# CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability

Fadi Boutros[1], Meiling Fang[1,2],Marcel Klemt[1], Biying Fu[1], Naser Damer[1,2]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2]Department of Computer Science, TU Darmstadt, Darmstadt, Germany
Email: fadi.boutros@igd.fraunhofer.de

## Abstract

*Face image quality assessment (FIQA) estimates the utility of the captured image in achieving reliable and accurate recognition performance. This work proposes a novel FIQA method, CR-FIQA, that estimates the face image quality of a sample by learning to predict its relative classifiability. This classifiability is measured based on the allocation of the training sample feature representation in angular space with respect to its class center and the nearest negative class center. We experimentally illustrate the correlation between the face image quality and the sample relative classifiability. As such property is only observable for the training dataset, we propose to learn this property by probing internal network observations during the training process and utilizing it to predict the quality of unseen samples. Through extensive evaluation experiments on eight benchmarks and four face recognition models, we demonstrate the superiority of our proposed CR-FIQA over state-of-the-art (SOTA) FIQA algorithms.* [1]

## 1. Introduction

Face image utility indicates the utility (value) of an image to face recognition (FR) algorithms [1, 19]. This utility is measured with a scalar, namely the face image quality (FIQ) score, following the definition in ISO/IEC 2382-37 [20] and the FR Vendor Test (FRVT) for FIQA [10].

As FIQA measures the face utility to FR algorithm, it does not necessary reflects, and does not aim at measuring, the perceived image quality, e.g. a profile face image can be of high perceived quality but of low utility to FR algorithm [35]. Assessing this perceived image quality has been addressed in the literature by general image quality assessment (IQA) methods [26, 29, 30] and is different than assessing the utility of an the image for FR . This is reflected by FIQA methods [28, 32, 36] significantly outperforming IQA methods [26, 29, 30] in measuring the utility [19] of face images in FR, as demonstrated in [8, 28, 36].

[1] https://github.com/fdbtrs/CR-FIQA

SOTA FIQA methods focused either on creating concepts to label the training data with FIQ scores and then learn a regression problem [14, 15, 32], or on developing a link between face embedding properties under certain scenarios and the FIQ [28, 34, 36]. Generally, the second approach led to better FIQA performances with most works mentioning the error-prone labeling of the ground truth quality in the first research direction as a possible reason [28, 36]. However, in the second category, transferring the information in network embeddings into an FIQ score is not a learnable process, but rather a form of statistical analysis, which might not be optimal.

This paper proposes a novel learning paradigm to assess FIQ, namely the CR-FIQA. Our concept is based on learning to predict the classifiability of FR training samples by probing internal network observations that point to the relative proximity of these samples to their class centers and negative class centers. This regression is learned simultaneously with a conventional FR training process that minimizes the distance between the training samples and their class centers. Linking the properties that cause high/low classifiability of a training sample to the properties leading to high/low FIQ, we can use our CR-FIQA to predict the FIQ of any given sample. We empirically prove the theorized link between classifiability (Section 3.3) and FIQ and conduct thorough ablation studies on key aspects of our CR-FIQA design (Section 5). The proposed CR-FIQA is evaluated on eight benchmarks along with SOTA FIQAs. The reported results on four FR models demonstrate the superiority of our proposed CR-FIQA over SOTA methods and the stability of its performance across different FR models. An overview of the proposed CR-FIQA is presented in Figure 1 and will be clarified in detail in this paper.

## 2. Related work

The recent SOTA FIQA approaches can be roughly grouped into two main categories. The first are approaches that learn a straight forward regressions problem to assess a FIQ score [1, 14, 15, 32, 38]. The second category uses prop-

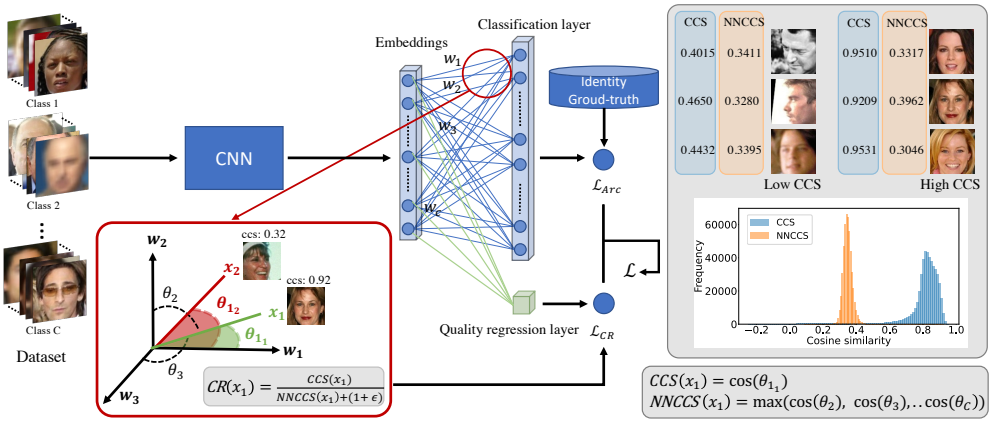



Figure 1. An overview of our CR-FIQA training paradigm. We propose to simultaneously learn to optimize the class center ($\mathcal{L}_{Arc}$), while learning to predict an internal network observation i.e. the allocation of the feature representation of sample $x$ in feature space, with respect to its class center $w_1$ and nearest negative class center $w_2$ ($\mathcal{L}_{\mathcal{CR}}$). The figure in the red rectangle illustrates the angle between two samples $x_1$ and $x_2$ (belong to identity 1) and their class center $w_1$. The plot on the right of the figure shows the distribution of the cosine similarity between training samples and their class centers (CCS) and nearest negative class centers (NNCCS) obtained from ResNet-50 trained on CASIA-WebFace [39]. The example images on the top-right of this plot are of high CCS values and the ones on the top-left are of low CCS values (notice the correspondence to perceived quality). These samples are selected from CASIA-WebFace [39]. During the testing mode, the classification layer is removed and the output of the regression layer is used to predict the FIQA of testing samples.

erties of the FR model responses to face samples to estimate the sample quality without explicitly learning a typical supervised regression that requires quality labels [28, 34, 36]. In the first category, the innovation focused on creating the FIQ labels for training. These quality labels included human-labeled quality labels [1], the FR genuine comparison score between a sample and an ICAO [18] compliant sample [14, 15], the FR comparison score involving the labeled sample (assumed to have the lower quality in the comparison pair) [38], and the Wasserstein distance between a randomly selected genuine and imposter FR comparisons with the labeled sample [32]. These solutions generally trained a regression network to predict the quality label, using both, trained-from-scratch networks in some cases [38], and pre-trained FR networks in other cases [14, 15, 32]. A slightly different approach, however also based on learning from labels, focuses on learning to predict the sample quality as a rank [22] based on FR performance-based training rank labels of a set of databases [4]. In the second category, the innovation was rather focused on linking face embedding properties under certain scenarios to the FIQ, without the explicit need for quality-labeled data. In [36], the assessed sample is passed through an FR network multiple times, each with a different random dropout pattern. The robustness of the resulting embeddings, represented by the sigmoid of the negative mean of the Euclidean distances between the embeddings, is considered the FIQ score. In [28], the FIQ score is calculated as the magnitude of the sample embedding. This is based on training the FR model using a loss that adapts the penalty margin loss based on this magnitude, and thus links the closeness of a sample to its class center to the unnormalized embedding magnitude. While in [34], the solution produces both, an FR embedding and a gaussian variance (uncertainty) vector, from a face sample. The inverse of harmonic mean of the uncertainty vector is considered as the FIQ score. Our CR-FIQA learns a regression problem to estimate the FIQ score, however, unlike previous works, without relying on preset labels, but rather learn a dynamic internal network observations (during training) that point out sample classifiability.

## 3. Approach

This section presents our proposed Certainty Ratio Face Image Quality Assessment (CR-FIQA) approach, which inspects internal network observations to learn to predict the sample relative classifiability. This classifiability prediction is then used to estimate the FIQ. An overview of the proposed CR-FIQA approach is presented in Figure 1. During the training phase of an FR model, the model can conveniently push the high-quality samples close to their class center and relatively far from other class centers. Conversely, the FR is not able to push, to the same degree, low-quality samples to their class center, and thus they will remain relatively farther from their class center than the high-quality ones. Based on this assumption, we theorize our approach by stating that the properties that cause a face sample to lay relatively closer to its class center during training are the ones that make it a high-quality sample, and vice versa. Therefore, learning to predict such properties in any given sample would lead to learning to assess this sample quality. To learn to perform such assessment, our training paradigm targets learning internal network observations that evolve during the FR training phase, where these observa-

tions act as a training objective. The predictions of such training paradigm can be simply stated as answering a question: if a given sample was hypothetically part of the FR model training (which it is not), how relatively close would it be to its class center? Answering this question would give us an indication of this sample quality as will be shown in detail in this paper.

In the rest of this section, We formalize and empirically rationalize our proposed CR-FIQA approach and its components. To do that, we start by shortly revisiting angular margin penalty-based softmax loss utilized to optimize the class centers of the FR model. Then present a detailed description of our proposed CR-FIQA concept and the associated training paradigm.

### 3.1. Revisiting Margin Penalty-based Softmax Loss

Angular margin penalty-based softmax is a widely used loss function for training FR models [3,5,17,28]. It extends over softmax loss by deploying angular penalty margin on the angle between the deep features and their corresponding weights. Margin penalty-based softmax loss aims to push the decision boundary of softmax, and thus enhance intra-class compactness and inter-class discrepancy. From this family of loss functions, this work utilizes ArcFace loss [5] to optimize the distance between the training samples and their class center. Our choice of ArcFace loss is based on the SOTA performance achieved by ResNet-100 network trained with ArcFace on mainstream benchmarks [5]. Formally, ArcFace loss is defined as follows:

$$\mathcal{L}_{Arc} = \frac{1}{N}\sum_{i\in N} -log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum\limits_{j=1,j\neq y_i}^{C} e^{s(cos(\theta_j))}}, \tag{1}$$

where $N$ is the batch size, $C$ is the number of classes (identities), $y_i$ is the class label of sample $i$ (in range $[1, C]$), $\theta_{y_i}$ is the angle between the features $x_i$ and the $y_i$-th class center $w_{y_i}$. $x_i \in R^d$ is the deep feature embedding of the last fully connected layer of size $d$. $w_{y_i}$ is the $y_i$-th column of weights $W \in R^d_C$ of the classification layer. $\theta_{y_i}$ is defined as $x_i w^T_{y_i} = \|x_i\|\|w_{y_i}\|cos(\theta_{y_i})$ [24]. The weights and the feature norms are fixed to $\|w_{y_i}\| = 1$ and $\|x_i\| = 1$, respectively, using $l_2$ normalization as defined in [24, 37]. The decision boundary, in this case, depends on the angle cosine between $x_i$ and $w_{y_i}$. $m > 0$ is an additive angular margin proposed by ArcFace [5] to enhance the intra-class compactness and inter-class discrepancy. Lastly, $s$ is the scaling parameter [37].

### 3.2. Certainty Ratio

In this section, we formulate and empirically rationalize the main concepts that build our FIQA solution. We derive our Certainty Ratio (CR) to estimate the sample relative classifiability. Additionally, we experimentally illustrate the strong relationship between our CR measure and FIQ.

**Certainty Ratio** During the FR model training phase, the model is trained to enhance the separability between the classes (identities) by pushing each sample $x_i$ to be close to its class center $w_{y_i}$ and far from the other (negative) class centers $w_j, j \neq y_i$. Based on this, we first define the Class Center Angular Similarity (CCS) as the proximity between $x_i$ and its class center $w_{y_i}$, as follows:

$$CCS_{x_i} = cos(\theta_{y_i}), \tag{2}$$

where $\theta_{y_i}$ is the angle between $x_i$ and its class center $w_{y_i}$, where the weights of the last fully connected layer of the FR model trained with softmax loss are considered as the centers for each class [5,25]. Then, we define the Closest Nearest Negative Class Center Angular Similarity (NNCCS) as proximity between $x_i$ and the nearest negative class center $w_j, j \neq y_i$. Formally, NNCCS is defined as follows:

$$NNCCS_{x_i} = \max_{j=1,j\neq y_i}^{C} (cos(\theta_j)), \tag{3}$$

where $\theta_j$ is the angle between $x_i$ and $w_j$. As we theorize, when the FR model converges, the high-quality samples are pushed closer to their class centers (high CCS) in relation to their distance to neighbouring negative class centers (low NNCCS). However, low-quality samples can not be pushed as close to their class centers. A sample able to achieve high CCS with respect to NNCCS is a sample easily correctly classified during training, and thus is relatively highly classifiable. We thus measure this relative classifiability by the ratio of CCS to NNCS, which we note as the Certainty Ratio (CR), as follows:

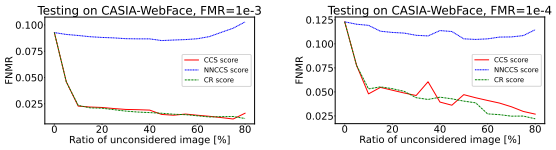
$$CR_{x_i} = \frac{CCS_{x_i}}{NNCCS_{x_i} + (1+\epsilon)}, \tag{4}$$

where the $1 + \epsilon$ term is added to insure a positive above zero denominator, i.e. shift the NNCCS value range from $[-1, +1]$ to $[\epsilon, 2 + \epsilon]$. This ensures that the CR of a sample with a lower NNCCS is relatively higher than a sample with a higher NNCCS, given the same CCS, i.e. NNCCS regulates the CCS value in relation to neighbouring classes. The $\epsilon$ is set to $1e-9$ in our experiments. The optimal CR is obtained when the CCS is approaching the maximum cosine similarity value (+1) and the NNCCS is approaching the minimum cosine similarity value (-1), i.e. the training sample is capable of being pushed to its class center, and far away from the closest negative class center, and thus it is highly classifiable.

### 3.3. Relation between the CR and FIQ

Here, we empirically prove the theorized relationship between the CR and FIQ (defined earlier as image utility). Namely, we want to answer: if the CR values achieved by training samples of an FR model were used as FIQ, would they behave as expected from an optimal FIQ? If yes, then the face image properties leading to high/low CR

do also theoretically lead to high/low FIQ. To answer this question we conducted an experiment on a ResNet-50 [13] FR model trained on CASIA-WebFace [39] with ArcFace loss [5] (noted as R50(CASIA)). Specifically, we calculate the CR, CCS, and NNCCS values from the trained model for all samples in the training dataset (0.5M images of 10K identities). An insight on the resulting CCS and NNCCS values (CR being a derivative measure) is given as value distributions in Figure 1, showing that these measures vary between different samples. Furthermore, based on the calculated scores, we plot Error vs. Reject Curves (ERC) (described in Section 4) to demonstrate the relationship between the CR, as an FIQ measure, and FR performance. To calculate the FR performance in the ERC curve, we extract the feature embedding of CASIA-WebFace [39] using a ResNet-100 model [13] trained on MS1M-V2 [5, 12] with ArcFace (noted as R100(MS1M-V2)). We utilize a different model (trained on a different database) to extract the embedding (R100(MS1M-V2)) than the one used to calculate CR, CCS, and NNCCS (R50(CASIA)) to provide fair evaluation where the FR performance is evaluated on unseen data. Then, we perform $n : n$ comparisons between all samples of CASIA-WebFace using feature embedding obtained from the R100(MS1M-V2).




(a) (b)

Figure 2. ERCs showing the verification performance as False None Match Rate (FNMR) at False Match Rate (FMR) of 1e-3 (2a) and 1e-4 (2b) with CCS, NNCCS and CR as FIQ vs. rejection ratio. This ERC plots show the effectiveness of rejecting samples with the lowest CCS and CR on the performance

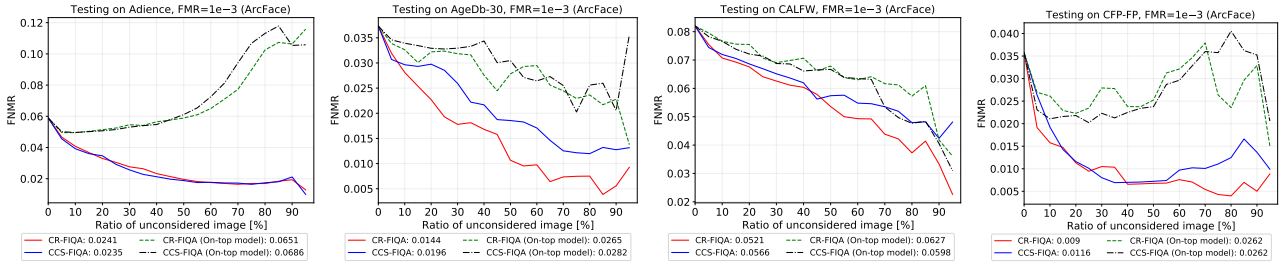
Figure 2 presents the ERC of CR, CCS, and NNCCS experimentally used as FIQ. An FIQ measure would cause the ERC to drop as rapidly as possible when rejecting a larger fraction of low-quality samples (moving to the right).

It can be clearly noticed in Figure 2 that the CCS and CR do behave as we would expect from a good performing FIQ, as the verification error value drops rapidly when rejecting low quality (low CCS and CR) samples. It can be also observed that the CR does that more steadily when compared to CCS. This points out that adding the scaling term NNCCS in CR calculation can enhance the representation of the CCS as an FIQ measure, which will be clearer later when we experimentally evaluate our CR-FIQA approach in Section 5. As expected, the NNCCS measure by itself does not strongly act as an FIQ measure would, demonstrated by the relatively flat ERC in Figure 2, as it only considers the distance to the nearest negative class. This empirical evaluation does provide a confirming answer to the previously stated question by affirming that the CR does act as expected from an FIQ measure and thus, theoretically, one can strongly link the image properties that cause high/low CR in the FR training data to these causing high/low FIQ.

### 3.4. Quality Estimation Training Paradigm

In the previous section, we proved that the CR does behave as an FIQ would, and thus, it can also relate to image properties that dictate FIQ. However, the CR measure is only observable for samples in the FR training dataset, where the class centers are known. In a real case scenario, the FIQ measure should be assessed to any single image, i.e. unseen evaluation data. Considering this, and in an effort to predict what the CR value would be for a given sample if hypothetically it was part of the FR training, we propose to simultaneously learn to predict the CR from the training dataset while optimizing the class centers (typical FR training) during the training phase, i.e. the CR-FIQA model. To enable this, we add a single regression layer to the FR model. The input of the regression layer is a feature embedding $x_i$ and the output is an estimation of the CR. The output of this regression layer is used later to predict the FIQ score of the unseen sample, e.g. from the evaluation dataset. Thus, we capture the properties that make the CR high/low to predict the FIQ of any given sample. Towards this goal, during the training phase, the model (in Figure 1) has two learning objectives: a) It is trained to optimize the distance between the samples and the class centers using ArcFace loss defined in Equation 1. b) It is trained to predict the internal network observation, CR, using Smooth L1-Loss [9] applied between the output of the regression layer (P) and the CR calculated as in Equation 5. Smooth L1-loss can be interpreted as a combination of L1 and L2-losses by defining a threshold $\beta$ that changes between them [9]. Our choice for smooth-l1 loss is based on: 1) It is less sensitive to outliers than l2. The derivative of L2 loss increases when the difference between the prediction and ground-truth label is increased, making the derivative of loss values large at the early stage of the training, leading to unstable training. Additionally, L2 loss can easily generate gradient explosion [9] when there are outliers in the training data. 2) L1 loss can lead to stable training. However, the absolute values of the difference between prediction and ground truth are small, especially in the later stage of the training. Therefore, the model accuracy can hardly be improved at a later stage of the training as the loss function will fluctuate around a stable value. Combining L1 and L2 as in Smooth L1-loss avoids gradient explosion, which might be caused by L2 and facilitates better convergence than L1. The loss leading to the second objective is then given as:

$$\mathcal{L}_{CR} = \frac{1}{N} \sum_{i \in N} \begin{cases} \frac{0.5 \times (CR_{x_i} - P_i)^2}{\beta} & \text{if } |CR_{x_i} - P_i| < \beta \\ |CR_{x_i} - P_i| - 0.5 \times \beta & \text{otherwise} \end{cases}. \quad (5)$$




(a) (b) (c) (d)

Figure 3. ERC comparison between CR-FIQA(S), CCS-FIQA(S), CR-FIQA(S) (On top) and CCS-FIQA(S) (On top). The plots show the effect of rejecting samples of lowest quality, on the verification error (FNMR at FMR1e-3). CR-FIQA(S) and CCS-FIQA(S) outperformed the on-top solutions, and CR-FIQA(S) performs generally better than CCS-FIQA(S) (curve decays faster with more rejected samples)

The final loss combining both objectives for training our CR-FIQA model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{Arc} + \lambda \times \mathcal{L}_{CR}, \tag{6}$$

where $\lambda$ is a hyper-parameter used to control the balance between the two losses. At the beginning of model training, the value range of $\mathcal{L}_{CR}$ is very small ($\leq 2$) in comparison to $\mathcal{L}_{Arc}$ ($\sim 45$). Setting $\lambda$ to a small value, the model will only focus on $\mathcal{L}_{Arc}$. Besides, setting $\lambda$ to a large value, i.e. $> 10$, we observed that the model did not converge. Therefore, we set $\lambda$ to 10 in all the experiments in this paper.

## 4. Experimental Setup

**Implementation Details** We demonstrate our proposed CR-FIQA under two protocols (small and large) based on the training dataset and the training model architecture. We utilize widely used architectures in the SOTA FR solutions, ResNet100 and ResNet50 [13], both modified as described in Section 3.4. For the small protocol, we utilize ResNet50 and the CASIA-WebFace [39] training data (noted as CR-FIQA(S)) and for the large protocol, we utilize ResNet100 and the MS1MV2 [5, 12] training data (noted as CR-FIQA(L)). The MS1MV2 is a refined version of the MS-Celeb-1M [12] by [5] containing 5.8M images of 85K identities. The CASIA-WebFace contains 0.5m images of 10K identities [39]. We follow the ArcFace training setting [5] to set the scale parameter $s$ to 64 and the margin $m$ to 0.5. We set the mini-batch size to 512. All models are trained with Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 1e-1. During the training, we use random horizontal flipping with a probability of 0.5 for data augmentation. We set the momentum to 0.9 and the weight decay to 5e-4. For CR-FIQA(S), the learning rate is divided by 10 at 20K and at 28K training iterations, following [5]. The training is stopped after 32K iterations. For CR-FIQA(L), the learning rate is divided by 10 at 100K and 160K training iterations, following [5]. The training is stopped after 180K iterations. All the images in evaluation and training datasets are aligned and cropped to $112 \times 112$, as described in [5]. All the training and testing images are normalized to have pixel values between -1 and 1. Both models are trained using the loss defined in Equation 6.

**Evaluation Benchmarks** We reported the achieved results on eight different benchmarks: Labeled Faces in the Wild (LFW) [16], AgeDB-30 [31], Celebrities in Frontal-Profile in the Wild (CFP-FP) [33], Cross-age LFW (CALFW) [41], Adience [6], Cross-Pose LFW (CPLFW) [40], Cross-Quality LFW (XQLFW) [23], IARPA Janus Benchmark–C (IJB-C) [27]. These benchmarks are chosen to provide a wide comparison to SOTA FIQA algorithms and give an insight into the CR-FIQA generalizability.

**Evaluation Metric** We evaluate the FIQA by plotting ERCs [11]. The ERC is a widely used representation of the FIQA performance [10, 11] by demonstrating the effect of rejecting a fraction face images, of the lowest quality, on face verification performance in terms of False None Match Rate [21] (FNMR) at a specific threshold calculated at fixed False Match Rate [21] (FMR). The ERC curves for all benchmarks are plotted at two fixed FMRs, 1e-3 (as recommended for border control operations by Frontex [7]) and 1e-4 (the latter is provided in the supplementary material). We also report the Area under the Curve (AUC) of the ERC, to provide a quantitative aggregate measure of verification performance across all rejection ratios.

Additionally, motivated by evaluating the FIQ as a weighting term for face embedding [32, 34], we follow the IJB-C 1:1 mixed verification benchmark [27] by weighting the frames such that all frames belonging to the same subject within a video have a combined weight equal to a single still image as described in IJB-C benchmark [27]. We do that by using the CR-FIQA quality scores as well as all SOTA methods. We report the verification performance of IJB-C as true acceptance rates (TAR) at false acceptance rates (FAR) of 1e-4, 1e-5, and 1e-6, as defined in [27].

**Face Recognition Models** We utilize four different SOTA FR models to report the verification performance at different quality rejection rate to inspect the generalizability of FIQA over FR solutions. The FR models are ArcFace [5], ElasticFace (ElasticFace-Arc) [3], MagFace [28], and CurricularFace [17]. All models process $112 \times 112$ aligned and cropped image to produce 512-D feature em-

bedding. We used the officially released pretrained ResNet-100 models trained on MS1MV2 released by the four FR solutions. Although, the presented solution in this paper does not aim, and is not presented as, a solution to extract face embeddings, but rather an FIQA solution, we opted to evaluate CR-FIQA(L) backbone as a FR model on mainstream FR benchmarks for sake of providing complete experiment evaluation and probe the possibility of simultaneously using it as both FIQA and FR model. The evaluation results of CR-FIQA(L) backbone as a FR model are provided in the supplementary material.

**Baseline** We compare our CR-FIQA approach with nine quality assessment methods. Three are general IQA methods that have been proven in [8] to correlate well to face utility i.e. BRISQUE [29], RankIQA [26], and DeepIQA [2], and six are SOTA face-specific FIQA methods, namely RankIQ [4], PFE [34], SER-FIQ [36], FaceQnet (v1 [14]) [14, 15], MagFace [28], and SDD-FIQA [32], all as officially released in the respective works.

## 5. Ablation Studies

This section provides experimental proof of the two main design choices in CR-FIQA.

**Does CR-FIQA benefit from the NNCCS scaling term?** To answer this, we conducted additional experiments using ResNet-50 model trained on CASIA-WebFace [39] using the experimental setup described in Section 4. This model is noted as CCS-FIAQ(S). The only difference from CR-FIAQ(S) is that the CCS-FIAQ(S) is trained to learn CCS (instead of CR) by replacing the $CR_{x_i}$ in Equation 5 with $CCS_{x_i}$, thus neglecting the NNCCS scaling term in the equation. Figure 3 presents the ERCs along with AUC using CR-FIAQ(S) and CCS-FIAQ(S) on Adience, AgeDb-30, CALLFW and CRF-FP. The verification error, FNMR at FMR1e-3, is calculated using ArcFace FR model (described in Section 4). The ERCs and AUC values show that the reduction in the error is more evident for CR-FIAQ(S) than CCS-FIQA(S). Thus, adding the scaling term NNCCS in CR enhanced the performance of the FIQA.

**Dose the simultaneous learning in CR-FIQA lead to better performance in comparison to on-the-top learning?** We consider the possibility of learning to estimate CR after finalizing the FR training, in comparison to the simultaneously learning in CR-FIQA. We conducted two additional experiments using pretrained ResNet-50 trained with ArcFace loss [5] on CASIA-WebFace [39]. Specifically, we add an additional single regression layer to this pretrained model. We freeze the weights of the pretrained model and train only the regression layer to learn the internal network observation of the pretrained model using only the $\mathcal{L}_{\mathcal{CR}}$ (Equation 5). Using this setting, we present two instances, CR-FIQA(S) (On top) and CCS-FIQA(S) (On top), that learned to predict CR and CCS, respectively. Each of CR-FIQA(S) (On top) and CCS-FIQA(S) (On top) is fine-tuned for 32K iteration with an initial learning rate of 0.01. The learning rate is divided by 10 at 20K and 28K training iterations, similarly to CR-FIAQ(S) and CCS-FIAQ(S). The results (ERCs and AUCs) of these models are compared to CR-FIQA(S) and CCS-FIQA(S) in Figure 3. The ERCs in Figure 3 presented the evaluation on Adience, AgeDb-30, CALFW and CFP-FP using the ArcFace FR model. The ERCs and AUCs in Figure 3 show that both CR-FIQA(S) and CCS-FIQA(S) lead to stronger reductions in the error than the CR-FIQA(S) (On top) and the CCS-FIQA(S) (On top), when rejecting low-quality samples. This supports our training paradigm that simultaneously learns the internal network observation, CR, while optimizing the class centers. This can be related to the step-wise convergence towards the final CR value in the simultaneously training. For both ablation study questions, ERCs and AUC for all the remaining benchmarks and FR models (mentioned in Section 4) lead to similar conclusions and are provided in the supplementary material.

## 6. Result and Discussion

All CR-FIQA performances reported in this paper are obtained under cross-model settings. The proposed CR-FIQA is used only to predict FIQ and not to extract feature representation of face images. None of the utilized FR models (ArcFace [5], ElasticFace [3], MagFace [28], and CurricularFace [17]) is trained with our paradigm. Instead, we trained a separate model for CR-FIQA and used the official pretrained FR models (as described in Section 4) for feature extraction. The verification performances as AUC at FMR1e-3 and FMR1e-4 are presented in Table 1. The visual verification performances as ERC curves (Figure 4) using ArcFace and ElasticFace FR models are reported at FMR1e-3. The ERC curves at FMR1e-4 and for MagFace and CurricularFace FR models at FMR1e-3 are provided in the supplementary material.

The ERC curves (Figure 4) and the AUC values (Table 1) show that our proposed CR-FIQA(S) and CR-FIQA(L) outperformed the SOTA methods by significant margins in almost all settings. Observing the results on IJB-C, Adience, CFP-FP, CALFW, and CPLFW at FMR1e-3 and FMR1e-4 (Figure 4 and Table 1), our proposed CR-FIQA outperformed all SOTA methods on all the considered FR models. On the AgeDB-30 benchmark, our proposed CR-FIQA ranked first in five out of eight settings and second in the other three settings (Table 1). On the LFW benchmark, our proposed CR-FIQA ranked behind the MagFace and the RankIQ. This is the only case that our models did not outperform all SOTA methods. However, it can be noticed from the ERC curves in Figure 4 that none of the SOTA methods were able to achieve stable behavior (smoothly decaying curve) on LFW. The main reason for such unstable ERC behavior on LFW is that the FR performance on LFW is nearly saturated (all models achieved above 99.80% accu-

Table 1. The AUCs of ERC achieved by our CR-FIQA and the SOTA methods under different experimental settings. CR-FIQA achieved the best performance (lowest AUC) in almost all settings. On XQLFW, the SER-FIQ (marked with *) is used for the sample selection of the XQLFW benchmark. The best result for each experimental setting is in bold and the second-ranked one is in italic. The notions of $1e-3$ and $1e-4$ indicate the value of the fixed FMR at which the ERC curve (FNMR vs. reject) was calculated at

| FR | | Method | Adience [6] | | AgeDB-30 [31] | | CFP-FP [33] | | LFW [16] | | CALFW [41] | | CPLFW [40] | | XQLFW [23] | | IJB-C [27] | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $1e-3$ | $1e-4$ | $1e-3$ | $1e-4$ | $1e-3$ | $1e-4$ | $1e-3$ | $1e-4$ | $1e-3$ | $1e-4$ | $1e^{-3}$ | $1e-4$ | $1e-3$ | $1e-4$ | $1e-3$ | $1e-4$ |
| ArcFace [5] | IQA | BRISQUE [29] | 0.0565 | 0.1285 | 0.0400 | 0.0585 | 0.0343 | 0.0433 | 0.0043 | 0.0049 | 0.0755 | 0.0813 | 0.2558 | 0.3037 | 0.6680 | 0.7122 | 0.0381 | 0.0656 |
| | | RankIQA [26] | 0.0400 | 0.0933 | 0.0372 | 0.0523 | 0.0301 | 0.0384 | 0.0039 | 0.0045 | 0.0846 | 0.0915 | 0.2437 | 0.2969 | 0.6584 | 0.7039 | 0.0385 | 0.0640 |
| | | DeepIQA [2] | 0.0568 | 0.1372 | 0.0403 | 0.0523 | 0.0238 | 0.0292 | 0.0049 | 0.0056 | 0.0793 | 0.0850 | 0.2309 | 0.2856 | 0.5958 | 0.6458 | 0.0383 | 0.0640 |
| | FIQA | RankIQ [4] | 0.0353 | 0.0873 | 0.0322 | 0.0420 | 0.0152 | 0.0260 | *0.0018* | *0.0024* | 0.0608 | 0.0672 | 0.0633 | 0.0848 | 0.2789 | 0.3332 | 0.0227 | 0.0342 |
| | | PFE [34] | 0.0212 | 0.0428 | 0.0172 | 0.0226 | 0.0092 | 0.0129 | 0.0023 | 0.0028 | 0.0647 | 0.0681 | 0.0450 | 0.0638 | 0.2302 | 0.2710 | 0.0176 | 0.0248 |
| | | SER-FIQ [36] | 0.0223 | 0.0434 | 0.0167 | 0.0223 | *0.0065* | *0.0103* | 0.0023 | 0.0028 | 0.0595 | 0.0627 | *0.0389* | 0.0584 | **0.1812*** | **0.2295*** | *0.0161* | *0.0241* |
| | | FaceQnet [14, 15] | 0.0346 | 0.0734 | 0.0197 | 0.0245 | 0.0240 | 0.0273 | 0.0022 | 0.0027 | 0.0774 | 0.0822 | 0.1504 | 0.1751 | 0.5829 | 0.6136 | 0.0270 | 0.0376 |
| | | MagFace [28] | *0.0207* | *0.0425* | *0.0156* | 0.0198 | 0.0073 | 0.0105 | **0.0016** | **0.0021** | *0.0568* | *0.0602* | 0.0492 | 0.0642 | 0.4022 | 0.4636 | 0.0171 | 0.0254 |
| | | SDD-FIQA [32] | 0.0248 | 0.0562 | 0.0186 | 0.0206 | 0.0122 | 0.0193 | 0.0021 | 0.0027 | 0.0641 | 0.0698 | 0.0517 | 0.0670 | 0.3090 | 0.3561 | 0.0186 | 0.0270 |
| | | CR-FIQA(S)(Our) | 0.0241 | 0.0517 | **0.0144** | **0.0187** | 0.0090 | 0.0145 | 0.0020 | 0.0025 | **0.0521** | **0.0554** | 0.0391 | *0.0567* | 0.2377 | 0.2740 | 0.0171 | 0.0250 |
| | | CR-FIQA(L)(Our) | **0.0204** | **0.0353** | 0.0159 | *0.0189* | **0.0050** | **0.0082** | 0.0023 | 0.0029 | 0.0616 | 0.0632 | **0.0360** | **0.0515** | *0.2084* | *0.2441* | **0.0138** | **0.0207** |
| ElasticFace [3] | IQA | BRISQUE [29] | 0.0644 | 0.1184 | 0.0375 | 0.0403 | 0.0281 | 0.0372 | 0.0034 | 0.0047 | 0.0726 | 0.0747 | 0.2641 | 0.4688 | 0.6343 | 0.6964 | 0.0357 | 0.0621 |
| | | RankIQA [26] | 0.0433 | 0.0862 | 0.0374 | 0.0436 | 0.0269 | 0.0318 | 0.0033 | 0.0045 | 0.0810 | 0.0835 | 0.2325 | 0.4306 | 0.6189 | 0.6856 | 0.0366 | 0.0599 |
| | | DeepIQA [2] | 0.0645 | 0.1203 | 0.0384 | 0.0411 | 0.0191 | 0.0256 | 0.0043 | 0.0056 | 0.0756 | 0.0772 | 0.2401 | 0.4541 | 0.5400 | 0.5832 | 0.0379 | 0.0590 |
| | FIQA | RankIQ [4] | 0.0400 | 0.0777 | 0.0309 | 0.0337 | 0.0149 | 0.0180 | **0.0013** | **0.0020** | 0.0598 | 0.0614 | 0.0581 | 0.0727 | 0.2468 | 0.2776 | 0.0226 | 0.0334 |
| | | PFE [34] | *0.0222* | *0.0381* | 0.0163 | 0.0172 | 0.0088 | 0.0113 | 0.0018 | 0.0025 | 0.0628 | 0.0643 | 0.0419 | 0.0895 | 0.2112 | 0.2436 | 0.0171 | 0.0247 |
| | | SER-FIQ [36] | 0.0240 | 0.0417 | 0.0163 | 0.0179 | *0.0061* | *0.0085* | 0.0021 | 0.0028 | 0.0574 | 0.0590 | 0.0387 | 0.0513 | **0.1576*** | **0.1868*** | *0.0156* | *0.0235* |
| | | FaceQnet [14, 15] | 0.0369 | 0.0667 | 0.0194 | 0.0207 | 0.0227 | 0.0247 | 0.0021 | 0.0026 | 0.0763 | 0.0777 | 0.1420 | 0.2880 | 0.5549 | 0.5844 | 0.0263 | 0.0370 |
| | | MagFace [28] | 0.0225 | 0.0385 | 0.0150 | **0.0158** | 0.0069 | 0.0095 | *0.0014* | *0.0021* | *0.0553* | *0.0563* | 0.0474 | 0.0597 | 0.3973 | 0.4282 | 0.0166 | 0.0243 |
| | | SDD-FIQA [32] | 0.0277 | 0.0512 | 0.0187 | 0.0200 | 0.0098 | 0.0118 | 0.0019 | 0.0027 | 0.0624 | 0.0638 | 0.0493 | 0.0634 | 0.3052 | 0.3562 | 0.0183 | 0.0266 |
| | | CR-FIQA(S)(Our) | 0.0257 | 0.0465 | **0.0146** | 0.0160 | 0.0070 | 0.0096 | 0.0015 | 0.0022 | **0.0509** | **0.0522** | *0.0383* | *0.0502* | 0.2093 | 0.2835 | 0.0167 | 0.0244 |
| | | CR-FIQA(L)(Our) | **0.0214** | **0.0357** | *0.0149* | *0.0159* | **0.0045** | **0.0065** | 0.0018 | 0.0025 | 0.0594 | 0.0608 | **0.0350** | **0.0462** | *0.1798* | *0.2060* | **0.0135** | **0.0203** |
| MagFace [28] | IQA | BRISQUE [29] | 0.0594 | 0.1308 | 0.0442 | 0.0799 | 0.0422 | 0.0589 | 0.0043 | 0.0058 | 0.0758 | 0.0788 | 0.4649 | 0.6809 | 0.6911 | 0.7229 | 0.0462 | 0.0787 |
| | | RankIQA [26] | 0.0407 | 0.0889 | 0.0370 | 0.0681 | 0.0369 | 0.0543 | 0.0041 | 0.0056 | 0.0829 | 0.0857 | 0.3251 | 0.6475 | 0.6706 | 0.7046 | 0.0461 | 0.0750 |
| | | DeepIQA [2] | 0.0571 | 0.1302 | 0.0417 | 0.0721 | 0.0322 | 0.0545 | 0.0048 | 0.0059 | 0.0787 | 0.0809 | 0.3672 | 0.6632 | 0.6162 | 0.6519 | 0.0474 | 0.0765 |
| | FIQA | RankIQ [4] | 0.0359 | 0.0837 | 0.0361 | 0.0531 | 0.0213 | 0.0332 | *0.0019* | *0.0027* | 0.0602 | 0.0629 | 0.0659 | 0.1642 | 0.3076 | 0.3475 | 0.0269 | 0.0383 |
| | | PFE [34] | 0.0215 | 0.0423 | 0.0192 | 0.0317 | 0.0107 | 0.0138 | 0.0023 | 0.0029 | 0.0640 | 0.0652 | 0.0449 | 0.1435 | 0.2615 | 0.2926 | 0.0200 | 0.0283 |
| | | SER-FIQ [36] | 0.0233 | 0.0451 | 0.0185 | 0.0293 | *0.0080* | 0.0139 | 0.0025 | 0.0033 | 0.0590 | 0.0607 | *0.0397* | *0.0821* | **0.2139*** | **0.2562*** | *0.0189* | *0.0270* |
| | | FaceQnet [14, 15] | 0.0365 | 0.0720 | 0.0217 | 0.0314 | 0.0271 | 0.0351 | 0.0022 | *0.0027* | 0.0763 | 0.0773 | 0.2988 | 0.5218 | 0.6016 | 0.6210 | 0.0305 | 0.0422 |
| | | MagFace [28] | *0.0212* | *0.0417* | **0.0159** | 0.0247 | 0.0085 | 0.0129 | **0.0017** | **0.0022** | *0.0562* | *0.0578* | 0.0506 | 0.0887 | 0.4478 | 0.4900 | 0.0195 | 0.0279 |
| | | SDD-FIQA [32] | 0.0253 | 0.0562 | 0.0216 | 0.0305 | 0.0146 | 0.0201 | 0.0021 | *0.0027* | 0.0643 | 0.0657 | 0.0525 | 0.1188 | 0.3404 | 0.3928 | 0.0215 | 0.0307 |
| | | CR-FIQA(S)(Our) | 0.0244 | 0.0507 | *0.0165* | **0.0234** | 0.0102 | *0.0121* | 0.0020 | 0.0028 | **0.0516** | **0.0528** | 0.0409 | 0.0840 | 0.2670 | 0.3336 | 0.0198 | 0.0284 |
| | | CR-FIQA(L)(Our) | **0.0211** | **0.0372** | 0.0174 | *0.0235* | **0.0062** | **0.0080** | 0.0023 | 0.0028 | 0.0614 | 0.0628 | **0.0374** | **0.0679** | *0.2369* | *0.2839* | **0.0162** | **0.0236** |
| CurricularFace [17] | IQA | BRISQUE [29] | 0.0502 | 0.1095 | 0.0433 | 0.0491 | 0.0323 | 0.0357 | 0.0041 | 0.0047 | 0.0755 | 0.0784 | 0.2709 | 0.5057 | 0.6146 | 0.6336 | 0.0362 | 0.0589 |
| | | RankIQA [26] | 0.0359 | 0.0752 | 0.0394 | 0.0510 | 0.0298 | 0.0356 | 0.0039 | 0.0045 | 0.0806 | 0.0865 | 0.2346 | 0.4654 | 0.5900 | 0.6212 | 0.0361 | 0.0556 |
| | | DeepIQA [2] | 0.0492 | 0.1070 | 0.0407 | 0.0476 | 0.0227 | 0.0278 | 0.0050 | 0.0056 | 0.0764 | 0.0786 | 0.2488 | 0.4961 | 0.5165 | 0.5526 | 0.0376 | 0.0571 |
| | FIQA | RankIQ [4] | 0.0314 | 0.0715 | 0.0365 | 0.0417 | 0.0186 | 0.0249 | *0.0018* | *0.0024* | 0.0590 | 0.0640 | 0.0541 | 0.0730 | 0.2449 | 0.2880 | 0.0220 | 0.0320 |
| | | PFE [34] | **0.0198** | 0.0365 | 0.0197 | 0.0227 | 0.0100 | 0.0134 | 0.0024 | 0.0028 | 0.0630 | 0.0657 | 0.0402 | 0.0983 | 0.1982 | 0.2220 | 0.0170 | 0.0238 |
| | | SER-FIQ [36] | 0.0211 | 0.0381 | 0.0167 | **0.0193** | *0.0074* | *0.0111* | 0.0025 | 0.0030 | 0.0587 | 0.0610 | 0.0356 | 0.0520 | **0.1558*** | **0.1866*** | *0.0153* | *0.0228* |
| | | FaceQNet [14, 15] | 0.0326 | 0.0626 | 0.0221 | 0.0267 | 0.0226 | 0.0274 | 0.0022 | 0.0027 | 0.0767 | 0.0799 | 0.1384 | 0.3229 | 0.5035 | 0.5411 | 0.0259 | 0.0354 |
| | | MagFace [28] | *0.0200* | *0.0364* | 0.0167 | 0.0195 | 0.0078 | 0.0111 | **0.0016** | **0.0021** | *0.0563* | *0.0590* | 0.0449 | 0.0607 | 0.3758 | 0.4178 | 0.0163 | 0.0232 |
| | | SDD-FIQA [32] | 0.0230 | 0.0462 | 0.0219 | 0.0254 | 0.0138 | 0.0185 | 0.0021 | 0.0027 | 0.0637 | 0.0675 | 0.0465 | 0.0671 | 0.2649 | 0.3053 | 0.0178 | 0.0254 |
| | | CR-FIQA(S)(Our) | 0.0227 | 0.0446 | **0.0156** | *0.0198* | 0.0097 | 0.0148 | 0.0020 | 0.0025 | **0.0513** | **0.0534** | *0.0340* | *0.0501* | 0.2101 | 0.2470 | 0.0165 | 0.0234 |
| | | CR-FIQA(L)(Our) | **0.0198** | **0.0336** | *0.0162* | 0.0200 | **0.0054** | **0.0080** | 0.0023 | 0.0029 | 0.0605 | 0.0618 | **0.0324** | **0.0462** | *0.1716* | *0.2318* | **0.0134** | **0.0194** |

Table 2. Verification performance on the IJB-C (1:1 mixed verification) [27]. CR-FIQA outperformed SOTA methods under all settings

| Quality Estimation | | 1:1 mixed Verification: TAR (%) at | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ArcFace [5] | | | ElasticFace [3] | | | MagFace [28] | | | CurricularFace [17] | | |
| | | FAR=1e-6 | FAR=1e-5 | FAR=1e-4 | FAR=1e-6 | FAR=1e-5 | FAR=1e-4 | FAR=1e-6 | FAR=1e-5 | FAR=1e-4 | FAR=1e-6 | FAR=1e-5 | FAR=1e-4 |
| | - | *89.85* | 94.47 | 96.28 | 89.15 | 94.54 | 96.49 | 85.67 | 93.08 | 96.65 | 90.46 | 94.89 | 96.58 |
| IQA | BRISQUE [29] | 86.65 | 93.62 | 95.98 | 85.68 | 93.51 | 95.65 | 81.11 | 90.64 | 94.82 | 88.16 | 93.98 | 96.29 |
| IQA | RankIQA [26] | 86.37 | 93.61 | 95.83 | 86.71 | 93.46 | 96.00 | 80.78 | 90.75 | 94.86 | 88.16 | 94.11 | 96.22 |
| IQA | DeepIQA [2] | 81.97 | 91.64 | 94.67 | 78.93 | 91.59 | 94.81 | 73.53 | 86.34 | 92.90 | 82.65 | 92.04 | 95.00 |
| FIQA | RankIQ [4] | 88.78 | 94.42 | 96.20 | 88.88 | 94.64 | 96.45 | 85.63 | 92.66 | 95.70 | 90.00 | 94.93 | 96.53 |
| FIQA | PFE [34] | 89.50 | 94.51 | 96.31 | 89.10 | 94.67 | 96.51 | 84.93 | 92.44 | 95.60 | 90.36 | 95.04 | 96.54 |
| FIQA | SER-FIQ [36] | 89.74 | 94.65 | 96.32 | *90.05* | 94.79 | 96.57 | 86.02 | 93.35 | 95.80 | 90.66 | 95.11 | 96.58 |
| FIQA | FaceQNet [14, 15] | 87.87 | 94.04 | 96.12 | 86.26 | 94.09 | 96.25 | 82.91 | 90.56 | 95.03 | 89.61 | 94.65 | 96.36 |
| FIQA | MagFace [28] | 89.49 | 94.41 | 96.22 | 89.37 | 94.69 | 96.46 | 85.75 | 92.71 | 95.54 | 90.34 | 95.02 | 96.50 |
| FIQA | SDD-FIQA [32] | 89.39 | 94.61 | 96.34 | 88.07 | 94.82 | 96.49 | 84.69 | 92.83 | 95.73 | 89.91 | *95.12* | *96.63* |
| FIQA | CR-FIQA(S)(Our) | 89.59 | **94.78** | *96.35* | **90.30** | **94.97** | **96.63** | *86.45* | *93.48* | **95.95** | **90.82** | **95.13** | **96.64** |
| FIQA | CR-FIQA(L)(Our) | **90.16** | *94.75* | **96.36** | 90.00 | *94.92* | *96.58* | **87.12** | **93.67** | *95.90* | *90.79* | *95.12* | 96.58 |

racy [3, 5, 17, 28]), leaving very few samples causing errors and thus lowering the significance of the measured FNMR.

The XQLFW benchmark is derived from LFW to contain pairs with a maximum difference in quality. The XQLFW images are chosen based on BRISQUE [29] and SER-FIQ [23] quality scores to be either extremely high or low. The use of SER-FIQ [36] in this selection might give a biased edge for SER-FIQ on this benchmark. On XQLFW, our CR-FIQA achieved very close performance to the selection method (SER-FIQ) and is far ahead of all other SOTA methods. Lastly, our proposed CR-FIQA(S) achieved very comparable performance to our CR-FIQA(L), pointing out the robustness of our approach, regardless of the training database and architecture complexity.

Table 2 presents the verification performance on the IJB-C 1:1 mixed verification benchmark [27] using quality scores as an embedding weighting term (as defined in [27]) under different experimental settings. For each of the FR models, we report in the first row the evaluation result of the corresponding FR model as defined in the protocol [27] and

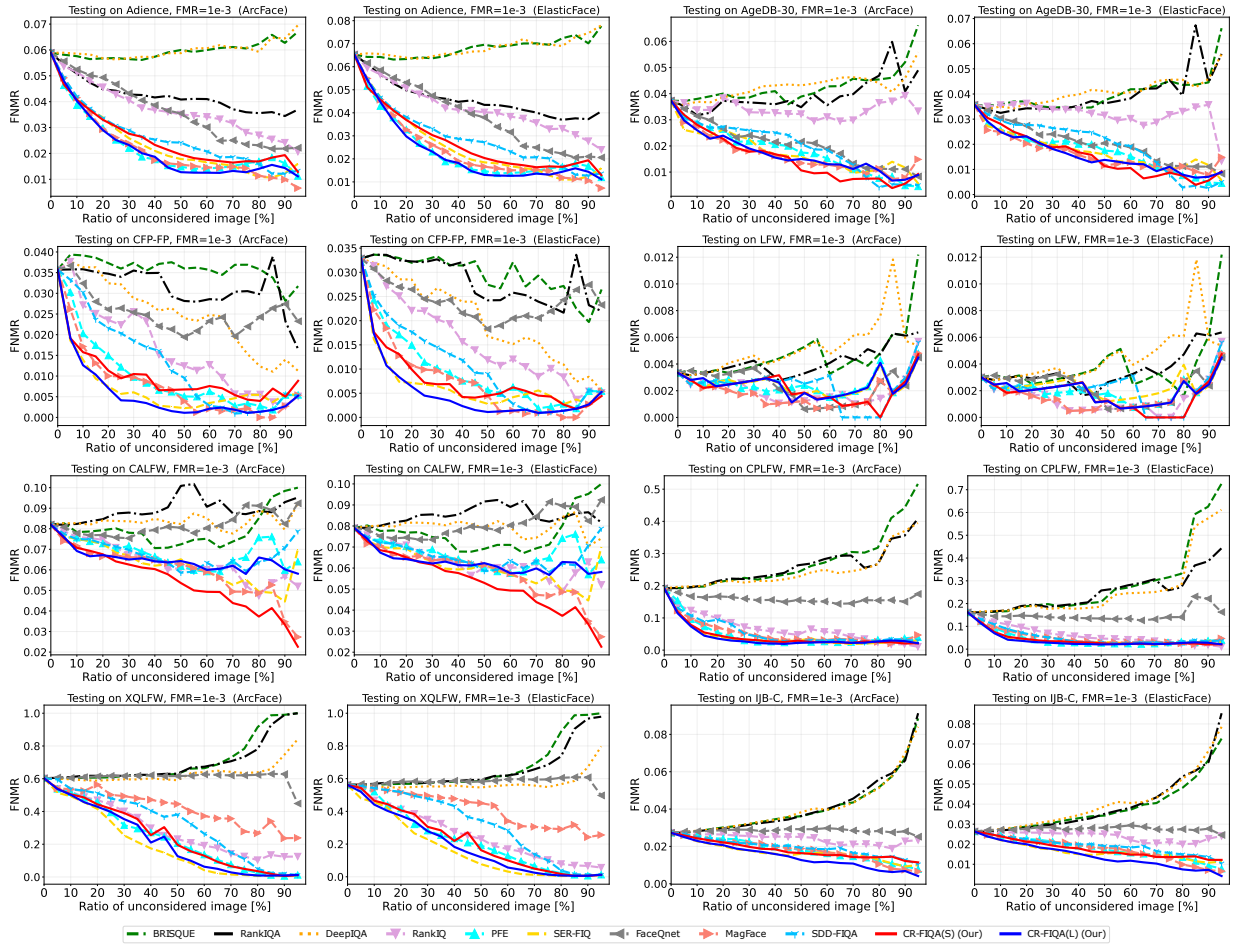



Figure 4. ERC (FNMR at FMR1e-3 vs reject) curves for all evaluated benchmarks using ArcFace and ElasticFace FR models corresponding to Table 1 results. The visual evaluation, ERC curves, using MagFace and CurricularFace FR models are provided in the supplementary material. The proposed CR-FIQA(L) and CR-FIQA(S) are marked with solid blue and red lines, respectively. CR-FIQA leads to lower verification error, when rejecting a fraction of images, of the lowest quality, in comparison to SOTA methods (faster decaying curve) under most experimental settings.

the corresponding released evaluation scripts [3, 5, 17, 28], i.e. without considering the FIQ. Our proposed CR-FIQA significantly leads to higher verification performance than all evaluated SOTA methods, when the quality score is used as an embedding weighting term (Table 2). This achievement is observable under all experimental settings (Table 2). Another outcome of this evaluation is that the integration of CR-FIQA leads to SOTA verification performance on one of the most challenging FR benchmarks, IJB-C [27].

## 7. Conclusion

In this work, we propose the CR-FIQA approach that probes the relative classifiability of training samples of the FR model and utilize this observation to learn to predict the utility of any given sample in achieving an accurate FR performance. We experimentally prove the theorized relationship between the sample relative classifiability and FIQ and build on that towards our CR-FIQA. The CR-FIQA training paradigm simultaneously learns to optimize the class center while learning to predict sample relative classifiability. The presented ablation studies and the extensive experimental results prove the effectiveness of the proposed CR-FIQA approach, and its design choices, as an FIQ method. The reported results demonstrated that our proposed CR-FIQA outperformed SOTA methods repeatedly across multiple FR models and on multiple benchmarks, including ones with a large age gap (AgeDb-30, Adience, CALFW), large quality difference (XQLFW), large pose variation (CPLFW, CFP-FP), and extremely large-scale and challenging FR benchmarks (IJB-C).

**Acknowledgment** This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. This work has been partially funded by the German Federal Ministry of Education and Research (BMBF) through the Software Campus Project.

## References


[1] Lacey Best-Rowden and Anil K. Jain. Learning face image quality from human assessments. *IEEE Trans. Inf. Forensics Secur.*, 13(12):3064–3077, 2018. 1, 2

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1):206–219, 2018. 6, 7

[3] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022. 3, 5, 6, 7, 8

[4] Jiansheng Chen, Yu Deng, Gaocheng Bai, and Guangda Su. Face image quality assessment based on learning to rank. *IEEE Signal Process. Lett.*, 22(1):90–94, 2015. 2, 6, 7

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 3, 4, 5, 6, 7, 8

[6] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.*, 9(12):2170–2179, 2014. 5, 7

[7] Frontex. Best practice technical guidelines for automated border control (abc) systems, 2015. 5

[8] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. A deep insight into measuring face image utility with general and face-specific image quality metrics. pages 1121–1130, 2022. 1, 6

[9] Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society, 2015. 4

[10] P. Grother, M. Ngan A. Hom, and K. Hanaoka. Ongoing face recognition vendor test (frvt) part 5: Face image quality assessment (4th draft). In *National Institute of Standards and Technology*. Tech. Rep., Sep. 2021. 1, 5

[11] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, Apr. 2007. 5

[12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016. 4, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4, 5

[14] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *CoRR*, abs/2006.03298, 2020. 1, 2, 6, 7

[15] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019. 1, 2, 6, 7

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5, 7

[17] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5900–5909. Computer Vision Foundation / IEEE, 2020. 3, 5, 6, 7, 8

[18] ISO/IEC JTC1 SC17 WG3. Portrait Quality - Reference Facial Images for MRTD. International Civil Aviation Organization, 2018. 2

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

[20] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 2382-37:2017 Information technology - Vocabulary - Part 37: Biometrics. International Organization for Standardization, 2017. 1

[21] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 19795-1:2021 Information technology — Biometric performance testing and reporting — Part 1: Principles and framework. International Organization for Standardization, 2021. 5

[22] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 133–142. ACM, 2002. 2

[23] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. Cross-quality LFW: A database for analyzing cross- resolution image face recognition in unconstrained environments. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–5. IEEE, 2021. 5, 7

[24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746. IEEE Computer Society, 2017. 3

[25] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 212–220, 2017. 3

[26] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1040–1049. IEEE Computer Society, 2017. 1, 6, 7

[27] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018. 5, 7, 8

[28] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021. 1, 2, 3, 5, 6, 7, 8

[29] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 1, 6, 7

[30] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 1

[31] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE CVPRW, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017. 5, 7

[32] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7670–7679. Computer Vision Foundation / IEEE, 2021. 1, 2, 5, 6, 7

[33] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9. IEEE Computer Society, 2016. 5, 7

[34] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019. 1, 2, 5, 6, 7

[35] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, Houston, TX, USA, September 28 - October 1, 2020*, pages 1–11. IEEE, 2020. 1

[36] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659. Computer Vision Foundation / IEEE, 2020. 1, 2, 6, 7

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. IEEE Computer Society, 2018. 3

[38] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face verification. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020. 1, 2

[39] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. 2, 4, 5, 6

[40] 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. 5, 7

[41] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 5, 7