GRAFIQS: Face Image Quality Assessment Using Gradient Magnitudes

Jan Niklas Kolf[1,2], Naser Damer[1,2], Fadi Boutros[1]
[1]Fraunhofer Institute for Computer Graphics Research IGD, Germany
[2]Technical University of Darmstadt, Germany
{jan.niklas.kolf, naser.damer, fadi.boutros}@igd.fraunhofer.de

Abstract

Face Image Quality Assessment (FIQA) estimates the utility of face images for automated face recognition (FR) systems. We propose in this work a novel approach to assess the quality of face images based on inspecting the required changes in the pre-trained FR model weights to minimize differences between testing samples and the distribution of the FR training dataset. To achieve that, we propose quantifying the discrepancy in Batch Normalization statistics (BNS), including mean and variance, between those recorded during FR training and those obtained by processing testing samples through the pretrained FR model. We then generate gradient magnitudes of pretrained FR weights by backpropagating the BNS through the pretrained model. The cumulative absolute sum of these gradient magnitudes serves as the FIQ for our approach. Through comprehensive experimentation, we demonstrate the effectiveness of our training-free and quality labeling-free approach, achieving competitive performance to recent state-of-the-art FIQA approaches without relying on quality labeling, the need to train regression networks, specialized architectures, or designing and optimizing specific loss functions.[1]

1. Introduction

Face image quality assessment (FIQA) estimates the utility of the captured sample for face recognition (FR) [1]. FIQA algorithms process an image to produce a scalar quality score. This score serves as a scalar measurement to quantify the quality of a face image in terms of its suitability for use in FR systems [15]. Ensuring high FIQ can improve the performance of FR and enhance their effectiveness in applications such as automated border control [15, 35]. FIQA primarily targets assessing the utility of a face image for automated FR [1, 15, 35] rather than assessing the perceived image quality [9]. Perceived image quality assessment has been addressed in the literature by general image quality assessment (IQA) methods [27, 29, 30], which provide insights into image quality from human perception. IQA does not necessarily reflect the utility of face image for FR [14], e.g., a face image may have high perceived quality according to IQA metrics; however, it may still be relatively less suitable for FR due to factors such as occlusion. On the other hand, FIQA algorithms provide a targeted assessment of the image's utility for FR tasks. Thus, FIQA approaches in the literature demonstrate superiority over IQA in assessing the utility of face image FR, as presented in [3, 10, 28, 32].

State-of-the-art (SOTA) FIQA can be categorized into two main categories. The approaches in the first category focus on labeling face images with quality labels and then training a regression network to predict the FIQ score of test samples [6, 11, 19, 32]. Best-Rowden et al. [6] proposed learning FIQ based on target quality labels obtained from either human assessment of FIQ or quality score labels computed from genuine comparisons. FaceQNet [19] proposed to label the images with quality scores based on the genuine comparison between a sample and an ICAO compliant sample and then used the labeled data to train a regression network. SDD-FIQ [32] generated quality labels to train a regression network using Wasserstein distance between genuine and imposter similarity distributions. RankIQ [11] proposed the learning-to-rank approach that predicts the sample quality as a rank using FRs performances on several datasets. The approaches in the second category [10, 28, 37, 38] assess the utility of face images by either learning to estimate the quality from face embedding properties during FR training or inspecting the robustness of embeddings. To assess the image quality, PFE [37] estimates an uncertainty of face embedding in the latent space and considers it as a reverse measure of face quality. MagFace [28] proposed to learn a universal feature embedding and then utilize the magnitude of the embedding to measure the quality of a given face image. SER-FIQ [38] inspected the robustness of face embedding and considered it as FIQ by passing the face images into an FR network multiple times, each with a different random dropout pattern, to output several embeddings of each sample. Then, the quality

[1] https://github.com/jankolf/GraFIQs

score is obtained by calculating the sigmoid of the negative mean of the Euclidean distances between the embeddings. CR-FIQA [10] proposed to estimate the FIQ of a sample by learning to predict its relative classifiability, which is measured based on the allocation of the sample feature representation in the embedding space to its class center and the nearest negative class center. DifFIQA [4] leverages a diffusion model to explore the stability of embeddings of face images through image perturbations caused by the processes of adding noise and denoising. eDifFIQA [5] applied knowledge distillation to DifFIQA [4], aiming at reducing the computational cost of DifFIQA [4].

This paper presents a pioneering approach, namely GRAFIQS, that leverages the gradient magnitude during the backpropagation step of pretrained FR model to assess the FIQ. Unlike recent high-performing FIQA approaches [10, 28, 32, 38] that rely on face embeddings, our approach does not require quality labeling and training of regression networks [19, 32], the need of specialized architectures [38], or designing and optimizing specific loss functions [28]. In contrast, our approach inspects the necessary changes in the pretrained FR model parameters to minimize the difference between test samples and the FR model training dataset distribution. To achieve this, we propose to measure the shift in Batch Normalization statistics (BNS), mean and variance, between the ones recorded during the FR training and those obtained by passing the test samples into the FR model. Subsequently, after extracting the BNS, we backpropagate the difference between BNS into the pretrained model to generate gradient magnitudes, whose absolute sum serves as FIQ. Note that the BNS of the training dataset are integral parts of the pretrained model parameters and do not require retraining the model to extract these values. Through extensive experiments, we prove that our training-free and quality labeling-free approach can achieve competitive results with recent SOTA FIQA methods. This novel perspective on FIQA provides a new way of assessing the quality of face images based on the gradient magnitudes of the pretrained FR model.

2. Methodology

This section presents our novel training-free and label-free FIQA technique that leverages the gradient magnitudes obtained from forwarding and backpropagating any test sample in a given pretrained FR model.

Conventionally, gradient optimization is only used during the training phase of a FR model to update the model parameters in the backpropagation step with respect to the training loss function [7]. The model parameters are iteratively updated until the model is converged, i.e., gradient descent converged to a local minimum. The parameters of the model are updated in a way that the loss function achieves the minimum value on the training dataset

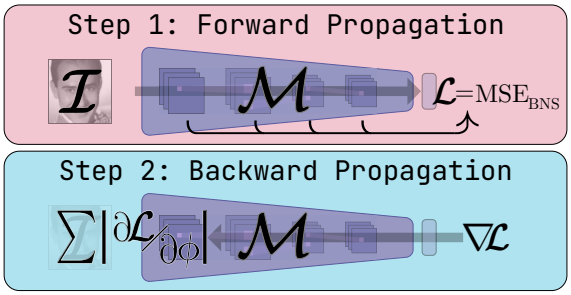


Figure 1. An overview of the proposed GRAFIQS for assessing the quality of unseen testing samples. Sample $\mathcal{I}$ is passed into the pretrained FR model and BNS are extracted. Then, the MSE between the BNS obtained by processing the testing sample and the one recorded during the FR training is calculated. The MSE is backpropagated through the pretrained FR to extract the gradient magnitudes of parameter group ϕ. Finally, the absolute sum of gradient magnitudes of ϕ is calculated and utilized as FIQ.

[2, 33]. Once the model is trained, the model weights are frozen, and no gradient optimizations are required or performed during the inference phase. In the case that the FR model is not designed and optimized to predict the FIQ, which is the case in our approach, the output of the FR can not be used directly to assess the utility of the test sample. To address the aforementioned challenge, we propose to assess the utility of any given test sample by calculating the required changes in the pretrained FR model weights to minimize the difference between the test sample and the model training data distribution. To achieve this objective, we calculate the difference between training data and test data distribution using mean squared error (MSE) between BNS [23, 39]. We then propose to backpropagate the MSE to generate gradient magnitudes from the pretrained model, in which their absolute sums are used to assess the sample quality, as detailed in Section 2.3.

Figure 1 presents an overview of our proposed approach to estimate the utility of test samples using a pretrained FR model. In this approach, a test sample is forwarded into the pretrained FR model, and a Batch Normalization (BN) loss is calculated (details in Section 2.2). Then, we calculate the gradient of the loss function with respect to the model parameters using backpropagation. The magnitude of the resulting gradients is used to measure the required changes in the model parameters to minimize the MSE between BNS of the training data and the BNS of the given sample. If high magnitudes are present within the gradients of the FR model parameters, this indicates that the parameters require large changes to minimize the MSE based on the input [33]. On the contrary, low magnitudes indicate that the parameters do not have to be updated to a large degree to achieve minimal MSE loss of BNS. We theorize that, given the BNS calculated on a training dataset and the BNS of an input image, high gradient magnitudes resulting from MSE loss indicate a low utility of the input image,

and vice versa, further explained in Section 2.1. Although calculating the BN loss would directly indicate the quality of the test samples, we show in this paper that the gradient magnitude (defined as how strong the required changes are) with respect to the model parameters is more informative in estimating the utility of the sample. This concept and the required loss function will be introduced in this section, and the success of the underlying approach will be empirically proven later in this work.

This section first presents fundamentals on gradient-based optimization. Then, we present and formalize the basic principles of GRAFIQS and the fundamental loss function introduced in this work.

2.1. Gradient-Based Optimization

In essence, the objective of a given optimization problem is to find an assignment for a given set of parameters $\theta \in \mathbb{R}^d$ that minimizes a given loss or cost function $\mathcal{L}(\theta)$ [2]:

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta) \tag{1}$$

A machine learning model $\mathcal{M}_\theta$, e.g. a FR model, is a single or a sequence of mathematical functions (also referred to as layers [7]) that maps an input, e.g. input image $\mathcal{I}$, to an output $y = \mathcal{M}_\theta(\mathcal{I})$, also called prediction, using the set of parameters θ [2].

For a simpler explanation, consider the supervised learning approach. In this approach, dataset $\mathcal{D}$ is given, which consists of input training samples $\mathcal{I}$ and their corresponding labels $\hat{y}$. The objective of optimization is to learn a set of parameters θ such that the prediction $y = \mathcal{M}_\theta(\mathcal{I})$ matches $\hat{y}$ as closely as possible [2]. The model $\mathcal{M}_\theta$ and therefore the parameters θ are fitted to $\mathcal{D}$ through this optimization process. The scalar-valued loss function $\mathcal{L}(\theta)$ is used to quantify the mismatch between the prediction $\mathcal{M}_\theta(\mathcal{I})$ and the desired target output $\hat{y}$ [33]. Therefore, minimizing $\mathcal{L}(\theta)$ minimizes the mismatch between prediction $y = \mathcal{M}_\theta(\mathcal{I})$ and target $\hat{y}$.

Since analytical solutions that minimize $\mathcal{L}(\theta)$ and yield θ^* are often not feasible, a commonly used method in machine learning is gradient descent [7, 33], a variation of gradient-based optimization. These optimization techniques require that the derivative of $\mathcal{L}(\theta)$ w.r.t. the model parameters θ, $\nabla_\theta \mathcal{L}(\theta)$, is defined [7].

The gradient $\nabla_\theta \mathcal{L}(\theta)$ is defined as the vector of partial derivatives of $\mathcal{L}(\theta)$ w.r.t. to each parameter $\theta_i \in \theta$:

$$\nabla_\theta \mathcal{L}(\theta) = \left[\frac{\partial \mathcal{L}(\theta)}{\partial \theta_1} \quad \cdots \quad \frac{\partial \mathcal{L}(\theta)}{\partial \theta_m} \right]^T \tag{2}$$

The gradient $\nabla_\theta \mathcal{L}(\theta)$ specifies the direction of the fastest increase in $\mathcal{L}(\theta)$ [33]. The magnitude of $\nabla_\theta \mathcal{L}(\theta)$ specifies the rate of change in the direction of the gradient [33]. Therefore, changing the parameters in the direction of $-\nabla_\theta \mathcal{L}(\theta)$ reduces $\mathcal{L}(\theta)$. As the gradient only gives a momentary and local rate of change that does not hold over larger distances, steps in the direction of the negative gradient have to be adjusted by a scaling parameter $\alpha > 0$ [2]. The parameter α is named the learning rate [2], as it specifies the rate of change in the direction of the negative gradient. To calculate θ^* that minimize $\mathcal{L}(\theta)$, an iterative update procedure with an initial set of parameters θ^0 and learning rate α^0 is used [2]:

$$\theta^{t+1} = \theta^t - \alpha^t \nabla_{\theta^t} \mathcal{L}(\theta^t) \tag{3}$$

With every parameter update, the overall mismatch between predictions $y = \mathcal{M}_\theta(\mathcal{I})$ and targets $\hat{y}$ of dataset $\mathcal{D}$, measured by $\mathcal{L}(\theta)$, is reduced. After sufficient update steps are performed, it is assumed that $\theta^{t_{\max}} \approx \theta^*$.

2.2. Batch Normalization

The previous section presented preliminary on FR model training and calculating gradient magnitudes, which serve as a basis for our concept to assess the FIQ. Details on FR training is provided in Section 3. Calculating gradient magnitudes, as we presented in Section 2.1, requires a loss function, e.g. multi-class classification loss, which is not feasible when training and test data are identity-disjoint (the case for the FR model). Alternatively, we propose to calculate MSE (as a loss function) between the BNS of the pretrained FR model and the BNS extracted by passing test samples into the pretrained FR. We then backpropagate MSE to calculate gradient magnitudes. Therefore, we briefly provide in this section insight into BN.

BN [23] is a technique applied in certain machine learning models to improve the speed, performance, and stability of the training phase [7, 33, 34]. It is included several times as a dedicated layer, the BN layer (BNL), within the sequence of layers [33]. The BNL receives a set (batch) of outputs (referred to as activations) from the previous layer as input [7]. The activations of the input batch are normalized using the mean and standard deviation calculated throughout the batch. During inference, mean μ and standard deviation σ are used for normalization that were calculated during the model training phase using exponential moving averages [7, 39], forming BNS.

To determine whether a particular input sample adheres to the same data distribution as the training data, the BNS, mean μ and standard deviation σ of the training data, can be compared to the BNS obtained from that sample during inference (μ' and σ') [39]. This technique serves as an effective method for domain adaptation, a process widely utilized in various applications [26, 39]. Note that BNS are part of the pretrained model parameters and do not require accessing the training data to extract these statistics [7].

To quantify the divergence between the BNS of the training data and that of an input sample, a loss function based on the mean squared error is devised. This loss function evaluates the discrepancy in BNS across all BNL in the model

and is defined as follows:

$$\text{MSE}_{\text{BNS}} = \frac{1}{|\text{BNL}|} \sum_{l \in \text{BNL}} \|\mu_l - \mu'_l\|_2^2 + \|\sigma_l - \sigma'_l\|_2^2, \quad (4)$$

where, μ_l and σ_l denote the mean and standard deviation of the BNS for the l-th layer recorded during training, while μ'_l and σ'^2_l represent the mean and variance for the same layer obtained from the input sample during inference. The norm $\|\cdot\|_2^2$ indicates the squared Euclidean (L2) norm.

A higher MSE_{BNS} signifies a greater disparity between the input sample and the data distribution of the training dataset, compared to an input sample that yields a lower MSE_{BNS}.

2.3. Deriving GraFIQs Approach

Although the MSE_{BNS} value can be directly used to evaluate the FIQ of a given input sample (as presented in Section 4), it does not reflect the required changes on the activation level to match the distribution between pretrained model and test sample.

Consider two different input samples, denoted as $\mathcal{I}_i$ and $\mathcal{I}_j$. At a given BNL l, BNL l receives activations as input that have respective mean and variance values $\mu'_{l,i}$, $\sigma'_{l,i}$ for $\mathcal{I}_i$ and $\mu'_{l,j}$, $\sigma'_{l,j}$ for $\mathcal{I}_j$, respectively. The MSE between the BNS of the training dataset and the BNS of sample $\mathcal{I}_m$, $m \in \{i, j\}$, is defined as $\text{MSE}_{l,m} = \|\mu_l - \mu'_{l,m}\|_2^2 + \|\sigma_l - \sigma'_{l,m}\|_2^2$. Assume that $\mu'_{l,i} = \mu'_{l,j}$ and $\sigma'_{l,i} = \sigma'_{l,j}$ and therefore $\text{MSE}_{l,i} = \text{MSE}_{l,j}$. This does not imply that the activations of the layer preceding the BNL, which result from the inputs $\mathcal{I}_i$ and $\mathcal{I}_j$, are identical. Different activations of the samples, which are generated by the parameters in the pretrained FR model, can lead to the same mean and standard deviation. However, the necessary changes within the parameters to minimize the difference between the BNS of the training set (μ'_l and σ'^2_l) and the BNS of the respective sample, $\mathcal{I}_i$ ($\mu'_{l,i}$, $\sigma'_{l,i}$) or $\mathcal{I}_j$ ($\mu'_{l,j}$, $\sigma'_{l,j}$), can differ significantly. Using only the mean and standard deviation, either at a single BNL or over the entire network, is not enough to fully capture the required changes of the parameters to output activations that match the BNS of the training dataset. To overcome the limitations that arise when the mean and standard deviations are used, we propose a new method that considers not only MSE_{BNS}, but also the necessary parameter changes that are described by gradient magnitudes.

This approach, gradient magnitude based FIQA (GraFIQs), is specified in the following. An input image $\mathcal{I}$ is fed into the network and the loss $\mathcal{L}_{\text{BNS}} = \text{MSE}_{\text{BNS}}$ is calculated. The gradients $\nabla_\theta \mathcal{L}_{\text{BNS}}$ are calculated by backpropagation through the network. The gradient magnitudes, i.e. the absolute value of the gradient, from $\partial \mathcal{L}_{\text{BNS}} / \partial \phi$ are extracted and summed to give the total required changes in the parameters ϕ. The FIQ is therefore calculated as $\sum |\partial \mathcal{L} / \partial \phi|$. Gradients can be extracted either at image level, $\phi = \mathcal{I}$, or for intermediate layers, $\phi = \text{B}i$, as detailed in Section 4.

3. Experimental Setup

Pretrained model architecture The proposed approach GraFIQs is demonstrated using two pretrained FR models, ResNet100 and ResNet50, trained with ArcFace loss on MS1MV2 [12, 17] and CASIA-WebFace [40], respectively. The pretrained models were released by [12] and they are publically available. The results of the pretrained ResNet50 model are provided in supplementary materials.

Evaluation Benchmarks We reported the achieved results on the following benchmarks: Labeled Faces in the Wild (LFW) [21], AgeDB-30 [31], Celebrities in Frontal-Profile in the Wild (CFP-FP) [36], Cross-age LFW (CALFW) [42], Adience [13], Cross-Pose LFW (CPLFW) [41] and Cross-Quality LFW (XQLFW) [25]. These benchmarks contain challenging pairs with large age variations (AgeDB-30 and CALFW), head-pose variations (CFP-FP and CPLFW), and face image quality variations (XQLFW). These benchmarks are chosen to be aligned with recent SOTA FIQ approaches [10].

Evaluation Metric We evaluate the FIQA by plotting Error-versus-Discard Characteristic (EDC) Curves [15, 16]. It should be noted that several SOTA approaches in the literature referred to EDC as ERC (Error-versus-Reject Curves) [4]. The EDC is a widely used metric for evaluating FIQA performance [15, 16]. EDC curve demonstrates the effect of discarding a fraction of face images, of the lowest quality, on face verification performance in terms of False None Match Rate [24] (FNMR) at a specific threshold calculated at fixed False Match Rate [24] (FMR). Following SOTA FIQA approaches [4, 10, 28], the EDC curves for all benchmarks are plotted at two fixed FMRs, $1e-3$ and $1e-4$. We also report the Area under the Curve (AUC) of the EDC, to provide a quantitative aggregate measure of verification performance across all rejection ratios.

FR Models To provide insight into the generalizability of our approach, we report the verification performance at different quality discard rates using four different FR models. The utilized models are: ArcFace [12], ElasticFace (ElasticFace-Arc) [9], MagFace [28], and CurricularFace [22]. We utilized the pretrained models provided by the corresponding authors [9, 12, 22, 28]. All models were originally trained on MS1MV2 [12, 17] and used ResNet100 [18] as network architecture. All models process 112×112 aligned and cropped images to produce feature embedding of size 512-D.

The results are reported under two protocols, same and cross model. The results reported using ArcFace [12] follow the same model protocol, where ArcFace is used to calculate the quality of images and for reporting the verification accuracies at different quality discard rates. The results

Table 1. Conceptual comparison on the design choices between our GRAFIQS and recent FIQA approaches in the literature.

FIQA	Quality Labels	Specific Architecture	Requires Training	Custom Loss	Inference: Feed-Forwards	Inference: Backwards	Inference: Embedding-Level	Inference: Gradient-Level
CR-FIQA [10]	✗	✗	✓	✓	1	0	✓	✗
DifFIQA [4]	✗	✗	✓	✓	1	0	✓	✗
eDifFIQA [5]	✓	✗	✓	✓	1	0	✓	✗
MagFace [28]	✗	✗	✓	✓	1	0	✓	✗
FaceQnet [19]	✓	✗	✓	✗	1	0	✓	✗
SDD-FIQA [32]	✓	✗	✓	✗	1	0	✓	✗
SER-FIQ [38]	✗	✓	✗	✗	100	0	✓	✗
PFE [37]	✗	✗	✓	✓	1	0	✓	✗
GRAFIQS (Our)	✗	✗	✗	✗	1	1	✗	✓

reported using ElasticFace [9], MagFace [28], and CurricularFace [22] are cross-model evaluation protocols, where ArcFace is used to calculate the quality of images and ElasticFace, MagFace, and CurricularFace were used to report the verification accuracies at different quality discard rates.

Baseline and Comparisons with SOTA FIQ The achieved results by our GRAFIQS are compared to the three general IQA methods, BRISQUE [29], RankIQA [27], and DeepIQA [8] and to nine SOTA FIQA methods, RankIQ [11], PFE [37], SER-FIQ [38], FaceQnet (v1 [19]) [19, 20], MagFace [28], SDD-FIQA [32], CR-FIQA [10], DifFIQA [4] and eDifFIQA [5].

4. Results

This section first provides empirical proof of the use of gradient magnitude to assess the utility of face images. Table 2 presents the verification performance as AUC at FMR= $1e-3$ and FMR= $1e-4$ reported on seven benchmarks. The results are reported using two protocols, same model and cross-model protocol. All quality values are calculated using the ArcFace model and the verification performances at different discard rates are provided using ArcFace (same model) as well as ElasticFace, MagFace and CurricularFace (cross-model). The corresponding EDCs are provided in Figure 2.

For each protocol, we first evaluate the use of the BNS based MSE_{BSN} function (Equation 4) as FIQ. The function MSE_{BSN} measures the change in the distribution between the model training data and test samples. In this experiment, we use MSE_{BSN} directly to assess the quality and we do not backpropagate MSE_{BSN} through the network. Higher values resulting from MSE_{BSN} indicate low quality and vice versa. It can be seen in Figure 2 (black, dashed line) that discarding samples with high MSE_{BSN} (low quality) improves the verification accuracies where EDC (in both same and cross model evaluation protocols) is dropped when discarding a fraction of low-quality samples.

In the next study, we evaluate the use of gradient magnitude to assess the quality of face images. As we discussed in Section 2, the MSE_{BSN} does not fully describe the strength and level of difference between the sample and the data distribution learned from the model. For each of the evaluation protocols considered, we provide detailed evaluations of the sums of absolute gradient magnitudes (as FIQ) extracted during the backpropagation step from four intermediate layers and on the image level (noted as as $\mathcal{I}$). The network architecture of utilized ArcFace models is ResNet100 which consists mainly of 4 stages, each consisting of several stacked residual blocks equal to 3, 4, 23, and 3, respectively. We consider the output of each of the stages (during the backpropagation step) as intermediate gradient magnitude, noted as B1, B2, B3, and B4, respectively. It can be clearly noticed that utilizing gradient magnitudes that result from $\mathcal{L}_{\text{BSN}}$ and are extracted from different intermediate layers to assess face image quality achieved higher verification than directly using MSE_{BSN} in most of the considered settings. Specifically, utilizing gradient magnitude from B2 achieved the best overall performance. The results are also supported by EDC curves in Figure 2, where EDC is dropped when discarding a fraction of low-quality (high gradient magnitude) samples.

4.1. Comparison with SOTA FIQ

We compared our achieved results with three IQA approaches, BRISQUE, RankIQA, and DeepIQA, as well as with seven SOTA FIQA approaches, RankIQ, PFE, SER-FIQ, FaceQnet, MagFace, SDD-FIQA, and CR-FIQA. Table 3 presents the achieved results as AUC of EDR calculated at two FMR thresholds, FRM$1e-3$ and FMR$1e-4$, using four different FR models. The corresponds EDC curves are presented in Figure 3. The results are reported on seven benchmarks described in Section 3. The results of our GRAFIQS approach are reported based on the best achieved settings from Table 2 (i.e. B2) which was previously discussed in Section 4.

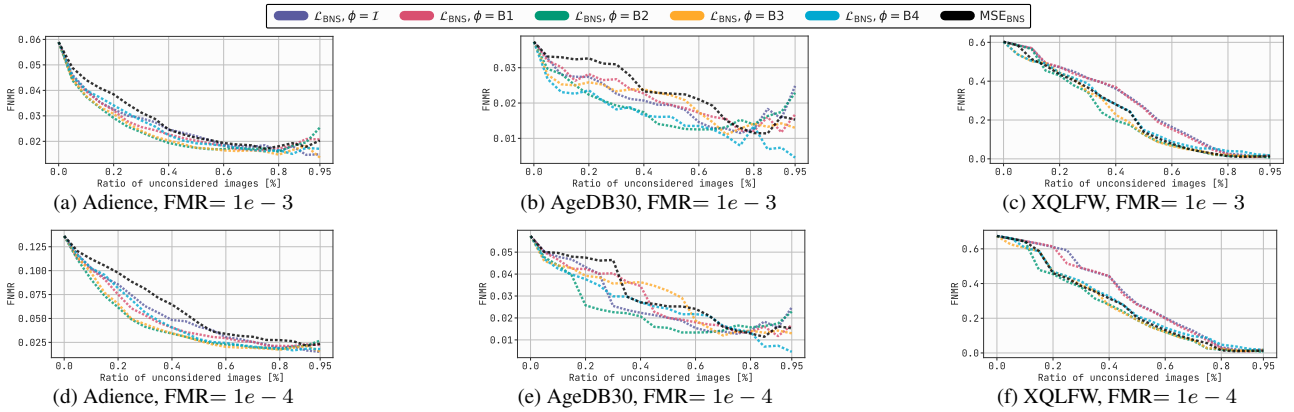
We made the following observations from the results reported in Table 3:

- In comparison to IQA approaches, our GRAFIQS outperformed all IQA approaches, BRISQUE [29], RankIQ [27] and DeepIQA [8], in all settings, as shown in Table 3 and Figure 3.
- On the benchmarks that include large age gaps (Adience, AgeDB-30 and CALFW) and in comparison to FIQA approaches, our GRAFIQS achieved competitive results to the SOTA approaches. For example, our GRAFIQS achieved competitive results, even outperformed them in many settings, to SDD-FIQA [32], FaceQNet [19], PFE [37], DifFIQA [4] and SER-FIQ [38]. Also, our GRAFIQS scored slightly behind SOTA approaches, eDifFIQA [5], CR-FIQA [10] and MagFace [10] on benchmarks with large age gaps, Table 3 and Figure 3.
- On the benchmarks that include large pose variations (CFP-FP and CPLFW) and compared to FIQA approaches, our GRAFIQS scored slightly behind CR-FIQA

Table 2. The achieved AUC of EDC by using two approaches presented in this paper, MSE of BNS (MSE_{BNS}) and gradient magnitudes ($\mathcal{L}_{\text{BNS}}$), and under different settings. The gradient magnitudes are extracted during the backpropagation step from different intermediate layers, B1, B2, B3 and B4 ($\phi = \text{B1} - \phi = \text{B4}$) as well as on the pixel level ($\phi = \mathcal{I}$). The results are reported under two operation threshold FMR= $1e-3$ and FMR= $1e-4$ and under two protocols, same model (ArcFace) and cross-model (ElasticFace, MagFace and CurricularFace). Utilizing gradient magnitudes from B2 achieved the best overall performance.

FR	Loss $\mathcal{L}$	FIQ	Gradient	Adience [13]		AgeDB30 [31]		CFP-FP [36]		LFW [21]		CALFW [42]		CPLFW [41]		XQLFW [25]		Mean AUC	
				$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$
ArcFace [12]	-	MSE_{BNS}	-	0.0262	0.0578	0.0223	0.0287	0.0151	0.0213	0.0029	0.0035	0.0631	0.0669	0.0475	0.0683	0.2207	0.2612	0.0568	0.0725
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \mathcal{I}$	0.0245	0.0502	0.0196	0.0252	0.0117	0.0168	0.0030	0.0039	0.0609	0.0658	0.0489	0.0665	0.2664	0.3200	0.0621	0.0783
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B1}$	0.0239	0.0464	0.0202	0.0269	0.0117	0.0170	0.0027	0.0034	0.0615	0.0656	0.0481	0.0677	0.2627	0.3130	0.0615	0.0771
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B2}$	0.0225	0.0403	0.0176	0.0219	0.0070	0.0111	0.0032	0.0038	0.0644	0.0692	0.0415	0.0612	0.2058	0.2447	**0.0517**	**0.0646**
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B3}$	0.0222	0.0406	0.0194	0.0276	0.0081	0.0125	0.0037	0.0041	0.0595	0.0636	0.0457	0.0675	0.2105	0.2514	0.0527	0.0668
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B4}$	0.0239	0.0458	0.0155	0.0246	0.0186	0.0277	0.0031	0.0037	0.0562	0.0598	0.0598	0.0853	0.2284	0.2721	0.0579	0.0741
ElasticFace [9]	-	MSE_{BNS}	-	0.0284	0.0540	0.0220	0.0236	0.0134	0.0168	0.0027	0.0035	0.0613	0.0630	0.0455	0.0595	0.1983	0.2335	0.0531	0.0648
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \mathcal{I}$	0.0262	0.0474	0.0202	0.0217	0.0116	0.0142	0.0032	0.0039	0.0589	0.0604	0.0455	0.0573	0.2515	0.2838	0.0596	0.0698
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B1}$	0.0254	0.0448	0.0202	0.0217	0.0114	0.0140	0.0026	0.0034	0.0595	0.0608	0.0452	0.0695	0.2501	0.2802	0.0592	0.0706
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B2}$	0.0233	0.0394	0.0182	0.0200	0.0070	0.0091	0.0029	0.0037	0.0614	0.0632	0.0393	0.0633	0.1930	0.2319	**0.0493**	**0.0615**
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B3}$	0.0232	0.0396	0.0193	0.0207	0.0074	0.0100	0.0035	0.0041	0.0574	0.0591	0.0422	0.0668	0.1947	0.2384	0.0497	0.0627
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B4}$	0.0250	0.0431	0.0158	0.0169	0.0133	0.0179	0.0030	0.0035	0.0539	0.0554	0.0498	0.0769	0.2058	0.2615	0.0524	0.0679
MagFace [28]	-	MSE_{BNS}	-	0.0272	0.0587	0.0222	0.0366	0.0178	0.0328	0.0033	0.0041	0.0635	0.0650	0.0496	0.1000	0.2568	0.3064	0.0629	0.0862
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \mathcal{I}$	0.0254	0.0508	0.0204	0.0355	0.0156	0.0290	0.0033	0.0043	0.0610	0.0630	0.0502	0.0984	0.2953	0.3355	0.0673	0.0881
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B1}$	0.0246	0.0472	0.0209	0.0395	0.0155	0.0287	0.0030	0.0037	0.0616	0.0634	0.0499	0.1355	0.2913	0.3384	0.0667	0.0938
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B2}$	0.0233	0.0419	0.0182	0.0253	0.0087	0.0186	0.0033	0.0041	0.0640	0.0652	0.0428	0.0987	0.2524	0.3018	0.0590	**0.0794**
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B3}$	0.0229	0.0420	0.0203	0.0409	0.0098	0.0187	0.0038	0.0045	0.0590	0.0603	0.0462	0.1073	0.2479	0.2905	**0.0586**	0.0806
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B4}$	0.0247	0.0459	0.0169	0.0391	0.0208	0.0380	0.0033	0.0039	0.0557	0.0569	0.0577	0.1266	0.2654	0.3187	0.0635	0.0899
Curricular-Face [22]	-	MSE_{BNS}	-	0.0245	0.0496	0.0212	0.0250	0.0145	0.0191	0.0029	0.0035	0.0621	0.0655	0.0426	0.0618	0.1863	0.2163	0.0506	0.0630
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \mathcal{I}$	0.0233	0.0429	0.0211	0.0238	0.0120	0.0161	0.0034	0.0039	0.0591	0.0623	0.0422	0.0578	0.2233	0.2652	0.0549	0.0674
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B1}$	0.0229	0.0410	0.0208	0.0237	0.0117	0.0160	0.0028	0.0034	0.0597	0.0624	0.0424	0.0728	0.2263	0.2639	0.0552	0.0690
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B2}$	0.0220	0.0365	0.0167	0.0200	0.0068	0.0099	0.0033	0.0038	0.0610	0.0641	0.0369	0.0663	0.1713	0.1959	**0.0454**	**0.0566**
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B3}$	0.0214	0.0357	0.0179	0.0211	0.0079	0.0117	0.0038	0.0043	0.0591	0.0616	0.0387	0.0694	0.1860	0.2143	0.0478	0.0597
	$\mathcal{L}_{\text{BNS}}$	$\sum \lvert\partial\mathcal{L}/\partial\phi\rvert$	$\phi = \text{B4}$	0.0225	0.0390	0.0151	0.0190	0.0159	0.0240	0.0033	0.0037	0.0550	0.0575	0.0448	0.0742	0.2025	0.2300	0.0513	0.0639

Figure 2. EDC for FNMR@FMR=$1e-3$ and FNMR@FMR=$1e-4$ of our proposed method using $\mathcal{L}_{\text{BNS}}$ as backpropagation loss and absolute sum as FIQ. The gradients at image level ($\phi = \mathcal{I}$), and block levels ($\phi = \text{B1} - \phi = \text{B4}$) are used to calculate FIQ. MSE_{BNS} as FIQ is shown in black. Results are shown on benchmarks Adience, AgeDB30 and XQLFW datasets using ArcFace model. The proposed GRAFIQS method leads to lower verification error when images with the lowest utility score estimated from gradient magnitudes are rejected. Furthermore, estimating FIQ by backpropagating $\mathcal{L}_{\text{BNS}}$ yields significantly better results than using MSE_{BNS} directly.



(a) Adience, FMR= $1e-3$ (b) AgeDB30, FMR= $1e-3$ (c) XQLFW, FMR= $1e-3$

(d) Adience, FMR= $1e-4$ (e) AgeDB30, FMR= $1e-4$ (f) XQLFW, FMR= $1e-4$

[10], DifFIQA [4], and eDifFIQA [5], and outperformed all other approaches in most settings.

- On LFW and in comparison to FIQA approaches, our GRAFIQS achieved slightly lower results than the SOTA FIQA approaches.
- The XQLFW benchmark, derived from LFW, includes pairs of images with maximal disparities in quality, which are selected based on quality scores from BRISQUE [29] and SER-FIQ [38], ensuring that the images are categorized as either exceptionally high or low quality. On XQLFW and in comparison to FIQA approaches, our GRAFIQS outperformed well-performing SOTA approaches, including MagFace [28], SDD-FIQA [32] and CR-FIQA [10] and achieved very close results to DifFIQA [4], eDifFIQA [5], and the labeling approach SER-FIQ [38].
- Under same-model (ArcFace) and cross-model (ElasticFace, MagFace, and CurricularFace) experimental settings, our GRAFIQS achieved consistent results, proving the generalizability of our concept for FIQA, as shown in Table 3.

To conclude, the presented novel concept proved to be highly effective. Our gradient magnitude-based approach GRAFIQS achieved very competitive results to the SOTA approaches without the need for quality labeling and regression network training [4, 5, 19, 32] or designing and optimizing special losses [10, 28], as summarized in Table 1. Unlike SER-FIQ which requires passing the sam-

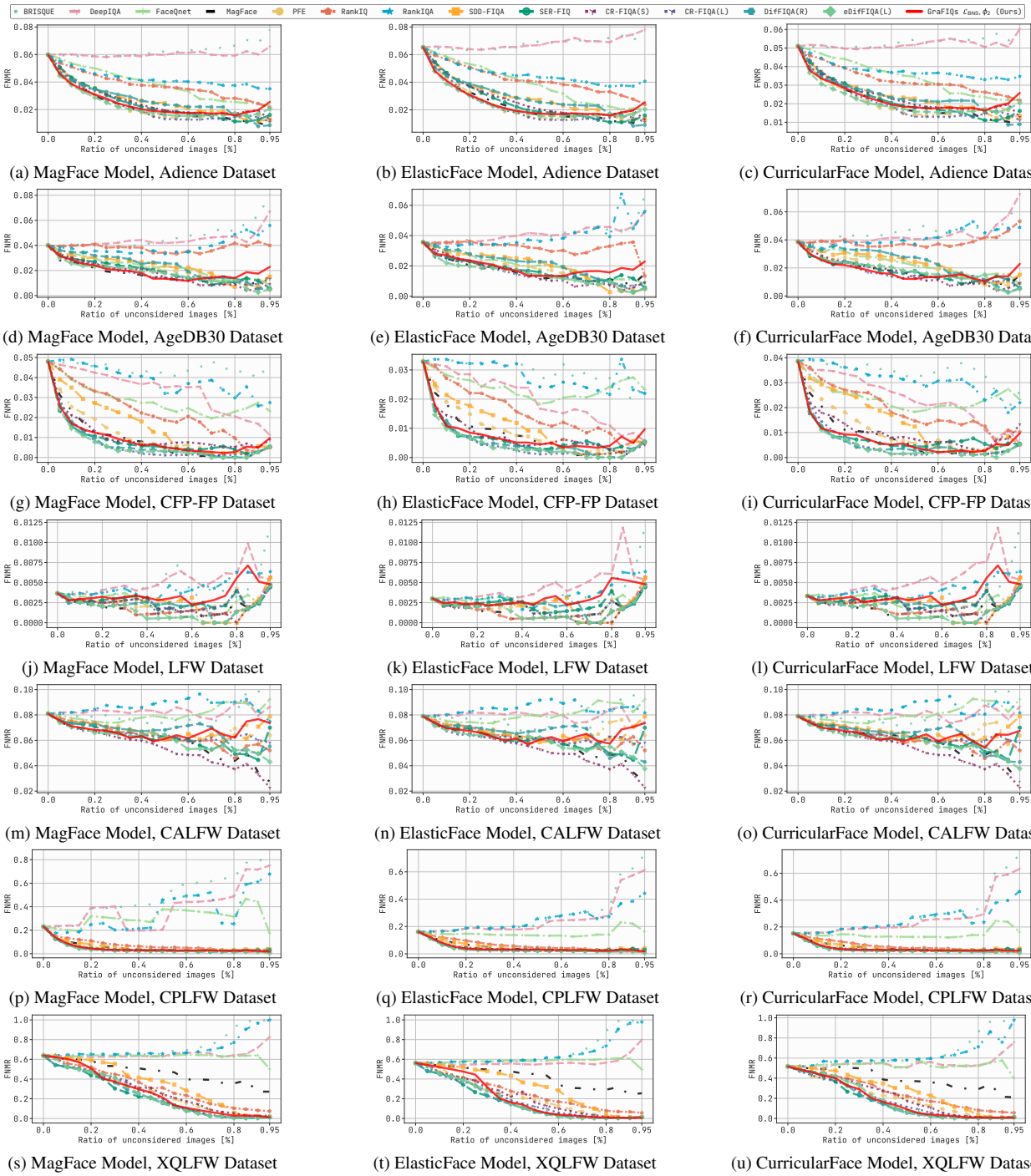
Table 3. The AUCs of EDC achieved by our GRAFIQS and the SOTA methods under different experimental settings. The notions of $1e-3$ and $1e-4$ indicate the value of the fixed FMR at which the EDC curves (FNMR vs. reject) were calculated. The results are compared to three IQA and nine ten FIQA approaches. The mean AUC overall evaluation datasets (except XQLFW as it was labeled by SER-FIQ [38]) at FMR= $1e-3$ and FMR= $1e-4$ per method are shown in the last column. The XQLFW dataset uses SER-FIQ (marked with *) as FIQ labeling method.

FR		Method	Adience[13]		AgeDB-30[31]		CFP-FP[36]		LFW[21]		CALFW[42]		CPLFW[41]		XQLFW[25]		Mean AUC	
			$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e^{-3}$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$
ArcFace[12]	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	0.0777	0.1034
		RankIQA[27]	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	0.0733	0.0962
		DeepIQA[8]	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	0.0727	0.0992
	FIQA	RankIQ[11]	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	0.0348	0.0516
		PFE[37]	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	0.0266	0.0355
		SER-FIQ[38]	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*	0.0244	0.0333
		FaceQnet[19, 20]	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	0.0514	0.0642
		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	0.0252	0.0332
		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	0.0289	0.0393
		CR-FIQA(S) [10]	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	*0.0234*	0.0333
		CR-FIQA(L) [10]	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	0.0235	**0.0300**
		DifFIQA(R) [4]	0.0251	0.0619	0.0194	0.0262	0.0053	0.0091	0.0020	0.0025	0.0629	0.0688	0.0365	0.0531	0.1847	0.2397	0.0252	0.0369
		eDifFIQA(L) [5]	0.0210	0.0402	0.0148	0.0176	0.0049	0.0083	0.0014	0.0019	0.0574	0.0627	0.0342	0.0500	0.1917	0.2469	**0.0223**	*0.0301*
		GRAFIQS $\mathcal{L}_{\text{BNS}}, \phi_2$ (Our)	0.0225	0.0403	0.0176	0.0219	0.0070	0.0111	0.0032	0.0038	0.0644	0.0692	0.0415	0.0612	0.2058	0.2447	0.0260	0.0346
ElasticFace[9]	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	0.0784	0.1240
		RankIQA[27]	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	0.0707	0.1134
		DeepIQA[8]	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	0.0737	0.1207
	FIQA	RankIQ[11]	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	0.0342	0.0443
		PFE[37]	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	0.0256	0.0372
		SER-FIQ[38]	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*	0.0241	0.0302
		FaceQnet[19, 20]	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	0.0499	0.0801
		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	0.0248	0.0303
		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	0.0283	0.0355
		CR-FIQA(S) [10]	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	0.0230	0.0295
		CR-FIQA(L) [10]	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	*0.0228*	*0.0279*
		DifFIQA(R) [4]	0.0278	0.0536	0.0194	0.0207	0.0050	0.0073	0.0019	0.0025	0.0616	0.0634	0.0330	0.0445	0.1599	0.1890	0.0248	0.0320
		eDifFIQA(L) [5]	0.0222	0.0374	0.0139	0.0148	0.0043	0.0066	0.0014	0.0019	0.0564	0.0576	0.0323	0.0440	0.1688	0.1996	**0.0218**	**0.0271**
		GRAFIQS $\mathcal{L}_{\text{BNS}}, \phi_2$ (Our)	0.0233	0.0394	0.0182	0.0200	0.0070	0.0091	0.0029	0.0037	0.0614	0.0632	0.0393	0.0633	0.1930	0.2319	0.0254	0.0331
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	0.1151	0.1725
		RankIQA[27]	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	0.0878	0.1584
		DeepIQA[8]	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	0.0970	0.1678
	FIQA	RankIQ[11]	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	0.0369	0.0666
		PFE[37]	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	0.0271	0.0499
		SER-FIQ[38]	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*	0.0252	0.0391
		FaceQnet[19, 20]	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	0.0771	0.1234
		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	0.0257	0.0380
		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	0.0301	0.0490
		CR-FIQA(S) [10]	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	*0.0243*	0.0376
		CR-FIQA(L) [10]	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	*0.0243*	**0.0337**
		DifFIQA(R) [4]	0.0256	0.0585	0.0223	0.0363	0.0066	0.0150	0.0020	0.0025	0.0638	0.0660	0.0371	0.0851	0.2177	0.2642	0.0262	0.0439
		eDifFIQA(L) [5]	0.0216	0.0403	0.0168	0.0246	0.0058	0.0121	0.0014	0.0019	0.0580	0.0595	0.0357	0.0810	0.2278	0.2792	**0.0232**	*0.0366*
		GRAFIQS $\mathcal{L}_{\text{BNS}}, \phi_2$ (Our)	0.0233	0.0419	0.0182	0.0253	0.0087	0.0186	0.0033	0.0041	0.0640	0.0652	0.0428	0.0987	0.2524	0.3018	0.0267	0.0423
CurricularFace[22]	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	0.0794	0.1305
		RankIQA[27]	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	0.0707	0.1197
		DeepIQA[8]	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	0.0738	0.1271
	FIQA	RankIQ[11]	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	0.0336	0.0463
		PFE[37]	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	0.0259	0.0399
		SER-FIQ[38]	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*	0.0237	0.0308
		FaceQNet[19, 20]	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	0.0491	0.0870
		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	0.0246	0.0315
		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	0.0285	0.0379
		CR-FIQA(S) [10]	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	*0.0226*	0.0309
		CR-FIQA(L) [10]	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	0.0228	*0.0288*
		DifFIQA(R) [4]	0.0230	0.0475	0.0227	0.0260	0.0055	0.0092	0.0020	0.0025	0.0608	0.0657	0.0305	0.0441	0.1600	0.1871	0.0241	0.0325
		eDifFIQA(L) [5]	0.0199	0.0338	0.0170	0.0195	0.0048	0.0084	0.0014	0.0019	0.0566	0.0601	0.0303	0.0455	0.1717	0.2080	**0.0217**	**0.0282**
		GRAFIQS $\mathcal{L}_{\text{BNS}}, \phi_2$ (Our)	0.0220	0.0365	0.0167	0.0200	0.0068	0.0099	0.0033	0.0038	0.0610	0.0641	0.0369	0.0663	0.1713	0.1959	0.0245	0.0334

ple 100 times into the pretrained FR model designed and trained with dropout layers, our GRAFIQS requires a single forward and backward pass. This work is the pioneer FIQA approach that assesses the quality of face images by inspecting the required changes in pretrained FR model weights to match pretrained FR model data distributions.

5. Conclusion

In this paper, we are the first to introduce a method that assesses the quality of face images by measuring the required changes in the pre-trained FR model weights to minimize dissimilarities between testing samples and the distribution of the FR training dataset. To accomplish this, the differences in BNS between those recorded during the training phase of FR and those acquired through the processing of testing samples using the pre-trained FR model were calculated. Then, the cumulative absolute sum of gradient magnitudes of the pre-trained FR weights was calculated by backpropagating the BNS through the pre-trained FR model. Unlike previous FIQA approaches, our approach did not rely on quality labeling, the necessity to train regression networks, specialized architectures, or the design and optimization of specific training loss functions. These find-



(a) MagFace Model, Adience Dataset (b) ElasticFace Model, Adience Dataset (c) CurricularFace Model, Adience Dataset

(d) MagFace Model, AgeDB30 Dataset (e) ElasticFace Model, AgeDB30 Dataset (f) CurricularFace Model, AgeDB30 Dataset

(g) MagFace Model, CFP-FP Dataset (h) ElasticFace Model, CFP-FP Dataset (i) CurricularFace Model, CFP-FP Dataset

(j) MagFace Model, LFW Dataset (k) ElasticFace Model, LFW Dataset (l) CurricularFace Model, LFW Dataset

(m) MagFace Model, CALFW Dataset (n) ElasticFace Model, CALFW Dataset (o) CurricularFace Model, CALFW Dataset

(p) MagFace Model, CPLFW Dataset (q) ElasticFace Model, CPLFW Dataset (r) CurricularFace Model, CPLFW Dataset

(s) MagFace Model, XQLFW Dataset (t) ElasticFace Model, XQLFW Dataset (u) CurricularFace Model, XQLFW Dataset

Figure 3. EDC curves for FNMR@FMR=$1e-3$ for all evaluated benchmarks using MagFace, ElasticFace, and CurricularFace FR models. AUC are shown in Table 3. EDC curves for ArcFace are provided in the supplementary material. The proposed GRAFIQS method, shown in solid red, utilizes gradient magnitudes and it is reported using the best setting from Table 2.

ings suggest potential directions for future research aimed at assessing FIQ from different perspectives.

Acknowledgments

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

References

[1] ISO/IEC JTC 1/SC 37 Biometrics. ISO/IEC 29794-1 Information technology Biometric sample quality Part 1: Framework. International Organization for Standardization, 2024. 1

[2] Charu C. Aggarwal. *Linear Algebra and Optimization for Machine Learning - A Textbook*. Springer, 2020. 2, 3

[3] Žiga Babnik, Naser Damer, and Vitomir Štruc. Optimization-based improvement of face image quality assessment techniques. In *11th International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*, pages 1–6. IEEE, 2023. 1

[4] Žiga Babnik, Peter Peer, and Vitomir Štruc. Diffiqa: Face image quality assessment using denoising diffusion probabilistic models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023. 2, 4, 5, 6, 7

[5] Žiga Babnik, Peter Peer, and Vitomir Štruc. eDifFIQA: Towards Efficient Face Image Quality Assessment based on Denoising Diffusion Probabilistic Models. *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 2024. 2, 5, 6, 7

[6] 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

[7] Christopher Michael Bishop and Hugh Bishop. *Deep Learning - Foundations and Concepts*. Springer, 1 edition, 2023. 2, 3

[8] 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. 5, 7

[9] 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. 1, 4, 5, 6, 7

[10] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5836–5845. IEEE, 2023. 1, 2, 4, 5, 6, 7

[11] 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. 1, 5, 7

[12] 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. 4, 6, 7

[13] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.*, 9(12):2170–2179, 2014. 4, 6, 7

[14] 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. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1121–1130. IEEE, 2022. 1

[15] 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, 4

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

[17] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In 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

[18] 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

[19] 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, 5, 6, 7

[20] 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. 5, 7

[21] 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. 4, 6, 7

[22] 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. 4, 5, 6, 7

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. 2, 3

[24] 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. 4

[25] 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. 4, 6, 7

[26] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 806–816. IEEE, 2023. 3

[27] 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, 5, 7

[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, 4, 5, 6, 7

[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, 5, 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. 4, 6, 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] Simon J.D. Prince. *Understanding Deep Learning*. MIT Press, 2023. 2, 3

[34] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2488–2498, 2018. 3

[35] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fiérrez, and Christoph Busch. Face image quality assessment: A literature survey. *ACM Comput. Surv.*, 54(10s):210:1–210:49, 2022. 1

[36] 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. 4, 6, 7

[37] 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, 5, 7

[38] 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, 5, 6, 7

[39] Shoukai Xu, Haokun Li, Bohan Zhuang, Jing Liu, Jiezhang Cao, Chuangrun Liang, and Mingkui Tan. Generative low-bitwidth data free quantization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 2020. 2, 3

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

[41] 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. 4, 6, 7

[42] 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. 4, 6, 7