provide a condensed overview of the literature and show the categorization of the works for every aspect listed in Figure 6. Table 2 additionally lists the datasets used to develop and evaluate the FIQA approaches of the surveyed literature. The implications of the dataset variety are discussed in Section 6.1.

Table 2. Datasets That Were Used in the Literature to Create or Evaluate FIQA Approaches

Dataset	Usage timespan	Used in
LFW [118]	2011 to 2021	17: B [7, 57, 60] E [3, 6, 22, 40, 42–46, 48–50, 52, 54]
FERET [119]	2007 to 2020	9: B [33, 58, 60, 64] C [50] E [12, 19, 22, 26]
VGGFace2 [120]	2019 to 2021	7: B [48, 53] C [47] E [40, 42, 46, 54]
CASIA-WebFace [121]	2017 to 2021	7: B [6, 51, 52] C [45, 57] E [3, 54]
CAS-PEAL [122]	2009 to 2018	6: B [12, 58, 61] C [19, 55] E [27]
FRGC [123]	2006 to 2018	6: B [10, 13, 34, 60] C [14, 55]
MS-Celeb-1M [124]	2019 to 2020	5: B [52] C [49, 50, 54] E [3]
CFP [125]	2019 to 2021	5: E [43, 44, 46, 49, 52]
IJB-C [126]	2019 to 2021	5: E [44, 45, 47, 49, 52]
YTF [127]	2014 to 2020	5: B [15] E [6, 11, 49, 52]
MS1MV2 [128]	2021	4: C [43–46]
IJB-A [129]	2017 to 2019	4: B [3] C [54] E [52, 57]
ChokePoint [64]	2011 to 2018	4: B [56, 59] E [55, 64]
SCface [130]	2011 to 2018	4: B [61] E [22, 55, 60]
Extended Yale [131]	2010 to 2018	4: B [23, 62, 65] C [55]
CPLFW [132]	2021	3: E [43, 44, 46]
IJB-B [133]	2021	3: E [43, 44, 46]
Adience [134]	2020 to 2021	3: E [43, 45, 50]
BioSecure [135]	2019 to 2021	3: E [40, 48, 53]
GBU [136]	2012 to 2014	3: B [12, 19] E [16]
AT&T [137]	2010 to 2016	3: B [58, 65] C [14]
CMU-PIE [138]	2009 to 2011	3: C [24] E [26, 64]
FRVT 2006 [139]	2008 to 2010	3: E [24, 25, 28]
Yale [140]	2007 to 2014	3: B [12, 19] E [32]
BANCA [141]	2006 to 2008	3: B [35, 36] E [29]
AgeDB [142]	2021	2: E [44, 46]
CALFW [143]	2021	2: E [44, 46]
MEDS-II [144]	2019 to 2020	2: B [2, 4]
MegaFace [145]	2019 to 2020	2: E [49, 52]
AR [146]	2014 to 2018	2: C [14, 55]
PaSC [147]	2013 to 2018	2: B [56] E [16]
MBGC [148]	2012 to 2014	2: E [12, 19]
Q-FIRE [149]	2012 to 2014	2: E [12, 18]

Dataset	Year	Used in
UTKFace [100]	2021	E [45]
CyberExtruder [101]	2020	E [48]
MEDS-I [102]	2020	B [1]
IJB-S [103]	2019	E [52]
ImageNet [104]	2019	C [54]
CMU-FIA [105]	2018	B [56]
NCKU face [106]	2018	C [55]
FIIQD [9]	2017	B [9]
Honda/UCSD [107]	2017	E [7]
FEI [108]	2016	B [58]
MIT [109]	2016	B [58]
AFLW [110]	2015	B [60]
BioLab-ICAO [17]	2012	B [17]
IIT-NRC [111]	2011	E [21]
Pointing'04 [112]	2011	C [21]
XM2VTS [113]	2010	B [23]
FRI-CVL [114]	2008	E [30]
HERMES project [115]	2008	E [30]
Cohn-Kanade [116]	2007	B [33]
WVU [117]	2007	B [33]

In-house datasets or datasets used only for other purposes (such as pure 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. The FIQA literature references in the rightmost columns are preceded by markers that denote the usage type: C - Dataset used only for FIQAA creation (model training or manual configuration); E - Only for FIQAA evaluation; B - Both creation & evaluation.

The surveyed FIQA works have been developed by a large variety of research groups. Independently of author relationships, various FIQA works are clearly based on prior work, which is noted both in the introductory literature text and the overview tables.

4.1 Factor-specific - Commonalities

The factor-specific approach commonalities can be described by the factor subcategories depicted in Figure 6:

- Size: Testing the inter-eye distance [1, 16, 17, 25, 28, 32, 34] or the image resolution [15, 21, 30, 31, 37] against some threshold is a comparatively simple approach to FIQA. It is present in various mostly older works alongside other FIQA methods. The referenced image resolution approaches mostly considered images cropped to the face and focused on

Table 3. Factor-specific FIQA Literature in Reverse Chronological Order

Reference	Aspects	Method(s)	Datasets
2020 [1]	DfrtFt	17 hand-crafted ISO/IEC TR 29794-5:2010 [66] related measures: Left-right symmetry ×7, capture-related ×10; 11 fused as random forests. *2 black-box COTS systems for FR.*	MEDS-I
2020 [2]	DlDuat	3 binary attributes (Eyes open, glasses, frontal); Non-DL: 23 models, i.a., SVMs; DL: Pretrained AlexNet [150], GoogLeNet [151]. *Best results via SVM+DL score-level fusion.*	In-house, MEDS-II
2019 [3]	DlDhgtFt	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
2019 [4]	DlDuat	Same as Reference [2], but, i.a., with smartphone images. *Continuation of Reference [2].*	In-house, MEDS-II
2019 [5]	Dhc	8 factors compared to mean scores from 26 humans. *Continuation of Reference [8].*	In-house (Smartphone)
2018 [6]	DlDfrtFt	CNN with MFM[152] & NIN[153] layers, trained using 15 synthetic degradation classes (5 types × 3 settings).	CASIA-WebFace, LFW, YTF
2017 [7]	DhgtFcFt	Subjective QS random forest, 7 hand-crafted features.	LFW, Honda/UCSD
2017 [8]	DfrtFt	9 factors plus random forest: Lighting symmetry, Pose symmetry, Brightness, Image contrast, Global Contrast Factor, Exposure, Blur, Sharpness, Vertical edge density.	In-house (Smartphone)
2017 [9]	DlDhgt	ResNet-50 trained on subjective illumination QSs. *Open source.*	FIIQD
2015 [10]	DfrtFt	AdaBoost on 3 "objective" measures [13] + optional training-set-"relative" measures. *Continuation of Reference [13].*	FRGC
2015 [11]	DuatV	Entropy, Viola-Jones [154] face detection confidence.	YTF
2014 [12]	DfrtFe	ANN on 5 factors/7 measures equivalent to Reference [19] vs. logistic regression, SVR, and 10 normalization/fusion combinations. *Continuation of Reference [19].*	CAS-PEAL, Yale, GBU, FERET, MBGC, Q-FIRE
2014 [13]	DuatFcV	Pose/Alignment (Reconstruction), Blur, Illumination.	FRGC
2014 [14]	DuatFcFtV	Two stages: 1. Pose (yaw/roll), 2. 12 GLCM features [155] fed into a GMM.	In-house (ABC), FRGC, AR, AT&T
2014 [15]	DfrtFtV	Facial symmetry, Illumination, Blur, Resolution.	YTF
2013 [16]	Dhc	9 FIQAA, i.a., Illumination (Direction), SEMC focus [24], Edge density [25], …, and SVM vs. GPO oracle. *Continuation of Reference [24].*	Unknown, GBU, PaSC
2012 [17]	Duat	30 factors, i.a., Hair Across Eyes, Looking Away, Varied Background.	BioLab-ICAO
2012 [18]	Dhc	Blur (MTF vs.: ED [156], LoG, SG, DCT).	Q-FIRE
2012 [19]	DuatFe	12 measures: Sharpness ×4, Contrast ×2, Illumination ×2, Focus ×2, Brightness ×2; Combined FIQAA with 7 factors.	CAS-PEAL, Yale, GBU, FERET, MBGC
2012 [20]	DhgtFc	5-class cascade SVM with Gabor magnitude features.	In-house
2011 [21]	DuatFeV	Pose (Linear Auto-associative Neural Networks), Illumination, Blur, Resolution. *QS relative to face image sequence. Continuation of Reference [30].*	In-house (100 videos), Pointing'04, IIT-NRC
2011 [22]	Duat	Landmark-based: Pose (Yaw/pitch/roll), Illumination (Histogram mass center variance), Symmetry (Lines).	FERET, LFW, SCface
2010 [23]	DhgtFeFt	Illumination of triangle mesh regions (Mean, ANN-weighted, Combined).	Extended Yale, XM2VTS
2010 [24]	Dhc	Illumination (Direction), SEMC focus, Edge density. *Continuation of Reference [25].*	FRVT 2006, CMU-PIE
2010 [25]	Dhc	Region density, Edge density, Eye distance. *Continuation of Reference [28].*	FRVT 2006
2009 [26]	Dhc	2 factors: Asymmetry (Imaginary Gabor filters), Sharpness ((I)DCT).	FERET, CMU-PIE
2009 [27]	Dhc	Symmetry (3 variations based on SIFT [157]).	CAS-PEAL
2008 [28]	Dhc	Edge density, Eye distance.	FRVT 2006
2008 [29]	DhcV	Blur (Sobel & Laplacian), Symmetry (Per-pixel).	BANCA
2008 [30]	DuatFeV	Pose (Center of mass distance), Illumination, Blur, Resolution.	FRI-CVL, HERMES project
2007 [31]	DuatFeV	Pose (Eye positions via gradient image), Illumination range & symmetry, Blur, Resolution, Skin content.	Unspecified (7 videos)
2007 [32]	Duat	6 factors: Lighting + Pose symmetry (LBP), Inter-eye distance, Illumination strength (Histogram), Contrast (Standard deviation), Blur (Gradient).	Yale
2007 [33]	DfrtFt	4 measures: Blur (Frequency kurtosis), Illumination (Weighted sum), Pose (Yaw), Expression (GMM).	FERET, WVU, Cohn-Kanade
2006 [34]	DhgtFt	27 factors listed, but few metric details; classification-error-based QS normalization; 3× QS fusion, i.a., ANN-based.	In-house (Passport database), FRGC
2006 [35]	Duat	Same as Reference [36], plus another score-level measure. *Continuation of Reference [36].*	BANCA
2006 [36]	Duat	Average face image correlation, Blur, Classification score sum of log-likelihoods.	BANCA
2005 [37]	DhcFc	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)
2004 [38]	DuatV	Pose (Haar features learned via SquareLev.R).	Unspecified (300 faces)
2004 [39]	DhgtFt	RBF-ANN on: Brightness, Spectrum ×7, Noise ×2.	Unspecified (850 images)

video-frame selection. Besides the face detection step, using the image resolution is trivial, while inter-eye distance requires eye landmark detection.

- Illumination: Many of the surveyed works included mostly simple illumination measures, comprising the brightness moments (mean, variance, skewness, or kurtosis) [1, 5, 7, 12, 16, 19, 21, 30, 33, 39], contrast measures [1, 5, 12, 16, 19, 32], dynamic range measures [7, 31, 34], entropy measures [10, 11, 13], or uniformity measures [9, 22–24, 34]. Note that "illumination" is of course also directly or indirectly measured by FIQAAs categorized under other parts of the taxonomy, which are consequently not listed here to avoid many duplicate listings of the same FIQAAs.

- Blur: Blur measures, or conversely sharpness measures, are also known as (de)focus measures. The measures can be subdivided into edge analysis approaches [1, 5, 8, 12, 15–19, 24, 25, 28, 29, 32, 34–36], frequency analysis approaches [10, 13, 18, 26, 31, 33], and low-pass filter approaches [18, 21, 26, 30]. Edge analysis involves image gradient computation, frequency analysis inspects the image transformed into the frequency domain, and low-pass filter approaches compare an artificially blurred version of the image with the original. Besides these more common subcategories, there were some comparatively opaque (i.e., less easily explainable) deep learning approaches among the FIQA literature that measured blur [3, 6].

- Symmetry: Holistic symmetry measures compare the entire left and right half of the face. The halves are defined either as fixed left/right splits of the whole input image [5, 8, 16, 26, 29], or are fitted to the face within the image [1, 5, 8, 31, 32]. Fixed left/right halves assume that the face is frontal without rotation, while fitted halves can account for some degree of head rotation. Besides these holistic methods there are localized symmetry measures that compare a number of paired local features within the left/right face halves, either based on general key points, e.g., via **SIFT (Scale Invariant Feature Transform)** [27] or based on facial landmarks [15, 22]. Of the surveyed methods, only the localized landmark-based measures inherently avoided the inclusion of image background information, although any of the methods could be extended to exclusively consider the facial area.

- Pose: Most pose FIQAAs were based on facial landmarks [7, 14, 17, 22, 31, 33, 37]. Others operated in a holistic manner by using appearance templates [158] to estimate pose angle ranges [21, 38], by assessing the frontal face reconstruction error [10, 13], by assessing the pose without angles in terms of the pixels' center of mass deviation [30], or via comparatively opaque machine learning approaches that assessed whether the pose is frontal or not, either with scalar [3] or binary [2, 4] output. Among the methods that did correspond to specific pitch/yaw/roll angles (see Figure 3) most did consider the yaw angle in addition to either the roll or pitch angle, while the rest considered either only the yaw angle or all three angles [22, 33, 37]. The one landmark-based method that did not correspond to any specific angle [31] computed the deviation of a landmark-derived point from the horizontal face center, which is closer to the holistic center of mass deviation approach of Reference [30].

- Other: There are some comparatively rare factor-specific FIQA approaches in the literature that were collected under the "Other" taxonomy category, namely, binary attributes such as with/without glasses in References [2, 4], noise measures in References [6, 39], skin tone measures in [31, 34, 37], deep learning "alignment" and "occlusion" measures based on human ground truth in Reference [3], and miscellaneous standard requirement check methods such as ink mark & crease detection in [17, 34, 37].

4.2 Factor-specific - Literature introductions

Luo [39] considered general IQA related to brightness, blur, and noise in the context of face images. Ten 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 resulted in worse performance. The IQAA was trained with and compared against the quality estimates of a single human on an unspecified dataset. The 10 features consisted 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).

The approach of Yang et al. [38] 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 be used in isolation for FIQA.

Subasic et al. [37] used 17 FIQAAs based on ICAO Doc 9303 [159] requirements. This includes measures that are less common among the literature, such as background uniformity and color balance. The 17 FIQAAs were integrated as part of a combined FIQA system, reusing background/skin/eye-segmentation for multiple measures, and hierarchically executed, i.e., resolution and sharpness are examined first. The combined FIQAA was used to determine whether an input image is ICAO-compliant or not, and the evaluation tested this binary prediction against 189 correspondingly labelled images of an unnamed database, correctly classifying 88%. Tolerance ranges were established based on a small subset of images where no quantitative ICAO requirements were available, and some existing ICAO tolerance ranges were relaxed.

In the approach of Kryszczuk and Drygajlo [36], two image-based ("signal-level") and one classification-score-based ("score-level") measure were used, and all three were combined by means of two GMMs with 12 Gaussian components each for binary assessments regarding "correct" and "erroneous" FR classifier decisions. The authors also added another score-level measure to the approach in Reference [35]. But the inclusion of measures based on FR classification scores means that the combined method can only be used after an FR comparison has taken place, so this component would have to be excluded to allow isolated single-image FIQA using the remaining two image-based methods. Of these, one measured sharpness as the mean of horizontal/vertical pixel intensity differences (corresponding to high-frequency features), and the other computed Pearson's cross-correlation coefficient between the face image and an average face image (corresponding to low-frequency features). The average face image was formed from the average of the first eight **PCA (Principal Component Analysis)** eigenfaces for a given training image set.

Hsu et al. [34] used 27 FIQA factors, which mostly relate to ISO/IEC 19794-5:2005 [160] requirements. While only very brief descriptions of the underlying FIQA approaches were provided, the work proposed quality score normalization and fusion with more details. The normalization per metric was based on the classification error against binary human quality labels ("good"/"poor"). Raw quality metric values were mapped to [0, 1] via five raw value thresholds, interpolated via sigmoid functions. The five raw threshold values were taken from five specific points of the false-accept/reject classification error curves, and corresponded to the quality scores 0, 0.4, 0.5, 0.7, 1.0. For FIQAA fusion, three models were trained, and the evaluation showed that a non-linear neural network obtained the best results in terms of correlation with FR performance.

Abdel-Mottaleb and Mahoor [33] proposed FIQAAs to assess blur, lighting, pose, and facial expression. Blur was measured as the kurtosis in the frequency domain. The lighting QS was formed by a weighted sum of the mean intensity values for 16 weight-defined regions to focus more on the center of the image. Pose was estimated as the yaw angle (see Figure 3), derived by comparing the amount of skin tone pixels between the left/right-side triangle, which in turn were defined by

the three center points of the eyes and the mouth. **Fisher Discriminant Analysis (FDA)** was employed to differentiate skin pixels from other regions. To assess whether the expression is good or bad in terms of quality, a GMM was trained based on the correct/incorrect decisions of an FR algorithm for a labeled facial expression dataset.

Gao et al. [32] proposed FIQAAs for asymmetry, inter-eye distance, illumination strength, contrast, and blur. Lighting/pose asymmetry was computed as the sum of the rectilinear distances between the histogram pairs for multiple **LBP (Local Binary Pattern)** features at designated locations in the face image halves. Illumination strength was proposed to be computed as the difference between a histogram for the input image and a fixed standard illumination histogram, contrast as the pixel value standard deviation, and sharpness as the gradient magnitude sum. The asymmetry metric was tested in terms of classification accuracy with a small labeled dataset. The methods have been incorporated into ISO/IEC TR 29794-5:2010 [66] (but the work is only cited directly for the lighting/pose symmetry part).

Fourney and Laganiere [31] 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 Reference [38]. The pose estimation in Reference [31] also worked 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 FIQA 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 Reference [31] were unrelated to the pose estimation: A normalized blur/sharpness QS was derived from the frequency domain; the face image resolution/pixel count was 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 detected whether human skin appears to be present in the image, which was 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 the number of satisfied per-factor thresholds, plus a weighted sum of the factor scores to break ties between video frames.

For References [30] and [21] 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 are not directly usable for single-image FIQA. We can remedy this obstacle using simple tricks, for example by choosing constant minima/maxima, hence why these works are still included here. The first of the two papers, [30], i.a., cited Reference [31] and directly adapted the face image resolution factor, but presented different approaches to measure the other shared factors: The FIQAA started with information gathered as part of the face detection stage, which determined 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 References [31] and [17]. Instead, i.a., the facial center of mass was 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 Reference [31]), 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 previously mentioned approaches that estimated 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 FIQA). Sharpness/blur was assessed using the approach presented in Reference [156], 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 FIQAA in Reference [21] can be seen as a continuation of Reference [30], with the sharpness, brightness, and resolution measures being almost identical. Brightness had now been more clearly defined as the Y component of the YCbCr color space and the resolution QS bound was removed (i.e., it became completely relative to an image sequence). The pose estimation was changed, stating that the prior center of mass approach in Reference [30] tended to be sensitive to environmental conditions. The new approach estimated actual angles and is adapted from Reference [161], using one auto-associative memory (an ANN without hidden layers) per detectable pose.

Rúa et al. [29] proposed three FIQA methods in the context of face video frame selection. One method measured symmetry by comparing the image against a horizontally flipped version of itself, calculating the per-pixel difference, meaning that this measure assumed a centered frontal pose. The other two FIQA methods assessed blur by computing the average value for either the Sobel or the Laplace operator over the entire input image.

Beveridge et al. examined the impact of a number of factors on FR verification performance in References [28] and [25] using **GLMMs (Generalized Linear Mixed Models)**. Taking the examined preexisting labels such as age or gender out of consideration, three described measurements were considered for automatic image-only quality assessment, one of which is the image resolution/eye distance. Two more complex measurements remain, with Reference [28] introducing an edge density metric consisting of the averaged Sobel filter pixel magnitude and Reference [25] 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 Reference [24] by comparing their edge density metric to two newly introduced FIQAAs. One was the **Strong Edge and Motion Compensated focus measure (SEMC focus)**, a successor to the edge density metric that was 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 FIQAA estimated to which degree a face is lit from the front (positive number output) or the side (negative number). Experiments in Reference [24] used GLMMs and FRVT 2006 test data/FR algorithms similar to References [28] and [25] and found that the illumination measure subsumed both the edge density and the SEMC focus measure regarding FR performance prediction. These measures were studied further in Reference [16], as described below.

Zhang and Wang [27] proposed three symmetry measure variations based on **SIFT (Scale Invariant Feature Transform)** [157]. The first variation counted the number of SIFT points in the left and right half of the image and divided 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 was 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 added 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.

Sang et al. [26] proposed a Gabor-filter-based asymmetry FIQAA to assess the illumination/pose and a sharpness FIQAA using **DCT+IDCT (Discrete Cosine Transform + Inverse DCT)**. The asymmetry FIQAA used the left/right halves of the input image, expecting an aligned frontal image. It was computed as the sum of the absolute difference between the left/right pixels for multiple filtered versions of the image halves, mirroring the right half for the comparison. The imaginary parts of Gabor filters were used with five orientations, also mirrored on the right half. To assess

sharpness, DCT followed by IDCT was applied to the input image to obtain a reconstructed version without high-frequency information, and the difference between both image variants was used to establish the sharpness value. The asymmetry FIQAA was examined via score plots for images with different lighting and pose conditions (of the same subjects) and the sharpness FIQAA similarly for either unmodified or synthetically blurred images, demonstrating classification potential for both. The asymmetry FIQAA approach from Reference [32] was included in the tests and produced similar output compared to the proposed FIQAA.

Rizo-Rodriguez et al. [23] presented a frontal illumination assessment method. First, a triangular mesh was fitted to the face in the input image. Then the mean luminance was 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. [22] proposed landmark-based measures for pose, illumination, and symmetry. For pose, the yaw/pitch/roll angles were 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 formed the pose QS, whereby the weights for yaw (0.6), pitch (0.3), and roll (0.1) were derived experimentally. Illumination was measured by applying a sigmoid function to the variance of the mass centers for eight gray-level histograms, which were computed for areas around eight landmarks (three on the nasal ridge, two on each cheek, one on the chin). Symmetry was measured by comparing the grayscale values of point pairs sampled along eight lines defined by landmark pairs on each side of the face. All three measures provided $[0, 1]$ scalar QS results. They were 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 differed depending on the used FR algorithm.

Liao et al. [20] trained an **SVM (Support Vector Machine)** cascade to predict subjective QS labels using Gabor filter magnitude values as features. The SVM cascade had four stages, each being a binary classifier, so the approach predicted integer QS levels from 1 to 5 (e.g., the first SVM decides whether the QS is 1 or whether it might be higher). Two of these 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 FIQA by Abaza et al. in Reference [19], and later Reference [12], i.a., incorporating synthetic image degradations regarding contrast, brightness, and blurriness for the evaluations. Of the 12 tested individual measures in Reference [19], 7 were retained to represent 5 input factors for a combined single-image FIQAA, 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 [12]), focus as the mean of the image gradient's L_1-norm and the Laplacian energy [162], sharpness as the mean of the two average gradient measures [32, 35], and illumination using the weighted sum technique proposed by Reference [33]. The five measures that were not used for the combined FIQAA comprise the Michelson contrast measure [163], the brightness measure from Reference [164], the Tenengrad sharpness measure plus an adaptive variant from Reference [165], and the luminance distortion [166] measure previously seen in

Reference [65] (but without the face average as the "reference"). Note that according to both References [19] and [12] the selected brightness measurement was chosen due to its reduced computational workload in comparison to the other tested method (which achieved better predictive performance). Continuing with Reference [12], the same 5 factors based on the chosen 7 (of 12) measures were presented as in Reference [19], 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 (×2, linear or Gaussian model) and a fusion (×5) part, including the previous method from Reference [19]. 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 was not used in isolation.

To measure blur in face images, Hua et al. [18] 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 3)** metric that consisted 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 first subtracting the 3×3 mean filtered image from the original, then taking the average of the result's absolute pixel values [156]. This measure also occurred in Reference [30], but is not to be confused with the previously mentioned Sobel filter edge density from References [28] and [25]. The correlation of these measures (applied to a face image) to a ground truth MTF applied to an optical chart was 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.

Ferrara et al. [17] introduced the "BioLab-ICAO" framework for ISO/ICAO-compliance assessment, comprising a database, a testing protocol, and a set of 30 FIQAAs for requirements derived from ISO/IEC 19794-5:2005 [160] (plus corrigenda/amendments). The 30 FIQA measures included factors that were less common in the surveyed literature, such as the detection of ink marks or creases. The evaluation individually tested 23 of the 30 measures, together with 2 unnamed **COTS (Commercial Off-The-Shelf)** FIQAAs, mostly in terms of compliance prediction accuracy (range [0, 100]) against ground truth labels. Most of the proposed BioLab-ICAO methods either outperformed both COTS systems or lacked a testable COTS counterpart, although the assessment of various requirements was still deemed to be difficult. BioLab-ICAO methods were later used in the training data preparation for FaceQnet v0 [53] and v1 [48].

Phillips et al. [16] examined 13 quality measures, including the edge density metric from References [28] and [25], plus the SEMC focus measure from Reference [24] (all four of these papers share authors). There also was an "illumination direction" measure that might correspond to Reference [24] as well, but this was not clarified. Similar to the two prior papers [28] and [25], the 13 quality measures in Reference [16] contained preexisting labels from **EXIF (Exchangeable Image File)** metadata, e.g., exposure time, leaving 9 measures that can clearly consist of FIQA approaches that use the actual image (pixel) data: Edge density [28], SEMC focus [24], illumination direction (possibly [24]), 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 was an SVM that summarized the other 12 measures. Pruning based on the 13 measures was compared against a **Greedy Pruned Order (GPO)** oracle that discarded images in an approximately optimal fashion to improve FR performance, thus representing an upper bound for FR performance improvements enabled by some FIQAA. 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 **FMR (False Match Rate)** results. Another FIQAA using PCA followed by **LDA (Linear Discriminant Analysis)** was trained, but it was observed to generalize poorly to the test set.

In the single-image FIQAA of Nikitin et al. [15] the resolution and illumination measurement, as well as the fusion to combine the factor-QSs, did not differ much from what has been mentioned previously (resolution QS relative to constants, illumination dynamic range usage QS, fusion via weighted sum). However, 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. [14], 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 was not detected as frontal, the overall FIQAA stopped, assigning the image to the poor QS bin. If the pose was detected as frontal, then the second stage decided between the fair/good QS bin assignment. It consisted of 12 **GLCM (Gray Level Co-occurrence Matrix)** features [155], which were 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.

The approach of Kim et al. [13] began 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 was measured as the kurtosis of the **CDF (Cumulative Distribution Function)** of the **DFT (Discrete Fourier Transform)** magnitude; and the last stage assessed brightness by comparing the histogram for the face image against a given reference histogram, whereby the latter simply was chosen to be the uniform histogram. Each of these three stages ended by comparing the error value result against a predefined threshold, aborting the overall FIQAA if the threshold was exceeded. This cascaded approach in Reference [13] was primarily meant to reduce the computational complexity for video processing. In the follow-up paper [10] the same three measures were utilized, but without the cascaded approach. Instead, the output of the three so called "objective" measures formed a QS vector. An additional "relative" quality measurement was conducted to assess the dissimilarity of the input image (e.g., from the test dataset) to the training dataset images. This was done via a multivariate Gaussian distribution for a six-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 FIQAA, the input thereof being a nine-dimensional vector made up of the three-dimensional "objective" and six-dimensional "relative" measure output. Note that the "relative" measure is entirely optional, but it did improve the quality assessment according to the evaluation in Reference [10]. In these evaluations both variants of the proposed FIQAA appeared superior to the also tested RQS [60], which seemed to actually degrade FR performance.

The work [11] by Damer et al. included three face frame selection methods, of which two could be considered for single-image FIQA. One method measured the entropy of the color channels, higher entropy being preferred. The other method calculated the confidence for a Viola-Jones [154] face detector as the sub-image classifier detection count, which can correspond, i.a., to pose and illumination.

Zhang et al. [9] 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 ResNet50 [167] 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.

Wasnik et al. [8] examined FIQA in the context of smartphone-based FR, evaluating eight FIQAAs based on ISO/IEC TR 29794-5:2010 [66] specifications and proposing a vertical edge density FIQAA for pose/lighting symmetry, plus a combined random forest FIQAA. The vertical edge density FIQAA computed the input image gradient, only keeping the magnitude for (vertical edge) pixels in a certain gradient phase range and used the mean of all magnitude pixel values to form a scalar result. The random forest FIQAA combined the eight ISO metrics, and a second variant replaced the ISO symmetry assessment part with the proposed vertical edge density FIQAA. To train the random forest algorithm, a database was first separated into good- and bad-quality images using a COTS system (VeriLook 5.4 [168]) plus subsequent manual checks by three trained experts. All nine individual FIQAAs, the two random forest FIQAAs, and the COTS FIQAA were evaluated by computing ERCs using an FR implementation from the same COTS suite [168]. The COTS FIQAA and the random forest algorithm incorporating the vertical edge metric provided the best results in terms of partial (20%) ERC AUC. The work by Khodabakhsh et al. [5] can be considered as a continuation of Reference [8], which examined the eight ISO FIQAAs in comparison to subjective quality assessments made by 26 human participants for smartphone images. It concluded, i.a., that the human FIQA highly correlated with FR performance, but not with the tested FIQAAs, indicating that the tested FIQAAs show limitations. Correlation between the metrics were also shown.

Wang [7] presented a hybrid approach to estimate subjective QSs using features consisting of seven factor-specific scores. The factors comprised 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 compared the predictive performance of this approach against the cascaded SVM method of Reference [20], with the results favoring the proposed approach for QSs 2 to 3.

Yu et al. [6] proposed using a CNN architecture with **MFM (Max-Feature-Map)** [152] and **NIN (Network In Network)** [153] layers for FIQA. Training used 16 classes: One represented the original unmodified training images, while the other 15 represented 5 types of synthetic degradation thereof, with 3 configurations of increasing severity each. These 5 degradation types comprised nearest-neighbor downscaling, Gaussian blur, **AWGN (Additive White Gaussian Noise)**, salt-and-pepper noise, and Poisson noise. This was sufficient to train a network to classify these degradations. To also estimate a scalar QS, an FR accuracy score was precomputed for each of the 16 classes, and the sum of the multiplication of those scores with the 16 classification probabilities formed the combined QS. The proposed CNN architecture was also used for the FR part (as a separately trained model), using the cosine distance as the similarity measure. Three variants of the network were evaluated for FIQA: One trained from scratch for FIQA, one first trained for FR before training for FIQA, and one that used 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 five 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.

Rose and Bourlai evaluated DL and non-DL methods to determine three binary facial attributes in References [2] and [4] (which was a continuation of Reference [2] 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 [150] and GoogLeNet [151] (an incarnation of the Inception architecture), pretrained on ImageNet [104] data. Their architectures were modified to classify two labels per attribute (i.e., six classes). And there were 23 non-DL models tested in Reference [2], 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 an SVM and either AlexNet or GoogLeNet led to the best results in Reference [2]. The evaluations in Reference [4] employed a smartphone (iPhone 5S) dataset in addition to the non-smartphone data used in Reference [2], the latter of which was only used for training. Of the non-DL methods, result values in Reference [4] were 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 was better varied between the experiments of Reference [4], which proposed to use the SVM trained on a combination of all used datasets (score-level fusion of the SVM and one of the DL networks was not tested).

Lijun et al. [3] proposed a multi-branch FIQA network, called MFQA, consisting of a feature extraction and a quality score part. The former was a CNN to derive image features. The latter fed these features into four fully connected branches for different quality properties and fused the output thereof into a final QS via another fully connected layer. These four branches corresponded to scores for alignment, visibility (i.e., occlusion), frontal pose, and clarity (i.e., blur). For training, 3,000 images were manually annotated with ground truth labels for the four factor scores and the overall QS.

Henniger et al. [1] examined 17 hand-crafted FIQA measures drawn from ISO/IEC TR 29794-5:2010 [66]. Of these, 7 measured symmetry of the left-right face image halves, whereof 1 measure summed the normalized pixel luminance differences, and the other 6 calculated the cross-entropy, Kullback-Leibler divergence, or histogram intersection for either the normalized or LBP-filtered pixel luminance values. The remaining 10 measures were capture-related methods found in ISO/IEC TR 29794-5:2010 [66], namely, general image contrast, global contrast factor, mean/variance/skewness/kurtosis of pixel luminance, exposure, sharpness, inter-eye distance, and blur. In the evaluation, the face image utility was first derived via FR comparisons using 2 unnamed black-box COTS FR systems, labeling images per system either as high-quality if their minimum mated comparison score was greater than a threshold for 60% FNMR [70], as low-quality if their maximum mated score was below a threshold for 30% FNMR, or leaving them unlabelled otherwise. The 17 measures were then examined in terms of their correlation with the FR-system-derived utility and by means of FNMR ERC plots. Based thereon, 11 measures were selected to create random forest models, namely, the 3 histogram symmetry measures and all capture-related measures except variance and skewness. Random forest training used the utility labels for the 2 individual systems, in addition to the union and intersection thereof.

4.3 Monolithic - Commonalities

The monolithic approaches do not have factor-specific subcategories by definition, but the dominant commonalities and differences can be highlighted via the data aspect instead:

- Utility-agnostic training (**Duat**): The most recent approach [42] evaluated a general IQA CNN from Reference [170] for the purposes of FIQA primarily on different facial areas. Besides Reference [42], all of the works that exclusively proposed monolithic **Duat** approaches [62, 64, 65] happened to rely on model data derived from a fixed set of training images and were non-DL. Both References [62] and [65] directly compared the input against an

Table 4. Monolithic FIQA Literature in Reverse Chronological Order

Reference	Aspects	Method(s)	Datasets
2021 [40]	DlDint	Evaluation of 6 monolithic FIQAAs, 10 IQAAs, and 9 factor-specific hand-crafted methods.	LFW, VGGFace2, BioSecure
2021 [41]	DlDint	Evaluation of monolithic FIQAAs [44, 48, 50, 60] on face images without masks, with real face masks, and images with synthesized masks.	In-house [169]
2021 [42]	DlDuat Fe	CNN IQA [170] on various facial areas (eyes, nose, mouth, averaged fusion) and on cropped or aligned images. Compared against monolithic FIQAAs [44, 50, 53, 60].	LFW, VGGFace2
2021 [43]	DlDfrt	Identification quality (IDQ) loss focused on the FR comparison threshold, used to train a FIQA branch in a frozen FR network, in turn used for knowledge distillation to train a separate lightweight FIQA network. *Open source.*	MS1MV2, LFW, CFP, CPLFW, IJB-B, Adience
2021 [44]	DlDint	Extends ArcFace [128] training loss with FR feature embedding magnitude-aware angular margin and regularization, so magnitude corresponds to quality. *Open source.*	MS1MV2, LFW, CFP, AgeDB, CALFW, CPLFW, IJB-B, IJB-C
2021 [45]	DlDfrt	The Wasserstein distance between FR comparison score sets for randomly selected mated and non-mated pairs is used to form ground truth QSs for FIQA network training with Huber loss. *Open source.*	MS1MV2, CASIA-WebFace, LFW, Adience, UTKFace, IJB-C
2021 [46]	DlDfri	Reduces the Reference [52] concept to a single uncertainty scalar, adding regularization relative to mini-batch uncertainty average, plus two uncertainty-aware identification loss variants, and the network uses multi-layer fusion. *Open source.*	MS1MV2, LFW, CFP, CALFW, CPLFW, AgeDB, IJB-B, VGGFace2
2020 [47]	DlDfrt	ResNet18 [167] FIQA model trained on ResNet34 [167] FR model ground truth scores, computing loss for the predicted QS minimum of each image pair.	VGGFace2, IJB-C
2020 [48]	DlDfrt	Same as Reference [53], but with dropout before the first fully connected layer, and multiple FR feature extractors to obtain the ground truth QSs. *Continuation of Reference [53]. Open source. Benchmarked in NIST FRVT QA [67].*	VGGFace2, LFW, CyberExtruder, BioSecure
2020 [49]	DlDint	Learns both uncertainty and FR features: Either 1. KL divergence loss to train an entire network, or 2. fixed FR network extension with loss relative to subject feature centers. Builds upon the uncertainty vector concept of Reference [52].	MS-Celeb-1M, LFW, MegaFace, CFP, YTF, IJB-C
2020 [50]	DlDfri	QS based on comparing embeddings from 100 random subnetworks; Works on FR networks trained with dropout or by adding a network on top. *Open source.*	MS-Celeb-1M, FERET, Adience, LFW
2019 [51]	DlDhgt	CNN trained on binary labels derived automatically based on fewer manual labels and non-DL methods. *Predicts scalar QSs after binary training.*	In-house, CASIA-WebFace
2019 [52]	DlDfri	Based on a pretrained FR network, trains separate two-layer perceptron network to measure per-feature-dimension uncertainty, compares via MLS (Mutual Likelihood Score). *Open source.*	CASIA-WebFace, MS-Celeb-1M, LFW, YTF, MegaFace, CFP, IJB-A, IJB-C, IJB-S
2019 [53]	DlDfrt	Frozen FR-pretrained ResNet-50 [167], training two new final layer replacements on QSs derived from FR features vs. BioLab-ICAO[17]-selected references. *Open source. Benchmarked in NIST FRVT QA [67].*	VGGFace2, BioSecure
2019 [54]	DlDhgt	"DFQA," a SqueezeNet[171]-based two-branch CNN; training with SVR loss; ground truth QSs generated by another CNN, in turn trained using 3,000 rule-guided human QS labels. *Direct SqueezeNet successor: SqueezeNext [172].*	ImageNet, IJB-A, MS-Celeb-1M, CASIA-WebFace, VGGFace2, LFW
2018 [55]	DlDhgt	14 methods: 2 FIQA CNNs, 5 non-FIQA CNNs, 3 non-FIQA mobile CNNs, 3 Hand-crafted, 1 COTS; Binary training labels (good/bad). *Considers FIQA in the context of smartphone FR in addition to non-smartphone data.*	In-house, CAS-PEAL, Extended Yale, AR, FRGC, NCKU face, ChokePoint, SCface
2018 [56]	DlDfrtV	CNN with inception module, trained using gallery FR comparison score minima for detected faces.	In-house, PaSC, ChokePoint, CMU-FIA
2017 [57]	DlDfrt Ft	5 methods, using either human or FR-based labels, DL or L2R [60] features, and SVR or L2R models. *Another paper version is Reference [173].*	LFW, IJB-A, CASIA-WebFace
2016 [58]	Dfrt	Kernel Partial Least Squares Regression using mean luminance and Laplacian of 10×10 image sub-blocks.	CAS-PEAL, FERET, MIT, FEI, AT&T
2015 [59]	DlDfrtV	CNN, PCA whitened input, QS labels via MSM.	ChokePoint
2015 [60]	DlDhgt Ft	2-stage learning to rank for five feature extractors: CNN (Landmarks), HOG, Gist [174], Gabor, LBP (per-feature-vector QS formed by weighted sum). *Can be considered non-DL by removing the CNN extractor. Open source.*	In-house/Unknown, FERET, FRGC, LFW, AFLW, SCface
2013 [61]	Dfrt	4-class SVM on Gist[174] or HOG.	SCface, CAS-PEAL
2012 [62]	Duat	Gaussian low-pass filter vs. fixed 38-image-average reference.	Extended Yale
2012 [63]	Dfri	"Impostor-based Uniqueness Measure," i.e., FR vs. fixed image set as FIQA.	In-house (Police)
2011 [64]	Duat	Per block low-frequency 2D DCT components compared to "ideal" frontal face.	FERET, CMU-PIE, ChokePoint
2010 [65]	Duat	Luminance distortion from Reference [166] against an averaged frontal face image.	Extended Yale, AT&T

averaged image, while Reference [64] compared against Gaussian distributions derived from the training images.

A few other works proposed both factor-specific and monolithic (Duat) FIQAAs, namely, References [35, 36], which contained a monolithic average face image correlation measure,

and Reference [11], which proposed to use the Viola-Jones face detection confidence as a monolithic FIQAA. To avoid duplicates, this literature is only listed in the factor-specific Table 3 and introduced in Section 4.2.

- Human ground truth training (**Dhgt**): These FIQA approaches were trained to estimate ground truth QS labels that stemmed from human assessments [51, 54, 55, 60]. Some of these works automatically transferred the human QSs to additional unlabeled images to extend the available training data [51, 54].
- FR-based ground truth training (**Dfrt**): These approaches obtained training data from FR models [43, 45, 47, 48, 53, 56–59, 61]. The majority of the monolithic approaches belong to this category.
- FR-based inference (**Dfri**): FIQAAs with FR-based inference utilize FR models as part of the quality assessment process even outside the training stage, but do not alter the FR model training [46, 50, 52, 63]. A notably early non-DL variant in this category is Reference [63], which directly compared the input image against a comparatively large fixed set of 1,000 images from different subjects with an FR system to assess the quality. The later approaches are DL-centric and estimate uncertainty for an FR model.
- FR-integration (**Dint**): FR-integrated FIQA methods not only use the FR model during inference, but are fully integrated into the FR model, meaning that FR and FIQA training are intertwined [44, 49]. This concept has emerged more recently than the others.

4.4 Monolithic - Literature introductions

Sellahewa and Jassim [65] used the luminance distortion component from the "universal image quality index" [166] to compare a face input image against a fixed average reference image generated from a training set (not to be confused with full-reference IQA, where a high-quality variant of the input image itself is known). This method worked by sliding a 8×8 window simultaneously over the input and reference image, computing $2L_{\text{input}}L_{\text{reference}}/(L_{\text{input}}^2 + L_{\text{reference}}^2)$ therein with L being the mean luminance, and using the mean of all window results as the final $[0, 1]$ QS.

Wong et al. [64] presented a FIQAA for frontal face images. Low-frequency 2D **DCT (Discrete Cosine Transform)** components were extracted for overlapping blocks of a normalized grayscale face image. Per block, these were compared against Gaussian distributions derived from a set of training images with frontal illumination, and a final QS was formed by fusing the resulting probabilities.

Klare and Jain [63] presented the **impostor-based uniqueness measure (IUM)**, an approach inherently adaptive to any used FR system. It was computed 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, Reference [63] proposed to use 1,000 feature vectors from different subjects to form this set. Note that the paper appeared to only utilize frontal face images (from an operational police dataset).

Qu et al. [62] proposed a FIQAA based on Gaussian blur face model similarity. The Gaussian blur was applied to the input image, which was then compared, in terms of the normalized correlation, against a fixed reference image formed by the average of 38 training images. The paper evaluated a range of sizes for the Gaussian blur. FR performance was not evaluated, but an evaluation can be found as part of the illumination methods considered in Reference [58].

Bharadwaj et al. [61] trained a one-vs-all SVM for four quality bins using either sparsely pooled **Histogram of Oriented Gradient (HOG)** or Gist [174] input features. The quality bin training labels were obtained using two COTS FR systems on training images that had a single designated good/studio quality image in addition to several probe images per subject.

Chen et al. [60] proposed the learning to rank approach with two stages. In stage one a number of preexisting feature extractors were used on the input image, and for each feature output vector thereof a **RQS (Rank-based Quality Score)** was derived as the features' weighted sum. Stage two applied a polynomial kernel to the RQS output vector of stage one and again used 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 the aforementioned weighted sums so each RQS differentiates between images from a number 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 does not have to use any deep learning, but the evaluated FIQAA implementation incorporated a CNN for facial landmark detection as one of five feature extractors. The other four (non-DL) feature extractors comprised Gist [174], HOG, Gabor, and LBP.

In Reference [59] by Vignesh et al., a CNN was utilized to directly output a final FR-performance-focused QS for a 64×64 face image input. The network had four convolutional layers and the face image input was preprocessed using PCA whitening. Training this approach required 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. [58] proposed to train a **KPLSR (Kernel Partial Least Squares Regression)** model for FIQA. Two features were derived for 10×10 sub-blocks of an image, forming a 200-dimensional feature vector as input for the KPLSR model. These features were the mean luminance and Laplacian gradient per sub-block. The training ground truth QSs were LBP-based FR comparison scores, whereby each image pair consisted 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 FIQAA and the FR performance was demonstrated in the evaluation.

In References [57] and [173], Best-Rowden and Jain presented multiple FIQAA variants partially based on DL. Five FIQAAs were evaluated, including the RQS approach of Reference [60]. Of the four newly proposed FIQAAs, three used training ground truth QSs derived from pairwise relative human assessments, and one derived the ground truth QSs from FR-method-dependent comparison scores with manually selected gallery images. Two of the methods used the 320-dimensional feature vector of an FR CNN [175] to train an 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 were also used in another variant of the human ground truth methods, which replaced the SVR with the **L2R (learning to rank)** approach of Reference [60] ("HQV-1"). The fourth method trained the L2R approach of Reference [60] with the features described therein, but for the human ground truth instead of the RQS dataset constraints [60] ("HQV-2"). In the evaluation, the CNN of Reference [175] was also used as one of the FR algorithms, in addition to two unnamed COTS systems. 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 [175], and RQS [60] for one of the COTS systems.

Qi et al. [56] used a CNN architecture with an inception module for FIQA. 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 FIQA network. A pretrained VGG-16 [176] and Inception-v3 [177] network was used for the FR part. The video frame FR performance improvement evaluation, i.a., compared against the

CNN approach of Vignesh et al. [59] and the learning to rank approach of Chen et al. [60], with the proposed CNN showing the best results.

Wasnik et al. [55] compared 14 methods for FIQA using seven publicly available datasets (plus in-house datasets) in the context of smartphone FR. Of the 14 methods, 10 were CNNs, 3 were hand-crafted, and 1 was a COTS system (VeriLook 5.4 [168]). Among the 3 hand-crafted methods, 2 were general IQAAs (BLIINDS-II [178], BRISQUE [79]), and 1 was Wasnik et al. [8]. Among the 10 pretrained CNNs, 2 were meant specifically for FIQA (the illumination-focused FIQA [9], and the general FIQA [56]), 3 were mobile networks (MobileNetV2 [179], DenseNet-169 [180], NASNet [181]), and the other 5 were AlexNet [150], VGG-16/VGG-19 [176], Inception [151], and Xception [182]. Of the two FIQA-specific CNNs, for Reference [9] a pretrained network provided by the authors was used, and for Reference [56] the network described therein was recreated while using the training dataset of Reference [55]. To adapt the non-FIQA CNNs for the FIQA task, the last three layers were replaced by fully connected layers of sizes 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 did not 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 FIQA-adapted CNNs AlexNet and Inception. This evaluation used five separate datasets, and the VeriLook SDK 5.4 [168] for FR comparisons.

Yang et al. [54] presented "DFQA," a FIQA CNN based on SqueezeNet [171], which itself was 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 [172]). However, it was not proven whether this performance equivalence is true for the biometric FIQA task here, since Reference [54] did not compare against any AlexNet-based FIQA variant, e.g., one analogous to their SqueezeNet-based approach, or the one used in Reference [55]. Most of the SqueezeNet architecture parts in the DFQA [54] network were represented in two functionally identical weight-sharing branches (also called "streams" in Reference [54]), each of which was followed by a (no longer weight-sharing) 1×1 kernel convolutional layer with 9×9 output. Then the mean of the two outputs was fed to an average pooling layer, resulting in the output feature vector. The paper compared 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 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 [124] dataset, which were then used to train the actual DFQA.

Hernandez-Ortega et al. created the open source FIQAA "FaceQnet" v0 [53] and v1 [48]. As part of the training data preparation for both FaceQnet versions, the BioLab-ICAO framework from Reference [17] was employed to select suitable high-quality images per subject, which were used to compute the ground truth QSs for the subjects' remaining training images. This ground truth QS computation consisted of the normalized Euclidean distances of embeddings produced by a number of FR feature extractors (three for v1; and only one, FaceNet [183], for v0). Both FaceQnet versions were based on a ResNet50 [167] model pretrained for FR using the VGGFace2 [120] dataset, replacing the final output layer with two fully connected layers. Only these two new layers were trained, the rest of the network weights were frozen. FaceQnet v1 extended 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. At the time of writing, FaceQnet v0 and v1 are the

only surveyed approaches that have been included in the report of the new NIST FRVT Quality Assessment campaign [67].

Shi and Jain [52] proposed **PFE (Probabilistic Face Embeddings)**, an approach 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 represent the Gaussian variance and mean, respectively. The work focused on using the uncertainty to improve the FR comparisons, so producing a single scalar QS was not the primary goal. It was nevertheless noted that the uncertainty could be used for FIQA purposes, and a part of the evaluations 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 Reference [52] used a fixed pretrained FR network as basis to compute the FR feature vector (i.e., Gaussian mean) and trained an additional module for the uncertainty vector (i.e., variance), on the same training dataset used for the FR network. The uncertainty module was 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, an **MLS (Mutual Likelihood Score)** was proposed by Reference [52], which weighed and penalized feature dimensions depending on the uncertainty. The uncertainty module training attempted to maximize this MLS for all genuine image pairs. In addition, Reference [52] explained how the uncertainty can be used to fuse embeddings for multiple images.

Zhao et al. [51] trained a CNN for FIQA in a semi-supervised fashion. First, binary labels (good/ bad) were manually assigned to a number of images to train a preliminary version of the DL model. This preliminary network then predicted labels for a different (larger) dataset in the second stage. The third stage updated 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 Reference [51] due to paper length limitations. For all "good" labels predicted by the preliminary network, the label were changed to "bad" if any of these binary constraints were "bad," but existing "bad" label predictions were not altered. This newly labeled dataset was then used in the fourth and final stage to fine-tune the model. Hinge loss was used during training for the binary classification task, but after training the network was modified to output a [0, 1] scalar QS prediction instead. It was noted that the CNN had better computational performance than the CNN proposed by Reference [6].

Terhörst et al. [50] 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 was 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 were used in Reference [50]). 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 Reference [50], the random subnetwork passes used the last two FR layers. The other variant was the "on-top model," meaning that a small additional network was trained with dropout on top of the FR model to transform its FR embeddings. Five layers with dropout were used in the implementation, which included the identity classification layer for training. Removing that, the first and last layer of the network had the same dimensions as the FR embedding. Evaluations used FaceNet [183] and ArcFace [128] for FR, and selected images using QSs from both SER-FIQ variants, FaceQnet v0 [53], an approach proposed by Best-Rowden in Reference [173], three general IQAAs (BRISQUE [79], NIQE [80], PIQE [81]), as well as a COTS system (Neurotec Biometric SDK 11.1 [168]). 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.

Extending the PFE concept of Shi and Jain [52], Chang et al. [49] proposed two methods to learn both uncertainty (variance) and feature (mean) at the same time, without a separate 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 is not 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 Reference [52] does not have to be used, thus avoiding increased computational complexity as evaluated in Reference [49]. One of the two methods in Reference [49] was "classification-based" and learned an entire FR network with both regular feature and uncertainty output, together forming a sampling representation for training, using the reparameterization trick [184] to enable backpropagation. Instead of using the MLS, the cost function consisted of a softmax classification loss, plus a regularization term to control the uncertainty aspect. The latter was 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 Reference [49] was "regression-based" and more akin to the separate uncertainty module training concept of Reference [52]: Similar to Reference [52], it began by using an FR feature network trained in isolation, then the weights were frozen and uncertainty output was added. But in contrast to Reference [52] the FR features (mean) were not frozen with the rest of the pretrained layers, and the method continued 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 Reference [49], the two methods of Reference [49] (using cosine similarity for comparisons) and the method of Reference [52] (using MLS for comparisons, including fusion where applicable) were compared. The "classification-based" method [49] was found to mostly result in better performance increases than the PFE method from Reference [52], while the "regression-based" method [49] appeared either worse or better, depending on the scenario (and further examination in the future was considered due to some observed performance regression with respect to the FR baseline).

Xie et al. [47] proposed the **"PCNet" (Predictive Confidence Network)** FIQAA, and evaluated it, i.a., against the conceptually similar FaceQnet v0 [53]. In contrast to FaceQnet, the network was trained from scratch, a more lightweight ResNet18 [167] was employed, and a different training scheme was used. To obtain comparison scores for FIQA training, a separate ResNet34 was first trained for FR, using cosine similarity for the comparisons. This was done twice, separately for both halves of a dataset, so FR comparison scores were not computed on FR training data. Only mated image pairs were used in the process. The FIQA model, PCNet, was then trained to predict a QS for each image of a pair, the loss being the squared difference between the pair's QS prediction minimum and the pair's previously computed FR comparison score. PCNet (using ResNet18) consistently outperformed FaceQnet v0 [53] and MNet [185] (both using ResNet50) in the evaluations, which, i.a., tested image-to-image verification improvements via ERC plots, and set-to-set verification, with set feature fusion weighted by the per-image quality. In the tests, three open source FR models and VGGFace2 [120] were used.

Chen et al. [46] proposed "ProbFace" based on the **PFE (Probabilistic Face Embeddings)** concept from Shi and Jain [52]. The FR base model was fixed during training, similar to Reference [52]. But instead of an uncertainty vector with the same dimension as the FR feature vector, ProbFace uses a single uncertainty scalar. As a result the required storage space was reduced and the **MLS (Mutual Likelihood Score)** comparison metric from Reference [52] was simplified to an uncertainty-scalar-adjusted cosine FR embedding comparison. In addition, the uncertainty

training was regularized relative to the average uncertainty of each mini-batch, and two uncertainty-aware identification loss variants were introduced to consider both mated and non-mated pairs during training. Of the latter, only one was used for the final ProbFace method configuration, namely, uncertainty-aware triplet loss. Furthermore, ProbFace derived the uncertainty from multiple fused FR base network layers, to more directly incorporate both low-level local texture information and high-level global semantic information. The uncertainty (i.e., quality) assessment aspect was studied mainly in terms of FR comparison improvements against other FR models, so no comparisons against pure FIQAAs were included. ProbFace was, however, also evaluated against the PFE approach from Reference [52] in terms of "risk-controlled face recognition," including an evaluation method akin to ERC, showing that both ProbFace and PFE can be effective in a more general FIQA context.

Ou et al. [45] proposed **SDD-FIQA (Similarity Distribution Distance for FIQA)**, an approach to generate ground truth QS training data by computing the Wasserstein distance between FR comparison score sets that include both mated and non-mated pairs. For this purpose an equal number of mated and non-mated comparison pairs were selected randomly, and the average of multiple computation rounds was used to obtain the final ground truth QS for each image. A FIQA network was then trained with Huber loss using such QS ground truth data. Similar to FaceQnet [48, 53], a pretrained FR network was taken to form the base of the FIQA network, replacing the embedding and classification layer with a fully connected layer for the quality score output and applying 50% dropout, except here the base network part was not frozen during training. The SDD-FIQA network was evaluated on various datasets against FaceQnet v0 [53] and v1 [48], PFE [52], SER-FIQ [50], PCNet [47], as well as three IQAAs (BLIINDS-II [178], BRISQUE [79], PQR [186]). The SDD-FIQA model showed superior performance in most cases. An ERC **"Area Over Curve" (AOC)** measure was also introduced as part of the evaluation, and the influence of the incorporation of non-mated pairs was demonstrated in an ablation study.

The "MagFace" approach from Meng et al. [44] expanded on the idea of FR with integrated FIQA. In contrast to previous approaches such as ProbFace [46], the data uncertainty learning approach from Reference [49], or PFE [52], MagFace does not have separate quality or uncertainty output at all. Instead the quality is directly indicated by the magnitude of the FR feature vector. The approach works by extending the ArcFace [128] training loss, changing the angular margin to a magnitude-aware variant and adding magnitude regularization. On one hand, the magnitude-aware angular margin increases the margin for larger magnitudes, penalizing higher magnitudes for lower-quality samples; and on the other hand, the regularization rewards higher magnitudes scaled by a hyperparameter. As a result, FR feature vectors for higher-quality images are pulled closer to the class center with larger magnitudes, and vice versa for lower-quality samples. The magnitude is bounded during training, so deriving a normalized quality score only requires linear scaling. Furthermore, the design also implies that the FR comparison function after training can be left unchanged from ArcFace [128], while other approaches such as ProbFace [46] and PFE [52] have to specifically include quality in the comparison function to introduce an effect. The magnitude quality aspect itself can be separately used, e.g., to facilitate weighted feature fusion, or for FIQA. MagFace was evaluated both in an FR context and in a FIQA context. The FIQA evaluation included ERC results on multiple datasets against FaceQnet v0 [53], SER-FIQ [50], the data uncertainty learning approach from Reference [49], and the three general IQAAs that were also used in the SER-FIQ [50] evaluation (BRISQUE [79], NIQE [80], PIQE [81]), showing that MagFace can achieve superior or similar FIQA results compared the other methods.

Chen et al. [43] proposed the **identification quality (IDQ)** training loss and the use of knowledge distillation to train a lightweight FIQA network called "LightQNet." The core idea of the IDQ training loss was to concentrate on the FR comparison threshold boundary. Thus, for a mini-batch

with comparison pairs of the same identity, the IDQ loss incorporated an FR threshold hyperparameter to compute pairwise ground truth labels, and the pairwise predicted QS was the minimum of each pair's predicted image QSs (compare with PCNet [47]). The pairwise ground truth labels could be "hard" binary labels, i.e., either above or below the FR threshold hyperparameter, but better performance was achieved with a "soft" exponential-based label variant that used the threshold as an offset, in addition to a scaling hyperparameter. A FIQA branch in a frozen FR network (similar to FaceQnet [48, 53]) and a separate lightweight FIQA network (LightQNet) were trained using IDQ loss. Additionally, using the FIQA branch as a teacher for the lightweight network was shown to improve the lightweight network's predictive performance over pure IDQ loss training. The proposed approach was evaluated against FaceQnet v1 [48], SER-FIQ [50], PFE [52], and PCNet [47] with better or competitive results on various datasets, and substantial (approximately threefold at the lowest) computational performance improvements for the lightweight network were observed.

Fu et al. [42] evaluated a no-reference general IQA CNN [170] for the purposes of FIQA in terms of FR utility. The IQA CNN was applied to various rectangular facial areas in particular, namely, the eyes, nose, and mouth, to examine the areas' individual usefulness for FIQA. These area-specific QSs were also fused by averaging them. In addition to the facial area assessments, the IQA CNN was tested with tightly cropped image variants and image variants aligned for FR input. A clear correlation of the IQA CNN output with FR utility for multiple FR models was demonstrated especially for the eyes area on the VGGFace2 [120] dataset, although results could not compete with the tested specialized monolithic FIQAAs (learning to rank [60], FaceQnet v0 [53], SER-FIQ [50], and MagFace [44]).

Fu et al. [41] investigated the effect of face masks on a number of monolithic FIQAAs, namely, the learning to rank approach [60], FaceQnet v1 [48], SER-FIQ [50], and MagFace [44]. The FIQAA performance was tested on regular face images without masks, on images with real masks of varying types, and on images with synthesized masks. Synthetic masks were automatically drawn based on detected facial landmarks in a variety of solid colors, i.e., without realistic shading, on top of images without masks. Results showed a drop in predicted QSs for images with masks for all tested FIQAAs, corresponding to reduced FR performance of both automatic systems and human experts, and the QS distributions for images with/without masks were especially distinct for MagFace [44] and the learning to rank approach [60]. Differences between results for the real and the synthetic masks were observed as well, indicating that improved synthesis realism may be desirable for this kind of evaluation in the future. Additionally, network attention visualizations were examined for FaceQnet v1 [48] and MagFace [44]. Note that this work was purely about the evaluation of existing FIQAAs, thus its categorization is based on the included FIQAAs instead of proposed FIQAAs.

Fu et al. [40] further evaluated the FR utility prediction performance of 6 monolithic FIQAAs [44, 45, 48, 50, 52, 60], 10 general IQAAs (i.a., the IQA CNN from Reference [170], BRISQUE [79], NIQE [80], PIQE [81]), and 9 factor-specific hand-crafted measures (i.a., for blur, symmetry, inter-eye distance). Most of the general IQAAs did improve FR performance in ERC tests, but overall they were outperformed by the best monolithic FIQAAs. The factor-specific hand-crafted measure results were inconsistent across datasets, indicating that these individual measures do not generalize sufficiently. Assessments from the hand-crafted measures also did not correlate strongly with the other IQAA/FIQAA assessments, while various IQAAs and FIQAAs did exhibit higher assessment overlaps. Network attention visualizations for some of the DL IQAAs and FIQAAs illustrated that the tested IQAAs incorporated more image background information than the FIQAAs, which concentrated more on the face region. Similarly to Reference [41], note that this work focused on the evaluation of existing FIQAAs, thus its categorization is based on the included FIQAAs instead of proposed FIQAAs.

5 EVALUATION

The first subsection hereunder introduces a common methodology to evaluate FIQAAs (or other biometric quality assessment algorithms) with respect to their ability to assess the biometric utility of samples for a given FR system and dataset. In the second subsection, we present a concrete evaluation for 14 FIQAAs and discuss the evaluation configuration, results, and limitations.

5.1 Error-versus-Reject-Characteristic

An **Error-versus-Reject-Characteristic (ERC)** can be plotted to evaluate the predictive performance of quality assessment algorithms, as proposed by Grother and Tabassi [187]. In the FIQA literature the "C" in ERC is occasionally also referred to as "Curve." It is currently intended to standardize the ERC concept in the next (third) edition of ISO/IEC 29794-1 [188] under a different name that replaces the "reject" term to avoid confusion with the meaning of "reject" in ISO/IEC 2382-37 [70].

In the context of FIQA, an FR system and a face dataset with subject identity labels is required in addition to the FIQAA to compute the ERC. The FR system compares face image pairs with a fixed comparison threshold [70] to decide between match [70] or non-match [70] (depending on the ERC error type) for each pair. QSs produced by the FIQAA 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 QSs, and an FR error measure is calculated for the resulting QS subsets. In Reference [187], it is suggested that the **FNMR (False Non-Match Rate)** [70] error measure should be used as the primary performance indicator. If desired, the FR threshold can then be derived for a fixed **FMR (False Match Rate)** [70] on the unfiltered image pairs—or vice versa if the FMR was plotted as the error measure. The error is typically plotted on the vertical axis. The rejected fraction, plotted on the horizontal axis, denotes the relative amount of images (0 to 100%) rejected based on the QS. Plotting this fraction instead of the increasing QS threshold normalizes the axis independently of the given FIQAA. This also means that QSs do not have to be constrained to a certain range, only their order is important.

Note that ERCs should usually represent the rejection of samples/images, not individual comparisons, so all comparisons with quality below the currently considered quality threshold have to be discarded simultaneously. This means that the horizontal axis actually denotes the maximum of the fraction of images rejected via the quality threshold, not the precise fraction of rejected images. This in turn means that ERC plots should prefer stepwise interpolation by continuing the error value from the last real ERC data point at which a batch of comparisons was rejected. Linear interpolation, as used by some works, can be misleading for rejection fraction ranges with low-quality granularity, which may occur for realistic evaluation configurations.

Olsen et al. [189] 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 concretely, Reference [189] proposed to compute the AUC for the full [0, 100%] range, and a **partial AUC (pAUC)** to focus only on the [0, 20%] range. The "area under theoretical best" term refers to the (unrealistic) best case where the error value decrease equals the rejected fraction percentage. Also note that the "area under theoretical best" is a constant value for a specific AUC range, so subtracting it from the FIQAAs' ERC curve areas will not alter their relative performance within that specific AUC range. Consequently, the subtraction can be omitted for AUC computations when only the per-AUC-range FIQAA ranking is analyzed (which is the case for Section 5.2).

A more realistic approximation of an optimal FIQAA may be achieved by means of an oracle, the concept of which was described by Phillips et al. [16]. Since more recent FIQA literature did not continue to explore this, future work could do so in an attempt to improve ERC evaluations. Conversely, the error at 0% rejection can be considered as the practical worst-case, because the average of many/infinite ERC curves for random QSs will approximately result in no error change for FNMR or FMR, and no real FIQAA should be worse than random QS assignment.

The FIQAA literature listed in this survey did not always provide ERC or AUC evaluation results. For example, some works evaluated the FIQAA in terms of quality label prediction performance and did not evaluate the FIQAA 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 6 for further discussions regarding this and other issues.

The ongoing NIST **Face Recognition Vendor Test (FRVT)** for face image quality assessment [67] evaluates FIQAAs combined with a number of FR algorithms and dataset types, showing results, i.a., in the form of ERC plots. Some noteworthy modifications to the usual ERC methodology were applied according to the current draft report [67]: To compute the FNMR at some rejection fraction, the evaluation divided by the count of comparisons at that rejection fraction (i.e., comparisons not removed by the quality threshold), instead of dividing by the total comparison count constant independently of the rejection fraction. QSs were perturbed with random uniformly distributed noise as a result tie breaker, and a logarithmic rejection axis was plotted to emphasize the results for smaller rejection fractions. The report furthermore introduced the **"Incorrect Sample Rejection Rate" (ISRR)** and **"Incorrect Sample Acceptance Rate" (ISAR)**, which are defined to incorporate both FR comparisons and QS rejections. A future goal of the project is to investigate (non-linear) calibration methods to map QSs of different FIQAAs to a common [0, 100] range with approximately equalized distribution.

5.2 Selective Evaluation

We conducted an FNMR ERC evaluation with 14 FIQA approaches, including both recent methods and general IQAAs, and at least one method for each data aspect category (described in Section 3.2), except for human quality ground truth training:

- Hand-crafted (**Dhc**):
  - Pose symmetry, Light symmetry, Blur, Sharpness, Exposure, **GCF (Global Contrast Factor)**: As described by Wasnik et al. [8] and ISO/IEC TR 29794-5:2010 [66].
  - PIQE [81]: Publicly available Python implementation.
- Utility-agnostic training (**Duat**):
  - BRISQUE [79]: Publicly available model (pybrisque implementation).
  - NIQE [80]: Publicly available model (scikit-video implementation).
- FR-based ground truth training (**Dfrt**):
  - FaceQnet v0 [53] & v1 [48]: Publicly available models.
  - PCNet [47]: Model provided by the authors.
- FR-based inference (**Dfri**):
  - SER-FIQ [50]: Publicly available model ("same model" variant on ArcFace, which is also used for FR).
- FR-integration (**Dint**):
  - MagFace [44]: Publicly available model ("iResNet100" backbone, trained on MS1MV2 [128]).
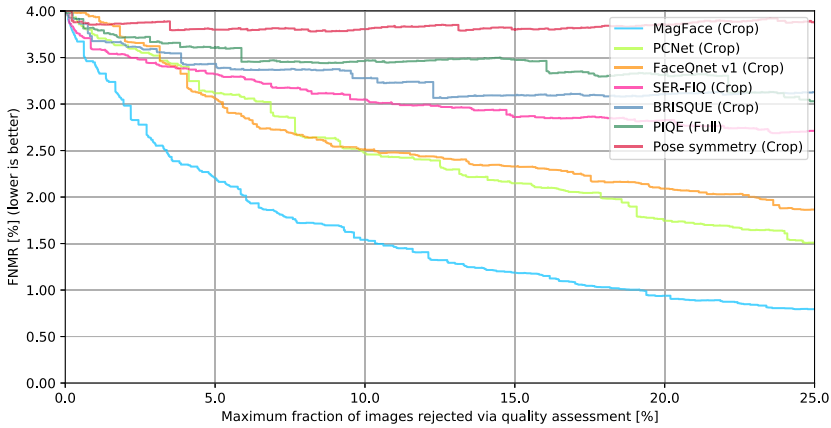
Fig. 7. ERC plot for a subset of the evaluated FIQAAs. EDC pAUC results for all evaluated FIQAAs are provided in Table 5.

The error at 0% rejection is set to 4% FNMR. ArcFace [128] was used for FR (cosine similarity), and RetinaFace [190] was used for face/facial landmark detection, employing the publicly available models for both (InsightFace's "LResNet100E-IR,ArcFace@ms1m-refine-v2" & "RetinaFace-R50").

We used **LFW (Labeled Faces in the Wild)** [118] as the evaluation dataset and consider all possible mated pairs therein. As shown by Table 2, LFW [118] is the dataset that has been employed by the greatest number of FIQA works, including recent ones. The FR performance on LFW [118] appears to already be almost saturated by the state-of-the-art systems, and the quality distribution correspondingly seems to be more narrow than in, e.g., IJB-C [126], as demonstrated most recently for MagFace [44]. This conversely means that LFW [118] is more challenging for FIQA ERC evaluations, since FIQAAs have to more effectively rank images in terms of biometric utility to decrease the error rate, especially for lower rejection fractions.

Figure 7 shows the ERC plot, but only with a subset of the FIQAAs for the sake of legibility, since multiple curves for the less-well-performing methods would approximately overlap graphically. Table 5, however, lists ranked ERC pAUC results for all 14 FIQAAs, which is a more useful representation for the analysis of many methods.

LFW [118] images depict a substantial amount of background information besides the actual face, so the type of preprocessing is relevant. For this evaluation, the ERC was computed for all FIQAA methods using both the full images (marked as "Full") and preprocessed variants (marked as "Crop"). The preprocessing variant used RetinaFace [190] to crop the images to the face and to subsequently align the images to the detected facial landmarks (there are no edge cases without a detected face or landmarks). Only the best-performing variants per FIQAA at 1% pAUC are shown, to avoid cluttered results. All **Dfrt**/**Dfri**/**Dint** approaches performed better with the preprocessed images, as to be expected due to their incorporation of FR during training and/or inference. A few of the **Dhc**/**Duat** approaches did, however, yield better results using the full images even for the considered 1% pAUC. This applies to more **Dhc**/**Duat** approaches for higher pAUC maxima (e.g., for "Light symmetry" at 20% pAUC), but the difference is never substantial enough to compete with the top-ranking FIQAAs regardless, so we do not include more detailed results regarding this.

As apparent in Table 5 and Figure 7, MagFace [44] has distinctly achieved the best results throughout this particular evaluation. The ranking of the other FIQAAs depends on the considered pAUC range: For 5% and 20% pAUC, the five **Dfrt**/**Dfri**/**Dint** approaches outperformed all

Table 5. ERC Evaluation Results in Terms of the Partial Area-Under-Curve (pAUC) Values
for Reject Fraction Ranges from 0% to 1%/5%/20%

FIQAA		1% pAUC		5% pAUC		20% pAUC
MagFace (Crop)	1:	0.00% (0.037%)	1:	0.00% (0.144%)	1:	0.00% (0.356%)
SER-FIQ (Crop)	2:	30.69% (0.038%)	3:	56.54% (0.176%)	5:	61.00% (0.625%)
FaceQnet v0 (Crop)	3:	51.52% (0.038%)	2:	54.85% (0.175%)	4:	57.92% (0.612%)
BRISQUE (Crop)	4:	59.35% (0.039%)	6:	66.27% (0.181%)	6:	69.18% (0.662%)
PCNet (Crop)	5:	61.10% (0.039%)	4:	60.35% (0.178%)	2:	40.46% (0.535%)
PIQE (Full)	6:	69.29% (0.039%)	8:	75.74% (0.186%)	8:	78.23% (0.702%)
Pose symmetry (Crop)	7:	71.78% (0.039%)	10:	87.60% (0.193%)	11:	92.63% (0.765%)
Blur (Crop)	8:	74.32% (0.039%)	7:	70.36% (0.183%)	7:	75.66% (0.690%)
Sharpness (Crop)	9:	82.52% (0.039%)	9:	78.77% (0.188%)	9:	86.17% (0.737%)
FaceQnet v1 (Crop)	10:	94.37% (0.040%)	5:	65.15% (0.180%)	3:	43.40% (0.548%)
Exposure (Crop)	11:	95.55% (0.040%)	13:	99.81% (0.200%)	14:	100.00% (0.798%)
NIQE (Full)	12:	95.74% (0.040%)	11:	96.42% (0.198%)	12:	96.40% (0.782%)
Light symmetry (Crop)	13:	97.94% (0.040%)	12:	97.78% (0.198%)	13:	97.94% (0.789%)
GCF (Full)	14:	100.00% (0.040%)	14:	100.00% (0.200%)	10:	91.82% (0.762%)

For each entry, e.g., "1: 0.00% (0.037%)," the first number denotes the ranking of the FIQAA ("1:" being
the best, "14:" the worst), the second number shows the relative performance ("0.00%" being the best,
"100.00%" the worst), and the bracketed third number is the actual absolute pAUC value (higher being
worse). ERC results are also plotted for a subset of the FIQAAs in Figure 7.

Dhc/**Duat** methods, but for 1% pAUC only three **Dfrt**/**Dfri**/**Dint** held the top rankings, BRISQUE
[79] being able to compete most closely with them. Note that we do not include results for higher
ERC rejection fractions, since these are less interesting from an operational perspective. In prac-
tice, one would not want to reject, e.g., every second image (50% rejection fraction) if the images
mostly have high FR utility—as is the case for LFW [118], according to the aforementioned high
state-of-the-art FR performance.

While issues and challenges in general are discussed in the subsequent section, it is also impor-
tant to highlight limitations of this particular evaluation, which can be used to show what future
work may want to consider:

- Only a low amount of FIQA/FR/dataset/preprocessing/hyperparameter configurations was
  tested in contrast to the available options. A more comprehensive literature evaluation will
  require re-implementation efforts for the many listed works that did not provide open refer-
  ence implementations, and automated configuration exploration may have to be employed
  to overcome a combinatorial explosion. Static ERC plots can quickly become too cluttered,
  as was already the case here with only 14 FIQAAs, but reduced or interactive plots can still
  be useful, and derived metrics such as pAUC can be used to analyze arbitrary configuration
  counts.
- No non-mated pairs were considered, since only the FNMR was used as ERC error. It could
  be interesting to test FMR, or possibly other metrics, as the error—especially because re-
  cent FIQA approaches have started to incorporate both mated and non-mated pairs during
  training.
- While the use of publicly available models is beneficial in terms of reproducibility, the com-
  parisons are not as fair as they could be due to differing training data. For **Dfrt**/**Dfri**/**Dint**,
  the different training data does imply that the results may not fairly reflect the potential
  of the underlying FIQA concepts or network architectures. Different preprocessing during
  training or different training time could likewise affect the performance. Note that this means
  that black-box FIQAAs (e.g., COTS systems) cannot be fairly compared by definition. Com-
  parisons between **Dfrt**/**Dfri**/**Dint** and **Dhc**/**Duat** approaches in this evaluation are, however,

rather unproblematic, since **Duat** approaches typically require different training data by design (e.g., general IQA training data instead of face images), and **Dhc** requires no training.

- The computational performance of FIQAAs could be relevant in practice, too, thus evaluations could be helpful. For anything except **Dfri**/**Dint** approaches, the computational performance should by definition be independent of the FR system choice, and dataset/preprocessing configuration may also be unimportant in this context. However, hardware configurations (i.a., CPU versus GPU) can matter instead, and implementation details have to be considered as well (i.e., concrete implementations may not utilize the given hardware as effectively as the underlying concepts would allow).

6 OPEN ISSUES AND CHALLENGES

An obvious challenge consists of the further improvement of FIQA methods in terms of predictive and computational performance. For deep learning FIQA 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 [191]. Naturally, FIQA with the goal of generating quality scores that predict FR utility [68] 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 7 highlights the identified key challenges.

6.1 Comparability and Reproducibility

As previously noted in Section 5, it would be challenging to comprehensively compare the performance of the surveyed FIQA approaches, since the evaluations presented in the literature differ in multiple aspects that would need to be aligned to facilitate fair direct comparisons:

- **Datasets:** As shown in Table 2, a variety of datasets were 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., Reference [48] or [54] regarding the VGGFace2 dataset [120]) or modified the data, e.g., by synthetically degrading images via increased blur or contrast (see, e.g., Reference [12]). Where training data is required for the FIQAA, the chosen subdivision of the datasets into training and test data also influences the evaluation results. Furthermore, various works assigned ground truth quality scores or labels to the dataset for FIQAA training and/or for the evaluation. When FIQA is evaluated in terms of FR performance improvements, the selection of image pairs that are considered initially for FR comparisons [70] (i.e., before filtering them via FIQA decisions) may alter the results as well. Another potentially interesting question is the degree of existing overlap between datasets regarding FR, which could be studied both in a general FR context and in the context of FIQA.

- **Evaluation methods:** Different evaluation methods and ways to report results are used among the literature. Some FIQA approaches are only tested by comparing predicted quality scores or labels against a given ground truth (e.g., assigned by humans), i.e., not all of the literature evaluates FIQA in terms of FR utility [68, 69] in the first place. Instead of evaluating the FIQA on its own, some literature that included image enhancement steps in the evaluation. For FR performance improvement evaluations via an ERC as described in Section 5.1, the FR comparison score threshold [70] and the error type configuration can differ between evaluations, which also applies to ERC-derived AUC results. Some of the works evaluated FIQA performance exclusively by means other than the ERC—for example, FR performance was evaluated for four FIQA-derived quality bins in Reference [61].

- **FR algorithms:** Evaluating FIQA in terms of FR performance improvement is desirable to examine how well quality scores of a FIQAA reflect FR utility [68], but this also introduces the FR algorithm choice for feature extraction [70] and comparison [70] as another evaluation factor. Furthermore, there are FIQA approaches among the literature that are conceptually based on FR models to begin with (see, e.g., Reference [50]), and FR algorithms are used by various works to establish ground truth quality scores/labels (see, e.g., Reference [48] for scores or Reference [61] for labels in the form of four quality bins). Last, some literature exclusively used anonymous and/or closed-source FR systems, which can limit reproducibility and expandability (see, e.g., Reference [61]).

Due to the amount of existing and possible FIQA evaluation configurations, the comparison of FIQAAs can be considered as a key challenge. This open issue could be limited in scope, e.g., by only considering FIQA 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 submit the presented FIQAAs to an evaluation campaign where all algorithms are assessed under the same benchmark, such as the previously mentioned NIST FRVT Quality Assessment evaluation [67]. Open evaluation protocols could be established as well.

Another solution is to publicly provide the FIQAA 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 algorithm 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 FIQA in particular depend on the used training data, publishing source code is preferable to pure black box releases. 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 FIQA 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 did not appear to publish accompanying source code (irrespective of the implementation complexity), more recent deep learning FIQA works tend to do so, with code being publicly available for, e.g., FaceQnet [48, 53], **PFE (Probabilistic Face Embeddings)** [52], SER-FIQ [50], and MagFace [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 FIQAAs, evaluating the computational performance may be of relevance as well. This is rarely considered in the surveyed FIQA literature. Computational performance tests 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 utility prediction performance evaluations. Other factors do, however, become relevant, namely, the computational optimization of the FIQAA, as well as the used hardware and the robustness of the time measurements. Besides measuring inference time, a different kind of computational performance tests could assess the efficiency of FIQA model training as well, although this is less relevant in an operational context as long as no frequent (re-)training is required.

6.2 Explainability and Interpretability

While the more recent monolithic deep learning FIQA approaches are trained specifically to output quality scores in terms of FR utility [69, 70], they are not 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, FIQA models should be able to predict FR utility [70] while also providing useful feedback regarding quality-degrading causes. Future work could thus attempt to improve upon this area, perhaps by adding visualizations based on 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 [192, 193], we expect growing interest in analyzing the behavior of FIQA methods for different population groups and the development of FIQA methods more transparent [194] and agnostic to selected covariates [195].

6.3 Use of Synthetic Data

For FIQA 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), such that the robustness can be verified for various cases with a high certainty. Existing images can also be degraded synthetically—this was done in a few works (e.g., Reference [12]). That is, both known techniques from prior work, such as Gaussian blurring, and more sophisticated techniques, such as deep learning style transfer, could be employed in the future. It is also possible to generate fully synthetic face images (see, e.g., StyleALAE [196]), which is a strategy that has not been used in the surveyed FIQA 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 [124], which has been used in some of the FIQA literature as well). This latter point is, however, not entirely clear, since deep learning face synthesis itself is typically trained on real face images.

6.4 Interoperability

Examining and improving interoperability in terms of FIQA FR utility prediction generality could be another goal for future work. While this may partially stand in conflict with the goal of maximizing 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 [53] to using three in v1 [48].

6.5 Vulnerabilities

Specific attacks on FIQA may be investigated in future works. For instance, the surveyed machine learning FIQA literature did not study adversarial attacks, i.e., attacks that specifically modify the input (physical [197] or digital after being captured and processed [198]) to confuse the FIQA model.

6.6 Standardization

ISO/IEC 29794-1:2016 [68] defines the notion of biometric sample quality, and a new edition is currently under development [188]. At the time of this writing, this new edition will, i.a., standardize ERC (Section 5.1) for FIQAA evaluation. ISO/IEC TR 29794-5:2010 [66] describes various

actionable FIQA measures, and the next edition is under development as an International Standard [199]. Current portrait quality specifications are established in ISO/IEC 39794-5:2019 [76], which contains content from the ICAO Portrait Quality TR [84], which in turn was based on parts of ISO/IEC 19794-5:2011 [86], ISO/IEC 19794-5:2005 [160], and ICAO Doc 9303 [159]. ISO/IEC 24358 ("Face-aware capture subsystem specifications") [200] is another relevant standard that is under development at the moment. An important future goal for FIQA is the standardization of some particular FIQA algorithm/model, analogous to the normative standardization of the open source **NIST Fingerprint Image Quality (NFIQ)** 2 as part of ISO/IEC 29794-4:2017 [201].

6.7 Further Applications

As described in Section 2.5, there are further application areas that were barely or not at all examined in the surveyed literature. For example, lossy compression control was not considered at all, although compression artifacts are mentioned as a quality degrading factor by various works. FIQA for other areas besides face recognition can also be explored further, including FIQA in the context of gender or other soft biometrics recognition [73], attention level estimation [72], emotion analysis [71], and so on.

Almost all of the found FIQA literature focused the visible spectrum. The exception is the work by Long et al. [202], which studied quality assessment for near-infrared face video sequences. They combined measures for sharpness, brightness, resolution, landmark-based head pose, and expression in terms of eyes/mouth being open/closed, but the evaluation was limited to comparisons against human rankings. Future work could thus quickly expand on FIQA for near-infrared images or for other spectra [203].

Furthermore, FIQA may also be relevant for face "depth" or "range" images, i.e., 2D images depicting 3D positions in terms of depth. But, similar to non-visible-spectrum FIQA, few works appear to exist that consider depth FIQA in a biometric context. The work by Lin and Chen [204] is one instance that included depth FIQA using a deformable shape model to identify excessive expression variations, and the FIQA part was used to improve 3D FR performance via sample rejection with a fixed quality threshold. Future work could further explore depth FIQA or 3D face quality assessment for other 3D representations. Combinations of, e.g., visible spectrum and depth images for FIQA, as well as FIQA for other application areas such as biometric depth image enhancement [205], could be investigated as well.

Another related field that future FIQA research may want to consider is face sketch recognition/synthesis, where literature published so far appears to be focused on perceptual measures instead of biometric utility prediction [206], a concrete recent example being the work by Fan et al. [207].

7 SUMMARY

Face image quality assessment is an active research area and can be used for a variety of application scenarios such as filtering and feedback during the acquisition process or for database maintenance and monitoring. The literature surveyed in this work predominantly focused on evaluating the proposed FIQA approaches either in terms of predictive performance with respect to given ground truth quality score labels or in terms of utility [69, 70] for the purpose of aiding face recognition by discarding images based on the assessed quality or some kind of quality-based processing or fusion [94]. Automatic face quality assessment is especially relevant for FR as part of large-scale systems, e.g., the European **Schengen Information System (SIS)**, the **VISA Information System (VIS)**, the **Entry Exit System (EES)**, or the US **ESTA (Electronic System for Travel Authorization)**, due to the amount of data and the multitude of different acquisition locations/devices.

A progression over time towards monolithic deep learning approaches was observed in the FIQA literature. Older methods were predominantly factor-specific and independent of concrete FR systems, while more recent methods tended to train on ground truth quality scores derived from FR comparisons. Some of the most recently emerging monolithic methods expanded on the FR focus, either by relying on FR systems during inference or by integrating FIQA into FR models.

One key challenge is to facilitate comparability of the FIQA evaluations, since many differing evaluation configurations were employed in the literature. Thus, future work should preferably provide the implementations of the proposed FIQAAs publicly, especially in the form of source code, enabling evaluations in later works to more easily include these FIQA approaches. The more recent works have begun to do so, but re-implementation efforts will be required if many of the older approaches are to be evaluated comprehensively. There also is the ongoing NIST FRVT Quality Assessment evaluation [67], to which FIQAAs can be submitted. Besides evaluating the predictive capabilities of FIQAAs, more attention could be paid to computational performance evaluations in the future.

Another key challenge is to improve the interpretability of deep learning–based FIQA, which so far mostly fell into the monolithic category of this survey, meaning that these modern approaches did not focus on providing extensive feedback for human operators to adjust acquisition conditions for increased biometric utility.

Of course, there also is the key challenge of further improving performance in terms of both utility and computational workload (e.g., with new deep learning network architectures), as well as improving robustness/decreasing bias [192, 193] (e.g., via the selection or synthetic extension of datasets for different quality degradation cases), which naturally is dependent on suitable evaluation methodologies.

In the long term, an important objective is the standardization of a specific FIQA approach, analogous to the normative standardization of the open source **NIST Fingerprint Image Quality (NFIQ)** 2 as part of ISO/IEC 29794-4:2017 [201], and advances regarding the aforementioned challenges can help to achieve this. Various other application scenarios can be explored further as well, e.g., FIQA-guided image enhancement or compression.

REFERENCES

[1] O. Henniger, B. Fu, and C. Chen. 2020. On the assessment of face image quality based on handcrafted features. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*. Gesellschaft für Informatik e.V., 273–280.

[2] J. Rose and T. Bourlai. 2020. 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*. Springer International Publishing, 61–76.

[3] Z. Lijun, S. Xiaohu, Y. Fei, D. Pingling, Z. Xiangdong, et al. 2019. Multi-branch face quality assessment for face recognition. In *Proceedings of the 19th International Conference on Communication Technology (ICCT)*. IEEE, 1659–1664.

[4] J. Rose and T. Bourlai. 2019. Deep learning based estimation of facial attributes on challenging mobile phone face datasets. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ACM, 1120–1127.

[5] A. Khodabakhsh, M. Pedersen, and C. Busch. 2019. Subjective versus objective face image quality evaluation for face recognition. In *Proceedings of the 3rd International Conference on Biometric Engineering and Applications (ICBEA)*. ACM Press, 36–42.

[6] J. Yu, K. Sun, F. Gao, and S. Zhu. 2018. Face biometric quality assessment via light CNN. *Pattern Recog. Lett.* 107 (May 2018), 25–32.

[7] C. Wang. 2017. A learning-based human facial image quality evaluation method in video-based face recognition systems. In *Proceedings of the 3rd International Conference on Computer and Communications (ICCC)*. IEEE, 1632–1636.

[8] P. Wasnik, K. B. Raja, R. Raghavendra, and C. Busch. 2017. Assessing face image quality for smartphone based face recognition system. In *Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[9] L. Zhang, L. Zhang, and L. Li. 2017. Illumination quality assessment for face images: A benchmark and a convolutional neural networks based model. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*. Vol. 10636. Springer International Publishing, 583–593.

[10] H. I. Kim, S. H. Lee, and Y. M. Ro. 2015. Face image assessment learned with objective and relative face image qualities for improved face recognition. In *Proceedings of the International Conference on Image Processing (ICIP)*. IEEE, 4027–4031.

[11] N. Damer, T. Samartzidis, and A. Nouak. 2015. 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, 85–98.

[12] A. Abaza, M. A. Harrison, T. Bourlai, and A. Ross. 2014. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biomet.* 3, 4 (Dec. 2014), 314–324.

[13] H. I. Kim, S. H. Lee, and Y. M. Ro. 2014. Investigating cascaded face quality assessment for practical face recognition system. In *Proceedings of the International Symposium on Multimedia (ISM)*. IEEE, 399–400.

[14] R. Raghavendra, K. B. Raja, B. Yang, and C. Busch. 2014. Automatic face quality assessment from video using gray level co-occurrence matrix: An empirical study on automatic border control system. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 438–443.

[15] M. Nikitin, V. Konushin, and A. Konushin. 2014. Face quality assessment for face verification in video. *Proceedings of the 24th International Conference on Computer Graphics and Vision (GraphiCon)*.

[16] P. J. Phillips, J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, et al. 2013. On the existence of face quality measures. In *Proceedings of the 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

[17] M. Ferrara, A. Franco, D. Maio, and D. Maltoni. 2012. Face image conformance to ISO/ICAO standards in machine readable travel documents. *IEEE Trans. Inf. Forens. Secur.* 7, 4 (Aug. 2012), 1204–1213.

[18] F. Hua, P. Johnson, N. Sazonova, P. Lopez-Meyer, and S. Schuckers. 2012. Impact of out-of-focus blur on face recognition performance based on modular transfer function. In *Proceedings of the 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 85–90.

[19] A. Abaza, M. A. Harrison, and T. Bourlai. 2012. Quality metrics for practical face recognition. *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)* (2012), 5.

[20] P. Liao, H. Lin, P. Zeng, S. Bai, H. Ma, et al. 2012. Facial image quality assessment based on support vector machines. In *Proceedings of the International Conference on Biomedical Engineering and Biotechnology (ICBEB)*. 810–813.

[21] K. Nasrollahi and T. B. Moeslund. 2011. Extracting a good quality frontal face image from a low-resolution video sequence. *IEEE Trans. Circ. Syst. Vid. Technol.* 21, 10 (Oct. 2011), 1353–1362.

[22] M. De Marsico, M. Nappi, and D. Riccio. 2011. Measuring measures for face sample quality. In *Proceedings of the 3rd International ACM Workshop on Multimedia in Forensics and Intelligence (MiFor)*. ACM Press, 7.

[23] D. Rizo-Rodriguez, H. Méndez-Vázquez, and E. Garcia-Reyes. 2010. An illumination quality measure for face recognition. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*. IEEE, 1477–1480.

[24] J. R. Beveridge, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, et al. 2010. Quantifying how lighting and focus affect face recognition performance. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 74–81.

[25] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, D. S. Bolme, et al. 2010. FRVT 2006: Quo vadis face quality. *Image Vis. Comput.* 28, 5 (May 2010), 732–743.

[26] J. Sang, Z. Lei, and S. Z. Li. 2009. Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5. In *Proceedings of the 3rd International Conference on Biometrics (ICB)*, Vol. 5558. Springer Berlin, 229–238.

[27] G. Zhang and Y. Wang. 2009. Asymmetry-based quality assessment of face images. In *Proceedings of the 5th International Symposium on Advances in Visual Computing (ISVC)*, Vol. 5876. Springer Berlin /, 499–508.

[28] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. 2008. Focus on quality, predicting FRVT 2006 performance. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. IEEE, 1–8.

[29] E. A. Rúa, J. E. L. Alba Castro, and C. García Mateo. 2008. Quality-based score normalization and frame selection for video-based person authentication. In *Biometrics and Identity Management (BioID)*, Vol. 5372. Springer Berlin, 1–9.

[30] K. Nasrollahi and T. B. Moeslund. 2008. Face quality assessment system in video sequences. In *Biometrics and Identity Management (BioID)*, Vol. 5372. Springer Berlin, 10–18.

[31] A. Fourney and R. Laganiere. 2007. Constructing face image logs that are both complete and concise. In *Proceedings of the 4th Canadian Conference on Computer and Robot Vision (CRV)*. IEEE, 488–494.

[32] X. Gao, S. Z. Li, R. Liu, and P. Zhang. 2007. Standardization of face image sample quality. In *Proceedings of the 2nd IAPR International Conference on Biometrics (ICB)*.

[33] M. Abdel-Mottaleb and M. H. Mahoor. 2007. Algorithms for assessing the quality of facial images. *IEEE Comput. Intell. Mag.* 2, 2 (May 2007), 10–17.

[34] R. L. V. Hsu, J. Shah, and B. Martin. 2006. Quality assessment of facial images. In *Proceedings of the Biometrics Symposium: Special Session on Research at the Biometric Consortium Conf. (BCC)*. IEEE, 1–6.

[35] K. Kryszczuk and A. Drygajlo. 2006. On combining evidence for reliability estimation in face verification. In *Proceedings of the 14th European Signal Processing Conference (EUSIPCO)*.

[36] K. Kryszczuk and A. Drygajlo. 2006. On face image quality measures. In *Proceedings of the 2nd Workshop on Multimodal User Authentication (MMUA)*.

[37] M. Subasic, S. Loncaric, T. Petkovic, H. Bogunovic, and V. Krivec. 2005. Face image validation system. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)*. IEEE, 30–33.

[38] Z. Yang, H. Ai, B. Wu, S. Lao, and L. Cai. 2004. Face pose estimation and its application in video shot selection. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*. IEEE, 322–325.

[39] H. Luo. 2004. A training-based no-reference image quality assessment algorithm. In *Proceedings of the International Conference on Image Processing (ICIP)*. IEEE, 2973–2976.

[40] B. Fu, C. Chen, O. Henniger, and N. Damer. 2022. A deep insight into measuring face image utility with general and face-specific image quality metrics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[41] B. Fu, F. Kirchbuchner, and N. Damer. 2021. The effect of wearing a face mask on face image quality. In *Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition*.

[42] B. Fu, C. Chen, O. Henniger, and N. Damer. The relative contributions of facial parts qualities to the face image utility. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)* (2021). Gesellschaft für Informatik e.V., 1–5.

[43] K. Chen, T. Yi, and Q. Lv. 2021. LightQNet: Lightweight deep face quality assessment for risk-controlled face recognition. *IEEE Signal Processing Letters*.

[44] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. 2021. MagFace: A universal representation for face recognition and quality assessment. (Mar. 2021). arXiv:cs/2103.06627.

[45] F. Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, et al. 2021. SDD-FIQA: Unsupervised face image quality assessment with similarity distribution distance. (Mar. 2021). arXiv:cs/2103.05977.

[46] K. Chen, Q. Lv, T. Yi, and Z. Yi. 2021. Reliable probabilistic face embeddings in the wild. (Feb. 2021). arXiv:cs/2102.04075.

[47] W. Xie, J. Byrne, and A. Zisserman. 2020. Inducing predictive uncertainty estimation for face recognition. *Proceedings of the British Machine Vision Conference (BMVC)*.

[48] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay. 2020. Biometric quality: Review and application to face recognition with FaceQnet. (June 2020). arXiv:2006.03298.

[49] J. Chang, Z. Lan, C. Cheng, and Y. Wei. 2020. Data uncertainty learning in face recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (Mar. 2020). arXiv:2003.11339.

[50] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. 2020. SER-FIQ: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (Mar. 2020). arXiv:2003.09373.

[51] X. Zhao, Y. Li, and S. Wang. 2019. Face quality assessment via semi-supervised learning. In *Proceedings of the 8th International Conference on Computing and Pattern Recognition (ICCPR)*. ACM, 288–293.

[52] Y. Shi and A. K. Jain. 2019. Probabilistic face embeddings. In *Proceedings of the International Conference on Computer Vision (ICCV)* (Aug. 2019). arXiv:1904.09658.

[53] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. 2019. FaceQnet: Quality assessment for face recognition based on deep learning. In *Proceedings of the 12th IAPR International Conference on Biometrics (ICB)* (Apr. 2019). arXiv:1904.01740.

[54] F. Yang, X. Shao, L. Zhang, P. Deng, X. Zhou, et al. 2019. DFQA: Deep face image quality assessment. In *Proceedings of the 10th International Conference on Image and Graphics (ICIG)*, Vol. 11902. Springer International Publishing, 655–667.

[55] P. Wasnik, R. Raghavendra, K. Raja, and C. Busch. 2018. An empirical evaluation of deep architectures on generalization of smartphone-based face image quality assessment. In *Proceedings of the 9th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE.

[56] X. Qi, C. Liu, and S. Schuckers. 2018. Boosting face in video recognition via CNN based key frame extraction. In *Proceedings of the International Conference on Biometrics (ICB)*. IEEE, 132–139.

[57] L. Best-Rowden and A. K. Jain. 2017. Automatic face image quality prediction. (June 2017). arXiv:cs/1706.09887.

[58] X. Hu, L. Zhuo, L. Zhang, and X. Li. 2016. Face image illumination quality assessment for surveillance video using KPLSR. In *Proceedings of the International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 330–335.

[59] S. Vignesh, K. Manasa Priya, and S. S. Channappayya. 2015. Face image quality assessment for face selection in surveillance video using convolutional neural networks. In *Proceedings of the 3rd IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 577–581.

[60] J. Chen, Y. Deng, G. Bai, and G. Su. 2015. Face image quality assessment based on learning to rank. *IEEE Sig. Process. Lett.* 22, 1 (Jan. 2015), 90–94.

[61] S. Bharadwaj, M. Vatsa, and R. Singh. 2013. Can holistic representations be used for face biometric quality assessment? In *Proceedings of the 20th International Conference on Image Processing (ICIP)*. IEEE, 2792–2796.

[62] F. Qu, D. Ren, X. Liu, Z. Jing, and L. Yan. 2012. A face image illumination quality evaluation method based on Gaussian low-pass filter. In *Proceedings of the 2nd International Conference on Cloud Computing and Intelligence Systems (CCIS)*. IEEE, 176–180.

[63] B. F. Klare and A. K. Jain. 2012. Face recognition: Impostor-based measures of uniqueness and quality. In *Proceedings of the 5th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 237–244.

[64] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 74–81.

[65] H. Sellahewa and S. A. Jassim. 2010. Image-quality-based adaptive face recognition. *IEEE Trans. Instrum. Measur.* 59, 4 (Apr. 2010), 805–813.

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

[67] P. Grother, A. Hom, M. Ngan, and K. Hanaoka. 2021. *Ongoing Face Recognition Vendor Test (FRVT) Part 5: Face Image Quality Assessment (4th Draft)*. Technical Report. National Institute of Standards and Technology.

[68] ISO/IEC JTC1 SC37 Biometrics. 2016. *ISO/IEC 29794-1:2016 Information Technology - Biometric Sample Quality - Part 1: Framework*. International Organization for Standardization.

[69] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. 2012. Quality measures in biometric systems. *IEEE Secur. Priv.* 10, 6 (Dec. 2012), 52–62.

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

[71] A. Pena, J. Fierrez, A. Lapedriza, and A. Morales. 2021. Learning emotional-blinded face representations. In *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*.

[72] R. Daza, A. Morales, J. Fierrez, and R. Tolosana. 2020. mEBAL: A multimodal database for eye blink detection and attention level estimation. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*.

[73] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez. 2018. Facial soft biometrics for recognition in the wild: Recent works, annotation, and COTS evaluation. *IEEE Trans. Inf. Forens. Secur.* 13, 8 (Aug. 2018), 2001–2014.

[74] S. Bharadwaj, M. Vatsa, and R. Singh. 2014. Biometric quality: A review of fingerprint, iris, and face. *EURASIP J. Image Video Process.* (July 2014), 28.

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

[76] ISO/IEC JTC1 SC37 Biometrics. 2019. *ISO/IEC 39794-5:2019 Information Technology - Extensible Biometric Data Interchange Formats - Part 5: Face Image Data*. International Organization for Standardization.

[77] H. Proença, M. Nixon, M. Nappi, E. Ghaleb, G. Özbulak, et al. 2018. Trends and controversies. *IEEE Intell. Syst.* 33, 3 (May 2018), 41–67.

[78] G. Zhai and X. Min. 2020. Perceptual image quality assessment: A survey. *Science China Information Sciences* 63, 11 (2020), 211301.

[79] A. Mittal, A. K. Moorthy, and A. C. Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* 21, 12 (Dec. 2012), 4695–4708.

[80] A. Mittal, R. Soundararajan, and A. C. Bovik. 2013. Making a "completely blind" image quality analyzer. *IEEE Sig. Process. Lett.* 20, 3 (Mar. 2013), 209–212.

[81] Venkatanath N, Praneeth D, Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani. 2015. Blind image quality evaluation using perception based features. In *Proceedings of the 21st National Conference on Communications (NCC)*. 1–6.

[82] R. A. Dihin, N. R. Hamza, and Z. H. Toman. 2020. Full-reference facial image quality assessment and identification by two proposed measures. *Journal of Southwest Jiaotong University* 55, 2 (2020), 12.

[83] J. Galbally and S. Marcel. 2014. Face anti-spoofing based on general image quality assessment. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 1173–1178.

[84] ISO/IEC JTC1 SC17 WG3. 2018. *Portrait Quality - Reference Facial Images for MRTD*. Technical Report. International Civil Aviation Organization.

[85] International Civil Aviation Organization. 2021. Retrieved from http://www.icao.int/publications/Documents/9303_p9_cons_en.pdf.

[86] ISO/IEC JTC1 SC37 Biometrics. 2011. *ISO/IEC 19794-5:2011 Information Technology - Biometric Data Interchange Formats - Part 5: Face Image Data.* International Organization for Standardization.

[87] E. Tabassi and P. Grother. 2007. *Quality Summarization - Recommendations on Biometric Quality Summarization Across the Application Domain.* NIST Interagency Report 7422. National Institute of Standards and Technology.

[88] Y. Song, J. Zhang, L. Gong, S. He, L. Bao, et al. 2019. Joint face hallucination and deblurring via structure generation and detail enhancement. *Int. J. Comput. Vis.* 127, 6-7 (June 2019), 785–800.

[89] K. Grm, W. J. Scheirer, and V. Štruc. 2020. Face hallucination using cascaded super-resolution and identity priors. *IEEE Trans. Image Process.* 29 (2020), 2150–2165.

[90] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. 2014. Incremental face alignment in the wild. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 1859–1866.

[91] L. Didaci, G. L. Marcialis, and F. Roli. 2014. Analysis of unsupervised template update in biometric recognition systems. *Pattern Recog. Lett.* 37 (Feb. 2014), 151–160.

[92] H. S. Bhatt, S. Bharadwaj, R. Singh, M. Vatsa, A. Noore, et al. 2011. On co-training online biometric classifiers. In *Proceedings of the International Joint Conference on Biometrics (IJCB).* 1–7.

[93] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez. 2010. Quality-based conditional processing in multi-biometrics: Application to sensor interoperability. *IEEE Trans. Syst., Man, Cyber. - Part A: Syst. Hum.* 40, 6 (Nov. 2010), 1168–1179.

[94] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho. 2018. Multiple classifiers in biometrics. Part 2: Trends and challenges. *Inf. Fusion* 44 (Nov. 2018), 103–112.

[95] M. Singh, R. Singh, and A. Ross. 2019. A comprehensive overview of biometric fusion. *Inf. Fusion* 52 (Dec. 2019), 187–205.

[96] A. Hadid, N. Evans, S. Marcel, and J. Fierrez. 2015. Biometrics systems under spoofing attack: An evaluation methodology and lessons learned. *IEEE Sig. Process. Mag.* 32, 5 (Sept. 2015), 20–30.

[97] B. Fu, N. Spiller, C. Chen, and N. Damer. The effect of face morphing on face image quality. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG).* Gesellschaft für Informatik e.V.

[98] J. Galbally, S. Marcel, and J. Fierrez. 2014. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Trans Image Process.* 23, 2 (2014), 710–724.

[99] P. Drozdowski, C. Rathgeb, and C. Busch. 2019. Computational workload in biometric identification systems: An overview. *IET Biomet.* 8, 6 (Nov. 2019), 351–368.

[100] Z. Zhang, Y. Song, and H. Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[101] CyberExtruder. Ultimate Face Matching Data Set. 2020. Retrieved from https://cyberextruder.com/face-matching-data-set-download/.

[102] S. Curry, D. Founds, J. Marques, N. Orlans, and C. Watson. 2009. *NIST Special Database 32 - Multiple Encounter Dataset I (MEDS-I).* Technical Report NIST IR 7679. National Institute of Standards and Technology.

[103] N. D. Kalka, B. Maze, J. A. Duncan, K. O'Connor, S. Elliott, et al. 2018. IJB-S: IARPA Janus surveillance video benchmark. In *Proceedings of the 9th International Conference on Biometrics: Theory, Applications and Systems (BTAS).* IEEE, 1–9.

[104] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, et al. 2009. ImageNet: A large-scale image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* 248–255.

[105] R. Goh, L. Liu, X. Liu, and T. Chen. 2005. The CMU face in action (FIA) database. In *Proceedings of the 2nd International Workshop on Analysis and Modelling of Faces and Gestures (AMFG).* 255–263.

[106] National Cheng Kung University. NCKU face database. 2020. Retrieved from http://robotics.csie.ncku.edu.tw/Databases/FaceDetect_PoseEstimate.htm.

[107] K. C. Lee, J. Ho, M. H. Yang, and D. Kriegman. 2005. Visual tracking and recognition using probabilistic appearance manifolds. *Comput. Vis. Image Underst.* 99, 3 (Sept. 2005), 303–331.

[108] C. E. Thomaz and G. A. Giraldi. 2010. A new ranking method for principal components analysis and its application to face image analysis. *Image Vis. Comput.* 28, 6 (June 2010), 902–913.

[109] M. Turk and A. Pentland. 1991. Eigenfaces for recognition. *J. Cog. Neurosci.* 3, 1 (1991), 71–86.

[110] M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the International Conference on Computer Vision Workshops (ICCVW).* IEEE, 2144–2151.

[111] D. O. Gorodnichy. 2005. Video-based framework for face recognition in video. In *Proceedings of the 2nd Canadian Conference on Computer and Robot Vision (CRV).* 330–338.

[112] N. Gourier and J. Letessier. 2004. The Pointing'04 data sets. In *Proceedings of the Pointing International Workshop on Visual Observation of Deictic Gestures*.

[113] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. 1999. XM2VTSDB: The extended M2VTS database. In *Proceedings of the 2nd International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)*.

[114] F. Solina, P. Peer, B. Batagelj, S. Juvan, and J. Kovac. 2003. Color-based face detection in the "15 seconds of fame" art installation. In *Proceedings of the International Conference Mirage*.

[115] P. T. Duizer, D. M. Hansen, and T. B. Moeslund. *Data Set: Head Direction*. Technical Report CVMT-07-02. HERMES project (FP6 IST-027110). Retrieved from http://www.cvmt.dk/projects/Hermes/head-data.html.

[116] T. Kanade, J. F. Cohn, and Y. Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition*. IEEE, 46–53.

[117] S. Mandala. 2005. *The Effect of Lighting Direction on Face Recognition Performance*. Master Thesis. West Virginia University.

[118] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report. University of Massachusetts, Amherst.

[119] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 10 (Oct. 2000), 1090–1104.

[120] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. arXiv:1710.08092.

[121] D. Yi, Z. Lei, S. Liao, and S. Z. Li. 2014. Learning face representation from scratch. (Nov. 2014). arXiv:cs/1411.7923.

[122] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, et al. 2008. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans. Syst., Man, Cyber. - Part A: Syst. Hum.* 38, 1 (Jan. 2008), 149–161.

[123] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, et al. 2005. Overview of the face recognition grand challenge. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 947–954.

[124] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. 2016. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the 14th European Conference on Computer Vision*. arXiv:1607.08221.

[125] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, et al. 2016. Frontal to profile face verification in the wild. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[126] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, et al. 2018. IARPA Janus benchmark - C: Face dataset and protocol. In *Proceedings of the 11th IAPR International Conference on Biometrics (ICB)*. IEEE, 158–165.

[127] L. Wolf, T. Hassner, and I. Maoz. 2011. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 529–534.

[128] J. Deng, J. Guo, and S. Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

[129] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, et al. 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1931–1939.

[130] M. Grgic, K. Delac, and S. Grgic. 2011. SCface - Surveillance cameras face database. *Multim. Tools Applic.* 51, 3 (2011), 863–879.

[131] K. C. Lee, J. Ho, and D. J. Kriegman. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 5 (May 2005), 684–698.

[132] T. Zheng and W. Deng. 2018. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Technical Report 18-01.

[133] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, et al. 2017. IARPA Janus benchmark-B face dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[134] E. Eidinger, R. Enbar, and T. Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forens. Secur.* 9, 12 (Dec. 2014), 2170–2179.

[135] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. R. Freire, et al. 2010. The multiscenario multienvironment BioSecure multimodal database (BMDB). *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (2010), 1097–1111.

[136] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, et al. 2011. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. IEEE, 346–353.

[137] F. S. Samaria and A. C. Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of the Workshop on Applications of Computer Vision (ACV)*. IEEE, 138–142.

[138] T. Sim, S. Baker, and M. Bsat. 2003. The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 12 (Dec. 2003), 1615–1618.

[139] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, et al. 2007. *FRVT 2006 and ICE 2006 Large-Scale Results*. Technical Report NIST IR 7408. National Institute of Standards and Technology.

[140] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 6 (June 2001), 643–660.

[141] E. Bailly-Bailliére, S. Bengio, F. E. D. E. R. Bimbot, M. Hamouz, J. Kittler, et al. 2003. The BANCA database and evaluation protocol. In *Audio- and Video-Based Biometric Person Authentication (AVBPA)*. Vol. 2688. Springer Berlin, 625–638.

[142] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, et al. 2017. AgeDB: The first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[143] T. Zheng, W. Deng, and J. Hu. 2017. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. (Aug. 2017). arXiv:cs/1708.08197.

[144] A. P. Founds, N. Orlans, G. Whiddon, and C. Watson. 2011. *NIST Special Database 32 - Multiple Encounter Dataset II (MEDS-II)*. NIST Interagency Report 7807. National Institute of Standards and Technology.

[145] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard. 2016. The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv:1512.00596.

[146] A. Martinez and R. Benavente. 1998. *The AR Face Database.* CVC Technical Report #24 (June 1998).

[147] J. R. Beveridge, K. W. Bowyer, P. J. Flynn, S. Cheng, P. J. Phillips, et al. 2013. The challenge of face recognition from digital point-and-shoot cameras. In *Proceedings of the 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–8.

[148] P. J. Phillips, P. J. Flynn, J. R. Beveridge, W. T. Scruggs, A. J. O'Toole, et al. 2009. Overview of the multiple biometrics grand challenge. In *Proceedings of the 3rd International Conference on Biometrics (ICB) (Lecture Notes in Computer Science)*. Springer, 705–714.

[149] P. A. Johnson, P. Lopez-Meyer, N. Sazonova, F. Hua, and S. Schuckers. 2010. Quality in face and iris research ensemble (Q-FIRE). In *Proceedings of the 4th International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 1–6.

[150] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90.

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

[152] X. Wu, R. He, Z. Sun, and T. Tan. 2015. A Light CNN for Deep Face Representation with Noisy Labels. arXiv:1511.02683.

[153] M. Lin, Q. Chen, and S. Yan. 2013. Network In Network. arXiv:1312.4400.

[154] P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 511–5181.

[155] R. M. Haralick, K. Shanmugam, and I. H. Dinstein. 1973. Textural features for image classification. *IEEE Trans. on Systems, Man, and Cybernetics* SMC-3, 6 (Nov. 1973), 610–621.

[156] F. Weber. 2006. Some quality measures for face images and their relationship to recognition performance. In *Proceedings of the NIST Biometric Quality Workshop*.

[157] G. D. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (2004), 91–110.

[158] E. Murphy-Chutorian and M. M. Trivedi. 2009. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 4 (2009), 20.

[159] International Civil Aviation Organization. 2015. Doc 9303 - Machine Readable Travel Documents.

[160] ISO/IEC JTC1 SC37 Biometrics. 2005. *ISO/IEC 19794-5:2005 Information Technology - Biometric Data Interchange Formats - Part 5: Face Image Data*. International Organization for Standardization.

[161] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley. 2007. Head pose estimation on low resolution images. In *Multimodal Technologies for Perception of Humans*, Vol. 4122. Springer Berlin, 270–280.

[162] P. T. Yap and P. Raveendran. 2004. Image focus measure based on Chebyshev moments. *IEEE Proc. - Vis. Image, Sig. Process.* 151, 2 (2004), 128–136.

[163] P. J. Bex and W. Makous. 2002. Spatial frequency, phase, and the contrast of natural images. *J. Optic. Societ. Amer. A* 19, 6 (June 2002), 1096–1106.

[164] S. Bezryadin, P. Bourov, and D. Ilinih. 2007. Brightness calculation in digital image processing. In *Proceedings of the International Symposium on Technologies for Digital Photo Fulfillment*. 10–15.

[165] Y. Yao, B. R. Abidi, N. D. Kalka, N. A. Schmid, and M. A. Abidi. 2008. Improving long range and high magnification face recognition: Database acquisition, evaluation, and enhancement. *Comput. Vis. Image Underst.* 111, 2 (Aug. 2008), 111–125.

[166] Z. Wang and A. C. Bovik. 2002. A universal image quality index. *IEEE Sig. Process. Lett.* 9, 3 (Mar. 2002), 81–84.

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

[168] NEUROtechnology. Face biometrics. 2020. Retrieved from http://www.neurotechnology.com/face-biometrics.html.

[169] N. Damer, F. Boutros, M. Süßmilch, F. Kirchbuchner, and A. Kuijper. 2021. Extended evaluation of the effect of real and simulated masks on face recognition performance. *IET Biometrics.* 10, 5 (2021), 548–561.

[170] L. Kang, P. Ye, Y. Li, and D. Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* 1733–1740.

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

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

[173] L. Best-Rowden and A. K. Jain. 2018. Learning face image quality from human assessments. *IEEE Trans. Inf. Forens. Secur.* 13, 12 (Dec. 2018), 3064–3077.

[174] A. Oliva and A. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175.

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

[176] K. Simonyan and A. Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556.

[177] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2818–2826.

[178] M. A. Saad, A. C. Bovik, and C. Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans Image Process.* 21, 8 (Aug. 2012), 3339–3352.

[179] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381.

[180] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. 2016. Densely Connected Convolutional Networks. arXiv:1608.06993.

[181] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. 2017. Learning Transferable Architectures for Scalable Image Recognition. arXiv:1707.07012.

[182] F. C. C. O. Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357.

[183] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* 815–823. arXiv:1503.03832.

[184] D. P. Kingma and M. Welling. 2013. Auto-Encoding Variational Bayes. arXiv:1312.6114.

[185] W. Xie and A. Zisserman. 2018. Multicolumn networks for face recognition. In *Proceedings of the British Machine Vision Conference (BMVC).*

[186] H. Zeng, L. Zhang, and A. C. Bovik. 2018. Blind image quality assessment with a probabilistic quality representation. In *Proceedings of the 25th International Conference on Image Processing (ICIP).* IEEE.

[187] P. Grother and E. Tabassi. 2007. Performance of biometric quality measures. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 4 (Apr. 2007), 531–543.

[188] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC WD 29794-1 - Information technology - Biometric sample quality - Part 1: Framework. 2021. Retrieved from https://www.iso.org/standard/79519.html.

[189] M. Olsen, V. Šmida, and C. Busch. 2016. Finger image quality assessment features - Definitions and evaluation. *IET Biomet.* 5, 2 (June 2016), 47–64.

[190] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. 2020. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE, 5202–5211.

[191] F. Hutter, L. Kotthoff, and J. Vanschoren (Eds.). 2018. *Automated Machine Learning: Methods, Systems, Challenges.* Springer. Retrieved from https://automl.org/book.

[192] I. Serna, A. Morales, J. Fierrez, M. Cebrian, N. Obradovich, et al. 2020. Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (SafeAI).*

[193] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch. 2020. Demographic bias in biometrics: A survey on an emerging challenge. *Trans. Technol. Societ.* 1, 2 (June 2020).

[194] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* 58 (June 2020), 82–115.

[195] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana. 2021. SensitiveNets: Learning agnostic representations with application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 6 (June 2021), 2158–2164.

[196] S. Pidhorskyi, D. Adjeroh, and G. Doretto. 2020. Adversarial Latent Autoencoders. arXiv:2004.04467.

[197] J. Galbally, S. Marcel, and J. Fierrez. 2014. Biometric antispoofing methods: A survey in face recognition. *IEEE Access* 2 (Dec. 2014), 1530–1552.

[198] J. Galbally, C. McCool, J. Fierrez, S. Marcel, and J. Ortega-Garcia. 2010. On the vulnerability of face verification systems to hill-climbing attacks. *Pattern Recog.* 43, 3 (Mar. 2010), 1027–1038.

[199] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC WD 29794-5 - Information technology - Biometric sample quality - Part 5: Face image data. 2021. Retrieved from https://www.iso.org/standard/81005.html.

[200] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC WD TS 24358 - Face-aware capture subsystem specifications. 2021. Retrieved from https://www.iso.org/standard/78489.html.

[201] ISO/IEC JTC1 SC37 Biometrics. 2017. *ISO/IEC 29794-4:2017 Information Technology - Biometric Sample Quality - Part 4: Finger Image Data*. International Organization for Standardization.

[202] J. Long and S. Li. 2011. Near infrared face image quality assessment system of video sequences. In *Proceedings of the 6th International Conference on Image and Graphics (ICIG)*. IEEE, 275–279.

[203] Miriam Moreno-Moreno, Julian Fierrez, and Javier Ortega-Garcia. 2009. Biometrics beyond the visible spectrum: Imaging technologies and applications. In *Proceedings of BioID-MultiComm (LNCS)*, Vol. 5707. Springer, 154–161.

[204] W. Y. Lin and M. Y. Chen. 2014. A novel framework for automatic 3D face recognition using quality assessment. *Multimedia Tools and Applications* 68, 3 (Feb. 2014), 877–893.

[205] T. Schlett, C. Rathgeb, and C. Busch. 2021. Deep learning-based single image face depth data enhancement. *Comput. Vis. Image Underst.* 210 (2021).

[206] H. Bi, Z. Liu, L. Yang, K. Wang, and N. Li. 2021. Face sketch synthesis: A survey. *Multim. Tools Applic.* (Feb. 2021).

[207] D. P. Fan, S. Zhang, Y. H. Wu, Y. Liu, M. M. Cheng, et al. 2019. Scoot: A perceptual metric for facial sketches. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 5611–5621.