FROQ[1]: Observing Face Recognition Models for Efficient Quality Assessment

Žiga Babnik[1], Deepak Kumar Jain[2], Peter Peer[1], Vitomir Štruc[1]
[1]University of Ljubljana, Ljubljana, Slovenia
[2]Dalian University of Technology, Dalian, China

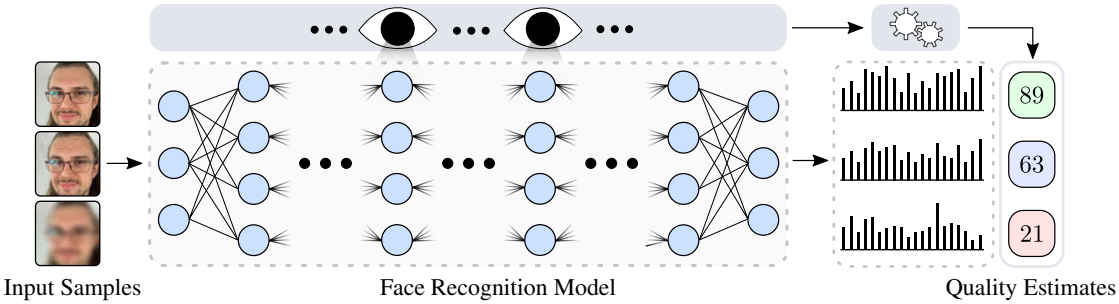

Figure 1. **Illustration of the concept behind the proposed FROQ[1] technique.** Face Recognition (FR) models condense face samples into feature vectors. In the process, they encode identity-specific information, but also other non-identifying cues, such as face-sample quality [4, 25, 40]. Unsupervised Face Image Quality Assessment (FIQA) techniques can extract quality information directly from FR models, but incur a high computational cost. Supervised techniques are efficient, but typically require extensive training with complex loss functions and dedicated (FIQA) model architectures. FROQ combines the best from both worlds and efficiently estimates face-image quality using only the given FR model, by observing a set of specific, carefully chosen intermediate representations, while avoiding costly training and the reliance on custom FIQA-model architectures.

Abstract

*Face Recognition (FR) plays a crucial role in many critical (high-stakes) applications, where errors in the recognition process can lead to serious consequences. Face Image Quality Assessment (FIQA) techniques enhance FR systems by providing quality estimates of face samples, enabling the systems to discard samples that are unsuitable for reliable recognition or lead to low-confidence recognition decisions. Most state-of-the-art FIQA techniques rely on extensive supervised training to achieve accurate quality estimation. In contrast, unsupervised techniques eliminate the need for additional training but tend to be slower and typically exhibit lower performance. In this paper, we introduce FROQ[1] (**F**ace **R**ecognition **O**bserver of **Q**uality), a semi-supervised, training-free approach that leverages specific intermediate representations within a given FR model to estimate face-image quality, and combines the efficiency of supervised FIQA models with the training-free approach of unsupervised methods. A simple calibration step based on pseudo-quality labels allows FROQ to uncover specific representations, useful for quality assessment, in any modern FR model. To generate these pseudo-labels, we propose a novel unsupervised FIQA technique based on sample perturbations. Comprehensive experiments with four state-of-the-art FR models and eight benchmark datasets show that FROQ leads to highly competitive results compared to the state-of-the-art, achieving both strong performance and efficient runtime, without requiring explicit training. The code for FROQ is available from:* https://github.com/LSIbabnikz/FROQ

[1] **FROQ** is pronounced as *FROG*.

1. Introduction

Face Recognition (FR) is an important research area with numerous real-world applications in security and surveillance, border control, police investigations, online banking, and mobile applications, among others [13]. The reliability of FR models in these applications is critical, as errors in the recognition process can compromise user privacy, result in monetary loss, or even lead to legal consequences. While significant advances have been made in FR technology over the years, FR systems still fail to accurately determine identity when deployed in challenging acquisition conditions [9,15,41], where variations in pose, illumination, or other environmental factors cannot be controlled for. To mitigate these issues, FR models often incorporate Face Image Quality Assessment (FIQA) techniques with the goal of assessing the fitness of the input images for recognition.

In accordance with ISO/IEC 29794-1 [21], modern FIQA techniques most often generate a *unified quality score* that corresponds to the utility of the given face sample for the task of recognition. Here, the utility is typically mea-

sured by how likely the sample is to cause false-match errors during the recognition process. In this manner, samples less likely to cause false-match errors are considered to be of higher quality. The quality (or utility) estimates allow FR systems to reject or recapture samples below a certain quality threshold, improving the system's reliability.

Existing FIQA techniques can be broadly categorized into: *unsupervised* and *supervised* methods. Unsupervised methods typically estimate sample quality by looking at the behavior of the FR model to perturbations applied to the input face sample [3–5, 40]. Supervised methods, on the other hand, commonly train a quality-regression model, use pseudo-quality labels [10, 31, 42], rely on a specific loss function [6, 23, 29], or external (often generative) proxy tasks [5, 16, 32]. Unlike supervised methods, unsupervised techniques are easily adapted to any target FR model, but they are noticeably slower when assessing quality, as they require several forward or even additional backward passes through the target FR model. Supervised methods are more efficient during inference, but often rely on dedicated model architectures and, hence, require more work to be adapted for a specific target FR model.

In this paper, we present a novel quality assessment technique, called **FROQ** (**F**ace **R**ecognition **O**bserver of **Q**uality), capable of accurately estimating face-sample quality that needs only a single forward pass through the FR model, as shown in Figure 1. The method can be easily adapted to any FR model and requires no supervised training or additional parameters to tune. The **main contribution** of the approach **is the Quality Observer**, whose goal is to closely monitor specific intermediate representations produced by the FR model during the recognition process. These representations are used as is to estimate the final quality score for a given input face sample. To discover useful intermediate representations, we present a simple semi- (or weakly) supervised approach, which evaluates the usefulness of individual representations for the task of quality estimation through the use of a small quality-labeled calibration set. In this way, FROQ combines the characteristics of both supervised and unsupervised techniques, achieving excellent runtime, estimating the quality within a single forward pass, without the need for any supervised training or additional FR-model parameters.

2. Related Work

In this section, we provide a brief overview of relevant work on face image quality assessment and discuss both *unsupervised* and *supervised* FIQA techniques. For a more comprehensive coverage of the topic, please see [36].

Unsupervised Methods. Unsupervised FIQA methods do not require any supervision when building FIQA models. Instead, they commonly estimate sample quality by observing the effects of various perturbations on the sample's representation within the latent space of the target FR model. One of the earliest methods, SER-FIQ [40], applied dropout to intermediate representations to estimate sample quality. The dropout layer removes certain values from the representation and can be seen as a type of random occlusion on the latent representation. FaceQAN [4] proposed an adversarial attack to predict quality. More specifically, it used the noise applied to the sample during an (adversarial) attack as the perturbation of choice. Recently, DifFIQA [5] introduced a combined approach, using two separate perturbations, encapsulated in the process of modern denoising diffusion probabilistic models (DDPMs). GraFIQs [25] presented a new type of unsupervised FIQA approach, focused on the statistics of the batch normalization layers during the backward pass through the target FR model. A common characteristic of unsupervised FIQA techniques is that they can typically be applied to any modern FR model without additional adjustments. However, they commonly require several forward or even backward passes to estimate quality, making them less computationally efficient.

Supervised Methods. Supervised FIQA techniques commonly require training (auxiliary) quality-regression networks or fine-tuning existing face recognition models to estimate face-sample quality and can be further subdivided by whether they need pseudo-quality labels or not.

Methods requiring pseudo-quality labels employ different annotation techniques to obtain the labeled data. The annotated data is then used to train quality regression models. One of the earliest methods in this category, by Best-Rowden and Jain [7], utilizes labels provided by human annotators, while FaceQNet, proposed by Hernandez-Ortega *et al.* [17, 18], was among the first employing automatically generated pseudo-quality labels, computed by comparing input face samples to the highest-quality images (references) of the same identity. A more robust approach, PCNet [42], used comparisons between several genuine (positive) samples to determine the pseudo labels. SDD-FIQA [31] extended the idea presented by PCNet by also considering information from imposter image pairs. A similar approach was later used in LightQnet [10], with an additional focus on minimizing the parameter count of the final quality extraction model. A quality-label optimization approach, applicable to any set of pseudo-quality labels, was proposed in eDifFIQA [6], leading to highly competitive results. Finally, CLIB-FIQA [32] presented a unique supervised technique that used labels of individual quality factors, such as blur, occlusion, lighting, *etc*., in combination with the CLIP encoder [34] to train a quality estimation model. While techniques from this group usually achieve good results, they need additional training on pseudo-quality labels and are limited by the expressiveness of the label generation process.

On the other hand, supervised FIQA methods that do not require pseudo-quality labels often use a custom loss function to train a model that can estimate both the feature and

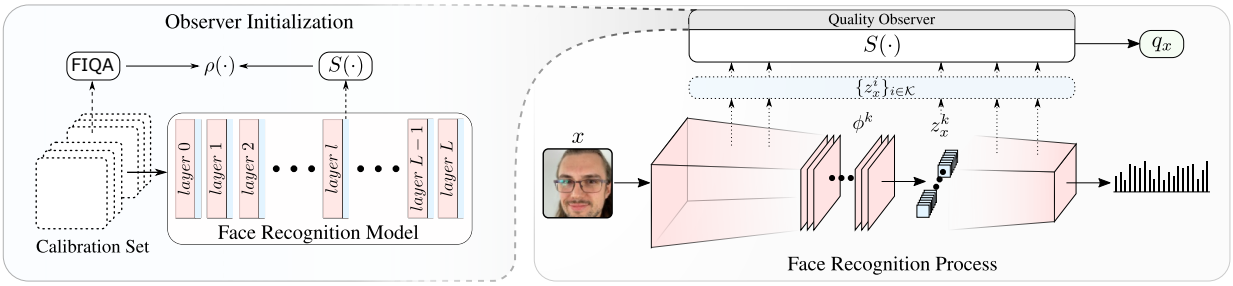

Figure 2. **High-level overview of FROQ.** FROQ estimates sample quality by observing specific intermediate representations produced by the recognition process. The quality score q_x for a given sample x is computed by applying an aggregation function $S(\cdot)$ on the values of the observed intermediate representations $\{z_x^i\}_{i\in\mathcal{K}}$. The set of intermediate representations $\mathcal{K}$ observed during the recognition process is determined in the *Observer Initialization* step. Here, each layer of a given FR model is evaluated using a calibration set of images and their corresponding quality scores obtained through an auxiliary FIQA approach, using Spearman's rank correlation $\rho(\cdot)$.

quality of a given face sample. For this reason, such methods are also called quality-aware FR methods. One of the earliest such methods, PFE [39], learns a mean and variance vector, corresponding to the feature and quality of the given sample. MagFace [29] proposes an extension of the popular ArcFace [12] loss function, with a magnitude-aware term, enabling the model to encode quality information into the magnitude of the feature vector. A special type of such a method, CR-FIQA [8], uses information encoded in the feature space of a trained model to estimate the quality of samples. Specifically, it uses the Certainty-Ratio, which compares the distances of the given sample to the positive class center and the nearest negative class center.

Our Contribution. The proposed FROQ technique leverages pseudo-quality labels; however, unlike supervised methods, the labels are not used to train a quality regressor. Instead, FROQ uses the labels to uncover intermediate representations of the target FR model that are useful for quality assessment. In this process, no new information is encoded into the FR model. For this reason, we categorize FROQ as a semi-supervised technique, combining characteristics of both unsupervised and supervised methods. It can compute the quality scores quickly, using a single forward pass through the FR model, without needing explicit training or limiting the target FR model to a specific loss.

3. Methodology

The main goal of FR models is to extract identity information from any given face sample, in the form of a dense identity-information-rich representation called a feature vector. It has been shown that FR models also encode non-identifying information about the samples, such as face pose, expression [22], as well as the quality of the sample [4, 31, 40]. In this section, we now introduce FROQ, a new semi-supervised FIQA approach that requires no training and can accurately predict sample quality using only a single forward pass through the target FR model. To estimate quality, FROQ uses a *Quality Observer*, which tracks values of predetermined intermediate representations of the given FR model, as shown in Fig. 2. The sample quality is then computed directly from the observed representations using a simple aggregation function. Both the set of observed intermediate representations and the aggregation function are determined in the *Observer Initialization* step.

3.1. Overview

Assume a FR model M, which consists of L layers and is parameterized by $\{\phi^l\}_{l=1}^L$, where ϕ^l are the parameters of the l-th layer. The goal of FROQ is to extract a quality estimate q_x of the input face sample x using the predefined *Quality Observer*. Here, the observer determines the quality q_x using a single forward pass of the sample x through the recognition network M. It is defined by two dependent components: (i) the aggregation function $S(\cdot)$, and (ii) the set of intermediate representations $\mathcal{K}$. The role of the aggregation function is to map any intermediate representation z^j into a single numerical value, where $z^j \in \{z^l\}_{l=1}^L$ is the intermediate representation produced by the j-th layer of the model M. The set $\mathcal{K}$ determines which representations will be observed and subsequently used for quality estimation.

3.2. Observer Initialization

The quality observer is at the core of the proposed FROQ technique. During inference, it accurately determines the quality of any input sample by simply looking at specific intermediate layers. The values of the intermediate representations are first condensed into a single numerical score using a dedicated aggregation function, and then combined across different representations to compute the final quality estimate. In the following section, we describe how the two main components of the observer are determined, i.e., the aggregation function $S(\cdot)$ and the set of representations $\mathcal{K}$.

The Aggregation Function. The main goal of the aggregation function $S(\cdot)$ is to map any intermediate representation z_x^l of the sample x into a single numerical score. There exist infinitely many such functions that make evaluating all possible solutions impossible. Additionally, the choice of the aggregation function also affects the set of observed repre-

sentations, which makes the combined search space of possible aggregation functions and sets of representations far too large to fully explore. For this reason, we hand-craft the aggregation function $S(\cdot)$ based on several insights about FR models, presented by prior works [23, 26, 29, 35].

Modern FR models are trained on medium-to-high quality samples [11, 12, 23], and so the learned parameters (filters) respond well to inputs corresponding to images of higher quality. This means that, generally, the amplitude of the individual layer's outputs (intermediate representations) should be correlated with the quality of the input samples. Following this insight, we design the aggregation function around the norm of intermediate values, formally:

$$S(z_x^l) = \|f_{flatten}(z_x^l)\|_2, \tag{1}$$

where z_x^l is the intermediate representation produced by the l-th layer of the input sample x, $f_{flatten}(\cdot)$ is a flattening operation and $\|\cdot\|_2$ is the L_2 norm. Since the intermediate representation z_x^l can be of arbitrary shape $(d_1, d_2, \ldots, d_D)$, we first reshape it into a single dimension with $(d_1 \times d_2 \times \cdots \times d_D)$ elements, using $f_{flatten}(\cdot)$.

Set of Intermediate Representations $\mathcal{K}$. To determine the set of layers (representations) $\mathcal{K}$, we propose a simple semi-supervised approach, which evaluates the usefulness of individual layers for quality estimation.

Given a calibration set of face images $\{x_i\}_{i=1}^N$, we compute pseudo-quality labels $\{\dot{q}_i\}_{i=1}^N$, using an auxiliary FIQA approach. Here, any preexisting FIQA technique could be used. However, to separate the proposed method from previous works, we devise a custom FIQA approach based on face-sample perturbations, presented in Sec. 3.4. For a given FR model M, we investigate the quality information of each layer $l \in [1, L]$ within M using:

$$c^l = \rho(\{\dot{q}_i\}_{i=1}^N, \{\hat{q}_i^l\}_{i=1}^N), \tag{2}$$

where $\rho(\cdot)$ is Spearman's rank correlation, and $\{\hat{q}_i^l\}_{i=1}^N$ the quality estimates obtained by applying the aggregation function $S(\cdot)$ on the intermediate representations produced by the l-th layer. The computed coefficient c^l shows the similarity of the ranking of sample qualities provided by the auxiliary FIQA approach and the intermediate representation l. A higher correlation coefficient (close to 1) shows that the two analyzed quality vectors have very similar quality rankings, indicating that the analyzed intermediate representations are suitable for the task of quality assessment. By evaluating all layers $l \in [1, L]$ in this manner, we obtain a set of layer correlation coefficients $\{c^l\}_{l=1}^L$.

To determine the final set of intermediate layers $\mathcal{K}$, we take into account only the best b layers, according to the correlation coefficients $\{c^l\}_{l=1}^L$ and collect these layers in $\mathcal{L}^b$. Since the search space of all possible combinations given b elements is $b!$ (factorial), we use an approximate greedy search algorithm limiting the number of combinations to $\frac{(b+1)\cdot b}{2}$. We initialize the set $\mathcal{K}^1$ to contain only the layer whose coefficient c^l is the highest among all layers in $\mathcal{L}^b$. At each subsequent step, the layer l, where $l \in \mathcal{L}^b \wedge l \notin \mathcal{K}^n$, which maximizes the join correlation coefficient $c^{\mathcal{K}^n \cup \{l\}}$, is added to the set $\mathcal{K}^n$. To compute $c^{\mathcal{K}^n \cup \{l\}}$ we use:

$$c^{\mathcal{K}^n \cup \{l\}} = \rho(\{\dot{q}_i\}_{i=1}^N, \{\hat{q}_i^{\mathcal{K}^n \cup \{l\}}\}_{i=1}^N), \tag{3}$$

where $\mathcal{K}^n \cup \{l\}$ is the union of the set $\mathcal{K}^n$ and the single element set $\{l\}$, and $\{\hat{q}_i^{\mathcal{K}^n \cup \{l\}}\}_{i=1}^N$, the set of joint quality scores of the calibration dataset, computed using:

$$\{\hat{q}_i^{\mathcal{K}^n \cup \{l\}}\}_{i=1}^N = \frac{1}{n+1} \sum_{j \in \mathcal{K}^n \cup \{l\}} \{\hat{q}_i^j\}_{i=1}^N, \tag{4}$$

where n is the cardinality of $\mathcal{K}$ in the current step. By performing $b-1$ steps, we exhaust all elements from $\mathcal{L}^b$ and obtain b possible solutions $\{\mathcal{K}^n\}_{n=1}^b$. The $\mathcal{K}^n$, with the highest correlation coefficient $c^{\mathcal{K}^n}$, is selected as the final set of observed representations $\mathcal{K}$.

3.3. Observer Usage

Once the observer has been initialized, i.e., the aggregation function $S(\cdot)$ and the set of representations $\mathcal{K}$ have been determined, for a given FR model M, the process of quality assessment is straightforward. During the recognition process, a sample x is run through the FR model M. The observer then applies the aggregation function $S(\cdot)$ on the values of the observed representations z_x^l, where $l \in \mathcal{K}$, and computes the quality score of the input sample x as:

$$q_x = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} S(z_x^k), \tag{5}$$

where $|\cdot|$ is the cardinality. Thus, the quality score q_x is the average aggregated value of all observed representations.

3.4. Auxiliary FIQA Approach

The proposed FROQ technique relies on a calibration set of face images and the corresponding pseudo-quality labels. To extract the labels, an auxiliary FIQA technique is needed. Here, we present a simple new unsupervised FIQA approach for this task. The auxiliary approach is based on three different types of face sample perturbations, i.e., horizontal flipping, Gaussian noising, and partial occlusions.

Given a face sample x from the calibration set, we compute the pseudo-quality label q_x using:

$$q_x = \frac{1}{3}(q_x^F + q_x^N + q_x^O), \tag{6}$$

where q_x^F, q_x^N, and q_x^O are the partial quality labels obtained using horizontal flips, Gaussian noise, and occlusions. The process to obtain the flip q_x^F and noise q_x^N labels is trivial.

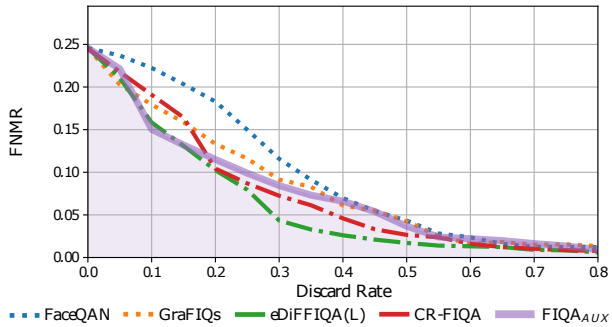

Figure 3. **Performance of the auxiliary FIQA technique.** We compare the performance of the auxiliary FIQA technique against two unsupervised and two supervised state-of-the-art techniques, on the XQLFW benchmark using the AdaFace FR model.

Table 1. **Comparison of benchmarks datasets.** The experiments are performed using eight commonly used benchmarks, which vary in size and focus on different aspects of face image quality.

Dataset	Images	IDs	Comparisons		Main Focus			Scale[†]
			Mated	Non-mated	Pose	Age	Quality	
LFW [19]	13,233	5,749	3,000	3,000	✗	✗	✗	M
Adience [14]	19,370	2,284	20,000	20,000	✗	✗	✗	M
CFP-FP [38]	7,000	500	3,500	3,500	✓	✗	✗	M
CPLFW [44]	11,652	3,930	3,000	3,000	✓	✗	✗	M
CALFW [45]	12,174	4,025	3,000	3,000	✗	✓	✗	M
AgeDB [30]	16,488	570	100,00	100,00	✗	✓	✗	M
XQLFW [24]	13,233	5,749	3,000	3,000	✗	✗	✓	M
IJB-C [28]	23,124[‡]	3,531	19,557	15,638,932	✗	✗	✗	L

[†]M - Medium; L - Large; Values estimated subjectively by the authors.

[‡] number of templates, each containing several images

Using cosine similarity, we compare the features of the original and a horizontally flipped or noisy sample, respectively. We apply noise to the sample according to:

$$x^N = (1-\alpha)\cdot x + \alpha \cdot \mathcal{N}(0,1), \tag{7}$$

where α is a hyperparameter of the approach, and $\mathcal{N}(0,1)$ is a normally distributed random variable, with a mean of zero and variance of one. For the occlusion perturbation, we compute the cosine similarity between the features of the original x and several occluded samples $\{x^{O_i}\}_{i=1}^{R}$. The sample x, which we assume is square (h equals w), is first divided into $R = (h/o)^2$ non-overlapping squares, where h is the height of the image and o the desired size of each square. To construct an occluded sample x^o, the o-th non-overlapping square in the original image is masked by setting all pixels to zero (black). By computing the similarity for all occluded samples and averaging their scores, we obtain the partial quality label q_x^O. The performance of the proposed auxiliary FIQA technique is shown in Fig. 3.

3.5. Calibration Dataset

Discovering informative intermediate representations of a given FR model, as described in Sec. 3.2, requires a small calibration set of face images. To guarantee a fair experimental evaluation, the set of images should not overlap with any of the benchmarks used in the experiments. Therefore, we construct our calibration set from the popular large-scale Glint360K [1,2] dataset. To create our subset, we randomly selected 500 samples from the original dataset. Since the images in the original dataset are mostly high- to medium-quality, we further degrade 33% of the images using the BSRGAN [43] degradation model. This means that the observer initialization step focuses on a wide range of image qualities when constructing the observed set of intermediate representations, ensuring good overall performance of the final quality observer.

Table 2. **Comparison of FIQA techniques.** We analyze ten methods, and compare their requirements pre- and during-inference to the proposed FROQ technique. The unsupervised methods are marked using BLUE, and the supervised using GREEN stripes.

Method	Quality Labels	Architecture Specific	Additional Training	Custom Loss	Inference: Feed-Forward	Inference: Backwards	Inference: Feature Level	Inference: Gradient Level	Inference: Representation Level
SER-FIQ [40]	✗	✓	✗	✗	100	0	✓	✗	✗
FaceQAN [4]	✗	✗	✗	✗	10	10	✓	✓	✗
GraFIQs [25]	✗	✗	✗	✗	1	1	✗	✓	✗
SDD-FIQA [31]	✓	✗	✓	✗	1	0	✓	✗	✗
LightQnet [10]	✓	✗	✓	✓	1	0	✓	✗	✗
PCNet [42]	✓	✗	✓	✗	1	0	✓	✗	✗
eDiFFIQA(L) [6]	✓	✗	✓	✓	1	0	✓	✗	✗
CLIB-FIQA [32]	✓	✓	✓	✓	1	0	✓	✗	✗
MagFace [29]	✗	✗	✓	✓	1	0	✓	✗	✗
CR-FIQA [8]	✗	✗	✓	✓	1	0	✓	✗	✗
FROQ	✓	✗	✗	✗	1	0	✗	✗	✓

4. Experiments & Results

4.1. Experimental Setup

Experimental Setting. We compare the performance of FROQ to 10 state-of-the-art FIQA techniques i.e.: (i) the unsupervised SER-FIQ [40], FaceQAN [4], GraFIQs [25] methods, and (ii) the supervised PCNet [42], SDD-FIQA [31], LightQnet [10], eDifFIQA [6], CLIB-FIQA [32], MagFace [29] and CR-FIQA [8] methods. All techniques are summarized in Table 2. We perform experiments using 4 commonly used state-of-the-art FR models, showing that our method can perform well using any modern FR model. Specifically we use the CNN-based: AdaFace[2] [23], ArcFace[3] [12], and CurricularFace[4] [20] models, as well as the Transformer based SwinFace[5] [33] model. The CNN models all use the ResNet100 backbone, while SwinFace uses the Swin Transformer [27]. All models are trained on the WebFace12M[2], MS1MV3 [3,5], Glint360k [3], and CASIA-WebFace [4] datasets. To show that FROQ can perform well in different scenarios, we repeat the experiments on 8 different face benchmarks, summarized in Table 1, i.e.: (i) LFW [19], Adience [14], (ii) cross-pose CPLFW [44], CFP-FP [38], (iii) cross-age CALFW [45],

[2] https://github.com/mk-minchul/AdaFace
[3] https://github.com/deepinsight/insightface
[4] https://github.com/HuangYG123/CurricularFace
[5] https://github.com/lxq1000/SwinFace

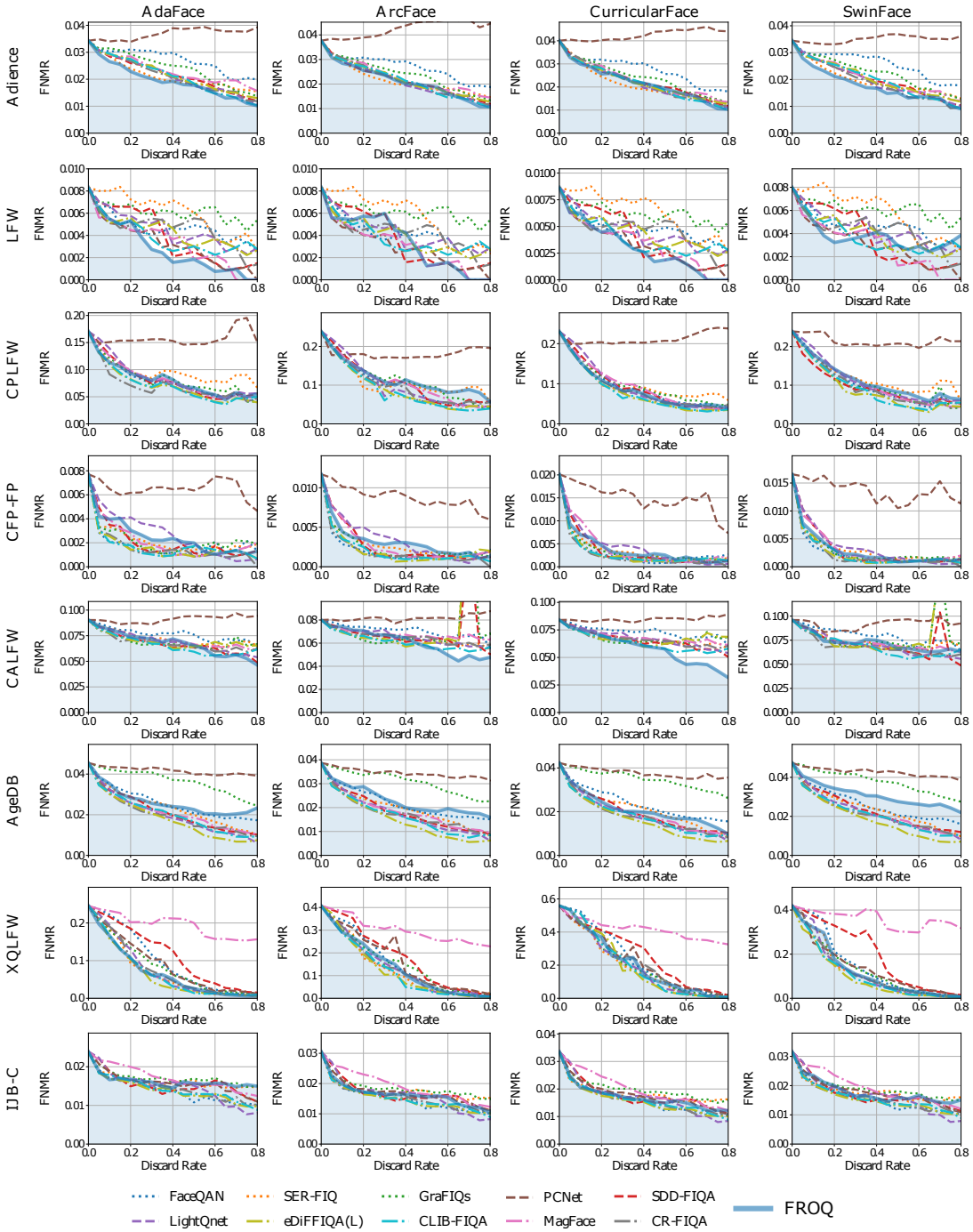

Figure 4. **Comparison to the state-of-the-art using EDC curves.** The performance of FROQ is compared against ten recent FIQA techniques, on eight benchmark datasets, using four FR models. The curves present FNMR values at FMR=$1e^{-3}$, for different discard rates. The unsupervised methods use dotted lines, the supervised methods dashed or dashed-dotted lines, and our method a full line.

AgeDB [30], (iv) cross-quality XQLFW [24] and (v) the large-scale IJB-C [28] benchmark.

Evaluation Methodology. To evaluate the methods, we use EDC (Error-versus Discard Characteristic) curves and the pAUC (partial Area Under the Curve) values calculated using the EDC curves. Both of these are regularly used to evaluate FIQA techniques [6, 8, 29, 36]. The EDC curves measure the FNMR (False Non-Match Rate) at a given FMR (False Match Rate), typically $1e-3$, for different discard rates of low-quality images. Overall, the EDC curves show how the performance of a given FR model improves when discarding some percentage of the lowest quality im-

ages from the dataset. The pAUC values condense the information shown by the EDC curves into a single numerical score by calculating the area under the curves up to some percentage of discarded images, typically set to 20%. For a more concise and clear presentation, we normalize the pAUC values using the value calculated at 0% discard rate.

Implementation Details. In the *Observer Initialization* step, we use the top b individual layers to construct the final set $\mathcal{K}$. To balance the complexity of the search algorithm and the completeness of the final set, we use $b = 10$. The proposed auxiliary FIQA technique uses three separate perturbations to produce the pseudo quality scores $\dot{q}$. For the noisy component q^N, we use the hyperparameter α, while for the occlusion component q^O, we use the hyperparameter o. The parameter α determines the amount of noise added to each sample. For the sake of quality assessment, the amount should be minimal, therefore, we set $\alpha = 0.001$. On the other hand, o determines the size of the square-structured occlusions applied to the sample, measured in pixels. For modern FR models, images are usually resized to 112×112 pixels, and we, therefore, choose $o = 14$, to divide the image into non-overlapping squares. All experiments are conducted on a PC with an Intel i9-10900KF CPU, 64 GB of RAM, and an Nvidia 3090 GPU. Observer initialization for a given FR model takes less than five minutes on the presented PC. For ArcFace, CurricularFace, and SwinFace, the process selects four and for the AdaFace model, five intermediate representation to be observed during recognition.

4.2. Comparison with the State-of-the-Art

In this section, we compare the experimental results of FROQ to other state-of-the-art methods, by looking at their: (i) performance, and (ii) inference runtime.

Performance Analysis. In Fig. 4, we show the comparison with state-of-the-art methods using EDC curves, and in Table 3 the corresponding comparison using the pAUC values. Following [4–6, 32, 37], we use a discard rate of 20% to compute the pAUC values. The results are presented for all four FR models and all included benchmark datasets. Observing the results, we see that the proposed FROQ technique performs well on all benchmark datasets and FR models. Looking at individual results, FROQ outperforms all methods on the Adience benchmark regardless of the given FR model. It achieves excellent (second-best) results on the XQLFW and the LFW benchmarks, using AdaFace, ArcFace, and SwinFace models, respectively. While on the CFP-FP and AgeDB benchmarks, FROQ is competitive compared to the unsupervised methods, and slightly behind the average of the state-of-the-art supervised methods. By observing the $\overline{pAUC}$ results, which combine the results of all benchmark datasets, we see that FROQ outperforms the best unsupervised method, FaceQAN, in all scenarios. The proposed method performs worse than the three best supervised methods: eDifFIQA(L), CR-FIQA, and CLIB-FIQA. Compared to FROQ, all three supervised methods require substantial computational resources and additional parameters to train their respective quality assessment models.

Table 3. **Comparison to the state-of-the-art.** The table shows pAUC scores calculated at a discard rate of 0.2, with FMR set to 10^{-3}, over all benchmarks using all four FR models. We mark the **best** and *second best* result of each dataset. The last column shows average results over all benchmarks, the best average result of the unsupervised methods is colored BLUE and the best of supervised methods is colored GREEN. Unsupervised methods are marked using BLUE, and supervised using GREEN stripes.

AdaFace [23] - $pAUC$(FMR= $1e^{-3}$)[↓]

Methods	Adience	LFW	CPLFW	CFP-FP	CALFW	AgeDB	XQLFW	IJB-C	$\overline{pAUC}$
SER-FIQ [40]	*0.871*	0.982	0.775	0.563	0.930	0.809	0.809	0.812	0.819
FaceQAN [4]	0.919	0.797	0.808	**0.474**	0.945	0.833	0.924	0.800	0.813
GraFIQs [25]	0.932	0.863	0.791	0.557	*0.893*	0.950	0.801	0.846	0.829
PCNet [42]	1.003	**0.730**	0.914	0.893	0.985	0.969	0.826	0.843	0.895
SDD-FIQA [31]	0.884	0.857	0.819	0.632	0.911	0.796	0.907	0.854	0.832
LightQnet [10]	0.890	0.837	0.854	0.711	0.925	0.792	0.835	0.846	0.836
eDiFFIQA(L) [6]	0.889	0.751	*0.736*	0.490	0.894	**0.756**	*0.759*	**0.791**	0.758
CLIB-FIQA [32]	0.893	0.762	0.746	*0.479*	0.897	0.770	0.770	0.807	0.766
MagFace [29]	0.890	*0.735*	0.805	0.632	0.900	*0.760*	0.958	0.915	0.824
CR-FIQA [8]	0.890	0.755	**0.699**	0.504	**0.887**	0.763	0.833	*0.796*	0.766
FROQ$_{ADA}$	**0.843**	0.754	0.764	0.646	0.925	0.822	**0.753**	0.798	0.788

ArcFace [12] - $pAUC$(FMR= $1e^{-3}$)[↓]

Methods	Adience	LFW	CPLFW	CFP-FP	CALFW	AgeDB	XQLFW	IJB-C	$\overline{pAUC}$
SER-FIQ [40]	*0.840*	0.982	0.797	0.539	0.934	0.790	0.828	0.732	0.805
FaceQAN [4]	0.864	0.775	0.826	**0.457**	0.962	0.835	0.883	0.731	0.792
GraFIQs [25]	0.872	0.863	0.814	0.538	*0.902*	0.946	0.818	0.786	0.817
PCNet [42]	1.012	**0.697**	0.810	0.920	0.998	0.965	0.860	0.770	0.879
SDD-FIQA [31]	0.841	0.857	0.829	0.649	0.931	0.801	0.935	0.806	0.831
LightQnet [10]	0.840	0.814	0.862	0.657	0.930	0.797	0.824	0.788	0.814
eDiFFIQA(L) [6]	0.842	0.751	**0.771**	0.497	0.904	**0.757**	0.810	*0.729*	0.758
CLIB-FIQA [32]	0.846	0.762	*0.778*	0.502	**0.900**	*0.762*	**0.789**	0.730	0.759
MagFace [29]	0.852	*0.712*	0.809	0.634	0.925	0.771	0.960	0.867	0.816
CR-FIQA [8]	0.861	0.732	0.791	*0.477*	0.912	0.764	0.814	**0.724**	0.759
FROQ$_{ARC}$	**0.827**	0.746	0.796	0.571	0.937	0.837	*0.805*	0.745	0.783

CurricularFace [20] - $pAUC$(FMR= $1e^{-3}$)[↓]

Methods	Adience	LFW	CPLFW	CFP-FP	CALFW	AgeDB	XQLFW	IJB-C	$\overline{pAUC}$
SER-FIQ [40]	0.832	0.986	0.764	0.493	0.926	0.794	**0.840**	*0.725*	0.795
FaceQAN [4]	0.855	0.786	0.804	0.453	0.953	0.830	0.931	0.730	0.793
GraFIQs [25]	0.857	0.882	0.785	0.477	*0.906*	0.950	0.887	0.780	0.815
PCNet [42]	1.000	**0.732**	0.902	0.931	0.993	0.969	*0.855*	0.776	0.895
SDD-FIQA [31]	0.838	0.865	0.812	0.556	0.932	0.793	0.867	0.806	0.809
LightQnet [10]	*0.827*	0.834	0.852	0.574	0.938	0.783	0.855	0.787	0.806
eDiFFIQA(L) [6]	0.831	0.763	**0.751**	0.448	0.906	**0.740**	0.883	**0.721**	0.755
CLIB-FIQA [32]	0.834	0.774	*0.756*	*0.446*	**0.905**	*0.749*	0.910	0.733	0.763
MagFace [29]	0.841	*0.736*	0.792	0.624	0.921	0.757	0.901	0.875	0.806
CR-FIQA [8]	0.859	0.746	0.765	**0.428**	0.908	0.751	0.901	0.734	0.761
FROQ$_{CURR}$	**0.821**	0.757	0.757	0.531	0.922	0.795	0.873	0.735	0.774

SwinFace [33] - $pAUC$(FMR= $1e^{-3}$)[↓]

Methods	Adience	LFW	CPLFW	CFP-FP	CALFW	AgeDB	XQLFW	IJB-C	$\overline{pAUC}$
SER-FIQ [40]	*0.840*	1.002	0.801	0.496	0.924	0.799	0.824	0.746	0.804
FaceQAN [4]	0.896	0.796	0.820	**0.441**	0.949	0.827	0.934	0.759	0.803
GraFIQs [25]	0.886	0.889	0.819	0.500	*0.889*	0.952	0.791	0.798	0.815
PCNet [42]	0.980	**0.716**	0.996	0.972	0.962	0.963	0.766	0.789	0.893
SDD-FIQA [31]	0.859	0.871	**0.738**	0.604	0.908	0.811	0.922	0.816	0.816
LightQnet [10]	0.867	0.837	0.863	0.603	0.914	0.798	0.815	0.799	0.812
eDiFFIQA(L) [6]	0.856	0.771	*0.784*	0.482	0.897	**0.753**	**0.715**	**0.743**	0.750
CLIB-FIQA [32]	0.861	0.784	0.791	0.494	0.891	0.765	*0.741*	0.749	0.759
MagFace [29]	0.862	0.732	0.819	0.591	0.894	*0.758*	0.960	0.881	0.812
CR-FIQA [8]	0.855	0.753	0.801	*0.466*	**0.863**	0.766	0.788	*0.743*	0.754
FROQ$_{SWIN}$	**0.802**	*0.726*	0.823	0.525	0.899	0.856	0.828	0.795	0.782

Runtime Analysis. In Table 4 we show the comparison with state-of-the-art methods in terms of the inference runtime. The results measure the mean μ and standard deviation σ of the inference runtime, for a single image, computed over a set of 10.000 images. To ensure a fair comparison, each image was processed individually (batch size of 1), the experiments were conducted on the same hardware, and the official implementations provided by the authors were used for all methods. The proposed FROQ method achieves similar runtime performance to other supervised methods, such as CR-FIQA and eDifFIQA(L). Compared to unsupervised methods, FROQ achieves a far

Table 4. **Runtime Complexity of FIQA techniques.** The table shows the method's inference runtime (in ms), for a single image, calculated over a set of 10.000 images. Unsupervised methods are marked using BLUE, and supervised using GREEN stripes.

FIQA Model	SER-FIQ [40]	FaceQAN [4]	GraFIQs [25]	PCNet [42]	SDD-FIQA [31]	LightQnet [10]	eDifFIQA(L) [6]	CLIB-FIQA [32]	MagFace [29]	CR-FIQA [8]	FROQ
Runtime ($\mu \pm \sigma$)	118.376 ± 29.240	$352.12313.515$	55.698 ± 32.328	13.913 ± 5.542	5.060 ± 1.300	4.929 ± 4.615	10.062 ± 1.342	80.346 ± 53.122	8.219 ± 0.228	9.381 ± 0.309	11.025 ± 1.261

Table 5. **Results of the Ablation Study.** We perform several ablation studies, thoroughly investigating the effects of individual components on the final experimental result. The table presents the $pAUC$ values, calculated at a discard rate of 20%, at FMR= $1e^{-3}$ for the ablation experiments using the AdaFace FR model.

	Changes	Adience	LFW	CPLFW	CFP-FP	CALFW	AgeDB	XQLFW	IJB-C	$\overline{pAUC}$
	Baseline	0.843	0.754	0.764	0.646	0.925	0.822	0.753	0.798	**0.788**
FIQA	**eDifFIQA(L)**	0.892	0.789	0.763	0.634	0.920	0.833	0.777	0.812	0.803
FIQA	**CR-FIQA**	0.891	0.754	0.760	0.651	0.924	0.816	0.771	0.806	*0.797*
FIQA	**CLIB-FIQA**	0.892	0.789	0.763	0.634	0.920	0.833	0.777	0.812	0.803
w.o. Opt.	**TOP-1**	0.848	0.801	0.773	0.593	0.926	0.844	0.799	0.837	0.803
w.o. Opt.	**TOP-5**	0.871	0.769	0.773	0.593	0.926	0.844	0.799	0.809	0.798
w.o. Opt.	**TOP-10**	0.873	0.777	0.768	0.593	0.922	0.833	0.785	0.814	*0.796*
	w.o. BSRGAN	0.851	0.931	0.888	0.737	0.966	0.935	0.763	0.814	*0.861*

better runtime, even against the least computationally complex GraFIQs, which is around five times slower. Unsurprisingly, LightQnet, which focuses on having a minimal computational footprint, is the fastest method. Overall, in terms of runtime, FROQ resembles supervised techniques, assessing the quality within a single forward pass, consequently achieving excellent performance.

4.3. Ablation Study

We perform an ablation study to investigate how individual components of the proposed FROQ technique contribute to the final performance, i.e., (i) use of specific auxiliary FIQA techniques, (ii) use of the greedy search algorithm, and (iii) use of the BSRGAN degradation process.

In Table 5, we present the results of the ablation study. The first row marked with *Baseline* presents the results of the FROQ technique; all other rows contain results of the ablation study, separated into the three groups. First, marked with *FIQA*, we present the results of using an alternate auxiliary FIQA technique to produce the pseudo-quality labels. To replace the base auxiliary FIQA, we chose the three best-performing state-of-the-art methods: eDifFIQA(L), CLIB-FIQA, and CR-FIQA. Using CR-FIQA as the auxiliary approach yields the best results, while eDifFIQA(L) and CLIB-FIQA achieve the same averaged result. Surprisingly, all three alternate auxiliary FIQA techniques perform worse than our proposed perturbation-based FIQA approach. Marked with *w.o. Opt.*, we present the results, where we forgo the greedy search for the set of observed intermediate representations $\mathcal{K}$ and instead use the top-n individual representations, specifically the top 1, 5, and 10 layers respectively. We observe that by increasing the number of representations used for the quality assessment task, the results slowly improve, however, they do not reach the performance achieved by the baseline approach. Finally, marked with *w.o. BSRGAN*, we present the results obtained using only high and medium quality images from Glint360k, without any additional degradation from BSRGAN. Here, the performance is significantly worse than that of the baseline, alluding to the importance of a wider range of quality values contained in the calibration set.

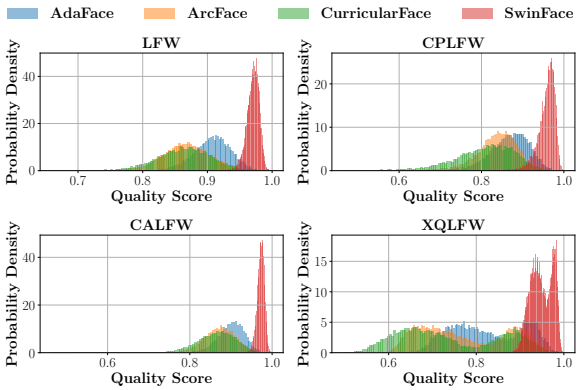

Figure 5. **Results of the Qualitative Evaluation.** We evaluate the quality score distributions of the presented technique, using different FR models over four distinct benchmark datasets.

4.4. Qualitative Evaluation

In this section, we present the results of the qualitative evaluation of the FROQ technique. In particular, we analyze the distribution of normalized quality scores produced by FROQ when using different FR models. We incorporate all four FR models and four distinct benchmarks, i.e., LFW, CPLFW, CALFW, and XQLFW, into the analysis presented in Fig. 5. From the results, differences between FR models can be easily spotted. While AdaFace, ArcFace, and CurricularFace achieve similar distributions, SwinFace exhibits a vastly narrower distribution of quality scores. The disconnect between the models is likely a consequence of the underlying architecture, as the three models are CNN-based, while SwinFace is a Transformer model.

5. Conclusion

In this paper, we introduced FROQ, a semi-supervised face image quality assessment method that estimates sample quality from intermediate representations within a face recognition (FR) model. Using a greedy search, it selects a subset of informative layers for quality estimation. Extensive experiments on multiple datasets have shown that FROQ outperforms all competing unsupervised FIQA methods and performs similarly to the best supervised techniques without requiring specialized training.

Acknowledgments. Supported by ARIS grants P2-0250, P2-0214, J2-2501 and the Young Researcher Program.

References

[1] X. An, J. Deng, J. Guo, Z. Feng, X. Zhu, Y. Jing, and L. Tongliang. Killing Two Birds with One Stone: Efficient and Robust Training of Face Recognition CNNs by Partial FC. In *Proceedings of the CVF/IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu. Partial FC: Training 10 Million Identities on a Single Machine. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV) Workshops*, 2021.

[3] Ž. Babnik, D. Naser, and V. Štruc. Optimization-Based Improvement of Face Image Quality Assessment Techniques. In *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2023.

[4] Ž. Babnik, P. Peer, and V. Štruc. FaceQAN: Face Image Quality Assessment through Adversarial Noise Exploration. In *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*, pages 748–754, 2022.

[5] Ž. Babnik, P. Peer, and V. Štruc. DifFIQA: Face Image Quality Assessment Using Denoising Diffusion Probabilistic Models. In *Proceedings of the IAPR/IEEE International Joint Conference on Biometrics (IJBC)*, pages 1–10, 2023.

[6] Ž. Babnik, P. Peer, and V. Štruc. eDifFIQA: Towards Efficient Face Image Quality Assessment Based on Denoising Diffusion Probabilistic Models. *Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 2024.

[7] L. Best-Rowden and A. K. Jain. Learning Face Image Quality from Human Assessments. *Transactions on Information Forensics and Security (TIFS)*, 13(12):3064–3077, 2018.

[8] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer. CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic Data for Face Recognition: Current State and Future Prospects. *Image and Vision Computing*, 135:104688, 2023.

[10] K. Chen, T. Yi, and Q. Lv. LightQNet: Lightweight Deep Face Quality Assessment for Risk-Controlled Face Recognition. *Signal Processing Letters*, 28:1878–1882, 2021.

[11] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, pages 20642–20653, 2023.

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[13] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei. The Elements of End-to-End Deep Face recognition: A Survey of Recent Advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022.

[14] E. Eidinger, R. Enbar, and T. Hassner. Age and Gender Estimation of Unfiltered Faces. *Transactions on Information Forensics and Security (TIFS)*, 9(12):2170–2179, 2014.

[15] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biometrics*, 7(1):81–89, 2018.

[16] J. Hernandez-Ortega, J. Fierrez, I. Serna, and A. Morales. FaceQgen: Semi-Supervised Deep Learning for Face Image Quality Assessment. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2021.

[17] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, and L. Beslay. Biometric Quality: Review and Application to Face Recognition with FaceQnet. *arXiv preprint arXiv:2006.03298*, 2020.

[18] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay. FaceQnet: Quality Assessment for Face Recognition Based on Deep Learning. In *Proceedings of the IAPR/IEEE International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[19] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[20] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020.

[21] ISO/IEC DIS 29794-1, Biometric Sample Quality. Standard, International Organization for Standardization (ISO), 2022.

[22] J. Jiang and W. Deng. Disentangling Identity and Pose for Facial Expression Recognition. *IEEE Transactions on Affective Computing (TAC)*, 13(4):1868–1878, 2022.

[23] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality Adaptive Margin for Face Recognition. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[24] M. Knoche, S. Hormann, and G. Rigoll. Cross-Quality LFW: A Database for Analyzing Cross-Resolution Image Face Recognition in Unconstrained Environments. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–5, 2021.

[25] J. N. Kolf, N. Damer, and F. Boutros. GraFIQs: Face Image Quality Assessment Using Gradient Magnitudes. In *Proceedings of the CVF/IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1490–1499, 2024.

[26] Q. Li, H. He, H. Lai, T. Cai, Q. Wang, and Q. Gao. Enhanced Nuclear Norm Based Matrix Regression for Occluded Face Recognition. *Pattern Recognition*, 126:108585, 2022.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[28] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. IARPA Janus Benchmark-C: Face Dataset and Proto-

col. In *Proceedings of the IAPR/IEEE International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[29] Q. Meng, S. Zhao, Z. Huang, and F. Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14225–14234, 2021.

[30] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. AgeDB: the First Manually Collected, in-the-Wild Age Database. In *Proceedings of the CVF/IEEE conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 51–59, 2017.

[31] F.-Z. Ou, X. Chen, R. Zhang, Y. Huang, S. Li, J. Li, Y. Li, L. Cao, and Y.-G. Wang. SDD-FIQA: Unsupervised Face Image Quality Assessment with Similarity Distribution Distance. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021.

[32] F.-Z. Ou, C. Li, S. Wang, and S. Kwong. CLIB-FIQA: Face Image Quality Assessment with Confidence Calibration. In *Proceedings of the CVF/IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1694–1704, 2024.

[33] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, and W. Deng. SwinFace: A Multi-Task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 34(4):2223–2234, 2023.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021.

[35] M. S. E. Saadabadi, S. R. Malakshan, A. Zafari, M. Mostofa, and N. M. Nasrabadi. A Quality Aware Sample-to-Sample Comparison for Face Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6129–6138, 2023.

[36] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch. Face Image Quality Assessment: A Literature Survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49, 2022.

[37] T. Schlett, C. Rathgeb, J. Tapia, and C. Busch. Considerations on the Evaluation of Biometric Quality Assessment Algorithms. *Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 6(1):54–67, 2023.

[38] S. Sengupta, J. C. Cheng, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to Profile Face Verification in the Wild. In *Proceedings of the CVF/IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.

[39] Y. Shi and A. K. Jain. Probabilistic Face Embeddings. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, pages 6902–6911, 2019.

[40] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020.

[41] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *Transactions on Technology and Society (TTS)*, 3(1):16–30, 2021.

[42] W. Xie, J. Byrne, and A. Zisserman. Inducing Predictive Uncertainty Estimation for Face Verification. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

[43] K. Zhang, J. Liang, L. Van Gool, and R. Timofte. Designing a Practical Degradation Model for Deep Blind Image Super-Resolution. In *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, pages 4791–4800, 2021.

[44] T. Zheng and W. Deng. Cross-Pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

[45] T. Zheng, W. Deng, and J. Hu. Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments. *CoRR*, abs/1708.08197, 2017.