DifFIQA: Face Image Quality Assessment Using Denoising Diffusion Probabilistic Models

Žiga Babnik[1], Peter Peer[2], Vitomir Štruc[1]
[1]University of Ljubljana, Faculty of Electrical Engineering
[2]University of Ljubljana, Faculty of Computer and Information Science
{ziga.babnik, vitomir.struc}@fe.uni-lj.si, peter.peer@fri.uni-lj.si

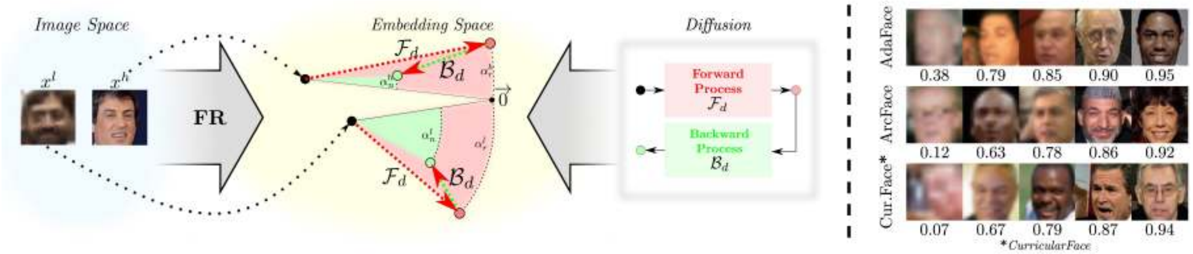

Figure 1. **High-level idea behind the proposed DifFIQA face image quality assessment (FIQA) approach.** The quality of face images corresponds to a considerable degree to the stability of the respective representations in the embedding space of a given face recognition (FR) model. DifFIQA utilizes a diffusion framework to explore the embedding stability through image perturbations caused by the noising and denoising processes. The intuition behind this approach is that the forward (noising) $\mathcal{F}_d$ and backward (denoising) $\mathcal{B}_d$ diffusion processes lead to larger embedding perturbations for lower-quality images (x^l) compared to facial images of higher quality (x^h). By analyzing the impact of both the forward and backward processes on the representation of a given image, DifFIQA is able to infer the corresponding quality and/or generate (FR model specific) quality rankings, as shown on the right. The figure is best viewed electronically.

Abstract

Modern face recognition (FR) models excel in constrained scenarios, but often suffer from decreased performance when deployed in unconstrained (real-world) environments due to uncertainties surrounding the quality of the captured facial data. Face image quality assessment (FIQA) techniques aim to mitigate these performance degradations by providing FR models with sample-quality predictions that can be used to reject low-quality samples and reduce false match errors. However, despite steady improvements, ensuring reliable quality estimates across facial images with diverse characteristics remains challenging. In this paper, we present a powerful new FIQA approach, named DifFIQA, which relies on denoising diffusion probabilistic models (DDPM) and ensures highly competitive results. The main idea behind the approach is to utilize the forward and backward processes of DDPMs to perturb facial images and quantify the impact of these perturbations on the corresponding image embeddings for quality prediction. Because the diffusion-based perturbations are computationally expensive, we also distill the knowledge encoded in DifFIQA into a regression-based quality predictor, called DifFIQA(R), that balances performance and execution time. We evaluate both models in comprehensive experiments on 7 datasets, with 4 target FR models and against 10 state-of-the-art FIQA techniques with highly encouraging results. The source code will be made publicly available.

1. Introduction

State-of-the-art face recognition (FR) models achieve near-perfect results on various benchmarks with high-quality facial images, but still struggle in real-world situations, where the quality of the input samples is frequently unknown [1, 13, 42]. For instance, surveillance, a common application of FR, often involves lower quality samples due to unconstrained and covert capture conditions. In such cases, assessing the quality of the face-image samples is crucial. Low-quality samples can mislead the FR models and cause catastrophic false-match errors, leading to privacy breaches or even monetary loss. By determining the quality of input samples and rejecting or requesting recapture of those below a given quality threshold, the stability and performance of FR models can typically be improved.

Face Image Quality Assessment (FIQA) methods provide FR methods with a quality estimate for each given face sample. In this context, the term *quality* can refer to either the character, fidelity or utility of the sample, as defined by ISO/IEC 29794-1 [22]. Similarly to most FIQA research, we focus on the biometric utility of the facial samples, rather than the visual quality (character and fidelity) as perceived by humans [36]. Such image characteristics are commonly evaluated by general-purpose Image Quality Assessment (IQA) techniques. Biometric utility encompasses several unknown aspects of the given face sample, including its visual quality, face-specific information, and the relative biases inherent to the targeted FR model. It can be interpreted as the usefulness (or fitness) of the sample for the recognition task. Several types of FIQA techniques have been proposed over the years. The largest group focuses on training regression models from calculated pseudo reference quality labels [7, 17, 33, 44], with differences between methods in how they calculate the labels. Other approaches include (unsupervised) analytical methods [3, 29, 40] that use reference-free approaches for quality prediction, and model-based solutions [5,31,39] that combine the face recognition and quality assessment tasks. While modern FIQA techniques have demonstrated impressive performance, providing reliable quality predictions for diverse facial characteristics is still a challenging task.

In this paper, we introduce a novel FIQA technique, called DifFIQA (**Dif**fusion-based **F**ace **I**mage **Q**uality **A**ssessment), that leverages the image-generation versatility of modern Denoising Diffusion Probabilistic Models (DDPMs) for face quality assessment and generalizes well across a wide variety of datasets and FR models. As shown in Figure 1, DifFIQA is based on the following two insights:

- **Perturbation robustness:** Images of higher-quality have stable representations in the embedding space of the given FR model and are less effected by noise perturbations introduced by the forward diffusion process.
- **Reconstruction quality:** High-quality samples are easier to reconstruct from partially corrupted (noisy) data with incomplete identity information and exhibit less disparity between the embeddings of the input and denoised samples than low-quality images.

Based on these observations, DifFIQA analyzes the embedding stability of the input image by perturbing it through the forward as well as backward diffusion process and infers a quality score from the result. To avoid the computationally expensive backward process and speed up computation, we also distill the DifFIQA approach into a regression-based model, termed DifFIQA(R). We evaluate both techniques through extensive experiments over multiple datasets and FR models, and show that both techniques lead to highly competitive results when compared to the state-of-the-art.

2. Related Work

In this section, we briefly review existing FIQA solutions, which can conveniently be partitioned into three main groups: (i) *analytical* techniques, (ii) *regression-based* approaches, and (iii) *model-based* methods.

Analytical methods. The vast majority of methods from this group can be viewed as *specialized* general-purpose IQA techniques that focus on quality predictions defined by (i) selected visual characteristics of faces, such as pose, symmetry or interocular distance, and/or (ii) general visual image properties, such as sharpness, illumination or noise. An early method from this group was presented by Raghavendra *et al.* in [34], where a three stage approach combining pose and image texture components was proposed. Another method by Lijun *et al.* [29] combined several face-image characteristics, including alignment, occlusion and pose, into a pipeline for quality score calculation. Several conceptually similar approaches that exploit different (explicit) visual cues have been presented in the literature over the years [12, 14, 24, 32]. However, the performance of such methods is typically limited, as they focus only on the characteristics of the input samples, with no regard to the utilized FR model. Nevertheless, a new group of analytical methods has recently emerged that incorporates information from both, the input face sample as well as the targeted FR system into the quality estimation process. An example of such an approach was presented by Terhörst *et al.* [40] in the form of the SER-FIQ technique. SER-FIQ calculates a quality score from the embedding variations of a given input face sample, caused by using different configurations of dropout layers. Another method, called FaceQAN by Babnik *et al.* [3], relies on adversarial attacks (which are harder to generate for high quality images) to calculate quality scores. Both of these methods achieve excellent results, but are also comparably computationally demanding, due to their reliance on running several instances of the same sample through the given FR model.

Regression-based methods. FIQA techniques from this group typically train a (quality) regression model using some sort of (pseudo) quality labels. Regression-based methods have received considerable attention over recent years, with most of the research exploring effective mechanisms for generating informative pseudo quality annotations. An early technique from this group, by Best-Rowden and Jain [4], for example, used human raters to annotate the (perceived) quality of facial images, and then trained a quality predictor on the resulting quality labels. Another technique, named FaceQnet [16, 18], relied on embedding comparisons with the highest quality image of each individual to estimate reference quality scores. Here, the highest quality images of each individual were determined using an external quality compliance tool and a ResNet-based regressor was then trained on the extracted quality labels. A more recent approach, called PCNet [44], used a large number of mated image pairs (i.e., a pair of distinct images of the same individual), to train a CNN-based regression model, where the quality labels were defined by the embedding similarity of the mated pairs. The SDD-FIQA approach, by Ou *et al.* [33], extended this concept to also include non-mated (impostor) pairs, (i.e., two unique images of different individuals), where the label for a single image was computed as the Wasserstein distance between the mated and non-mated score distributions. LightQNet, by Chen *et al.* [7], trained a lightweight model, by employing an identification quality loss using quality scores computed from various image comparisons. While regression-based methods in general perform well over a variety of benchmarks and state-of-the-art FR models, their main weakness is the lack of specialization. As the optimal quality estimate for a given input image, is *by definition* FR model specific [2,22], regression-based techniques may require retraining towards the targeted FR model to ensure ideal performance.

Model-based methods. The last group of techniques com-

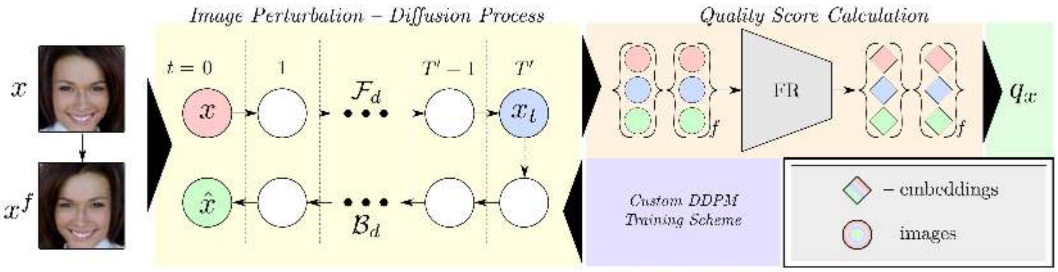

Figure 2. **Overview of DifFIQA.** The proposed quality assessment pipeline consists of two main parts: the *Diffusion Process* and the *Quality-Score Calculation*. The diffusion process uses an encoder-decoder UNet model (D), trained using an extended DDPM training scheme that helps to generate higher-quality (restored) images. The custom DDPM model is used in the Diffusion Process, which generates noisy x_t and reconstructed $\hat{x}$ images using the forward and backward diffusion processes, respectively. To capture the effect of facial pose on the quality estimation procedure, the process is repeated with a horizontally flipped image x^f. The Quality Score Calculation part then produces and compares the embeddings of the original images and the images generated by the diffusion part.

bines face-image quality assessment and face recognition into a single task. One such technique, PFE by Shi and Jain [39], learned to predict a pair of vectors from the input image, i.e., a mean and a variance vector. The mean vector can be thought of as the embedding of the input sample image, while the variance vector represents the sample's variability, and can be used to calculate the sample quality. The presented method inspired several new (uncertainty-aware) methods [8, 28, 46], further improving on the performance of PFE. Another notable technique, called MagFace [31], extended the popular ArcFace loss [10] by incorporating a magnitude-aware angular margin term, which dynamically adjusts class boundaries. The embeddings produced by MagFace encode quality in the magnitude of the embedding itself and can hence be easily inferred. A powerful FIQA technique, called the CR-FIQA, was recently proposed by Boutros *et al*. in [5]. CR-FIQA calculates the quality of the input samples as the ratio between the positive class center and nearest negative class center in a classification task setting, and was demonstrated to produce highly competitive results across various datasets and settings.

Our contribution. The DifFIQA technique, the *main contribution* of this work, can be seen as an analytical method that relies on the capabilities of a DDPM in combination with a chosen FR model. From a conceptual point of view, it is most closely related to FaceQgen [15], a FIQA technique that uses a (GAN-based) generator model for synthesizing high-quality versions of the input samples and the respective discriminator (that aims to distinguish between genuine high-quality images and poorly restored ones) for quality scoring. Unlike FaceQgen, which analyzes the differences between the original and restored images independently of the target FR model, DifFIQA utilizes results from the forward (i.e., noising/degradation) as well as backward (denoising/restoration) diffusion processes and quantifies the embedding variability/uncertainty in the embedding space of a selected FR model for quality estimation. As we show later in the experimental section, this leads to highly competitive FIQA results when compared to the state-of-the-art.

3. Methodology

The stability of the image representations in the embedding space of a given FR model is highly indicative of the input-image quality, as demonstrated by the success of various recent FIQA techniques [3, 40]. One way to explore this stability is by causing perturbations in the image space and analyzing the impact of the perturbations in the embedding space of the targeted FR model. This can, for example, be achieved by using the forward and backward processes of modern diffusion approaches where: the forward process adds some amount of noise to the sample, and the backward process tries to remove the noise, by reconstructing the original. Our main contribution, the DifFIQA technique, takes advantage of the proposed idea, as illustrated in Figure 2, and employs a custom DDPM model for the generation of noisy and reconstructed images. The generated images are then passed through a chosen FR model to explore the impact of the perturbations on the variability of the embedding corresponding to the input image.

3.1. Preliminaries

To make the paper self-contained, we briefly present the main concept behind denoising diffusion probabilistic models (DDPMs), with a focus on their application within our approach. More information on the theoretical background and applications of diffusion models can be found in [9].

In general, DDPMs represent a special type of generative model that learns to model (image) data distributions through two types of processes: a forward (noising) process and backward (denoising) process [9, 23]. The **forward diffusion process** $\mathcal{F}_d$ iteratively adds noise to the given input image x_0, by sampling from a Gaussian distribution $\mathcal{N}(0, I)$. The result of this process is a noisy sample x_t, where t is the number of time steps chosen from the sequence $\{0, 1, \ldots, T\}$. The whole forward process $\mathcal{F}_d$ can be presented as a Markov chain given by

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t|x_{t-1}\sqrt{1-\beta_t}, \beta_t I), \tag{1}$$

where β_t is a variance parameter that defines how much noise is added to the sample at the time instance t of the forward process. By making use of the reparameterization trick [19, 26], any sample x_t can be obtained directly from the input sample x_0, i.e.:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I), \tag{2}$$

where $\overline{\alpha}_t = \prod_{i=0}^{t}(1-\beta_i)$.

The **backward diffusion process** $\mathcal{B}_d$ attempts to iteratively denoise the generated samples x_t, using a deep neural network model D_θ parameterized by θ, according to

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t I), \tag{3}$$

where $t = T, \ldots, 0$, $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, and

$$\mu_\theta(x_t, t) = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha_t}}x_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t. \tag{4}$$

The network is trained to optimize μ_θ, by minimizing the $\mathcal{L}_2$ loss

$$\mathcal{L}_2 = \mathbb{E}_{t,x_0}||D_\theta(x_t, t) - x_0||^2, \tag{5}$$

where $D_\theta(x_t, t)$ is the reconstructed and x_0 the input image. In the remainder of the paper, we drop the subscript θ and use D to denote the deep neural network, which is represented by an unconditional UNet model.

3.2. Overview of DifFIQA

Given a face image x, the goal of DifFIQA is to estimate the quality score $q_x \in \mathbb{R}$, by exploring the effects of the forward and backward diffusion processes of a custom DDPM model D on the image representation e_x in the embedding space of a given FR model M. DifFIQA consists of two main steps, dedicated to: (i) *image perturbation* and (ii) *quality-score calculation*. The image perturbation step uses the forward diffusion process $\mathcal{F}_d$ to create a noisy sample x_t from the input image x and the backward process $\mathcal{B}_d$ to generate the restored (denoised) image $\hat{x}$. In the quality-score calculation step, the representations e_x, e_{x_t}, $e_{\hat{x}}$ corresponding to the input x, noisy x_t and restored image $\hat{x}$, are calculated using the FR model M and then analyzed for disparities to infer the final quality score q_x of the input sample x. To also capture pose-related quality information, DifFIQA repeats the entire process using a horizontally flipped version x^f of the input image x, as also illustrated in Figure 2.

3.3. Extended DDPM Training

To train the DDPM model D (i.e., a UNet [35]) needed by DifFIQA, we extend the standard training process of diffusion models to incorporate time dependent image degradations, as illustrated in Figure 3. These additional (time-dependent) degradations allow the model to learn to gradually reverse these degradations and, in turn, to construct higher quality images during the backward process. Formally, given an input face image $x_0 = x$, the training procedure first constructs a degraded image $x' = d(x)$, where $d(\cdot)$ is some degradation function. A time step $t \in [0, T]$ is then selected for the given sample, from which a time-dependent degraded image is computed as follows:

$$x'_t = Y(x_0, x', t) = (1-\ddot{\alpha}_t)x_0 + \ddot{\alpha}_t x' \tag{6}$$

where $\ddot{\alpha}_t$ is calculated as $\sin(\frac{t}{T} \cdot \frac{\pi}{2})$. The degraded image x'_t is then used to produce the noisy sample x_t using (2). Here, $\ddot{\alpha}_t$ is a time-dependent variable that monotonically increases on the interval $t \in [0, T]$, such that $\ddot{\alpha}_{t=0} = 0$ and $\ddot{\alpha}_{t=T} = 1$. In other words, at time step 0 only the non-degraded image is considered, while at time step T only the degraded image is considered. To implement the degradation function $d(\cdot)$, we use part of the BSRGAN [45] framework that creates a random sequence of image mappings that imitate real-life degradations.

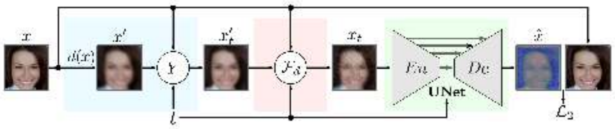

Figure 3. **Presentation of the extended DDPM learning scheme.** Given a training sample x and a time step t, the proposed approach generates a time step dependent degraded image x'_t, by combining the original with a degraded image using the function Y. The image x'_t is then used to generate a noisy sample x_t using the standard forward diffusion approach $\mathcal{F}_d$. A UNet (D) is then trained to reconstruct the input sample in the backward process $\mathcal{B}_d$.

Diffusion models are commonly trained on the full range of time steps $[1, T]$ and learn to generate images from pure noise. However, such a setting is not relevant in the context of quality assessments, as the generated (denoised) images have to exhibit a sufficient correspondence with the input samples x. The easiest solution to this issue is to limit the number of time steps, on which the model is trained $t \in [1, T']$, where $T' < T$, and, in turn, ensure that the noisy image is properly conditioned on the input x. The extended training procedure then minimizes (5) until convergence.

3.4. Generating Noisy and Reconstructed Images

To estimate the quality of a given face image x, DifFIQA makes use of the forward and backward diffusion processes of the trained DDPM. Because head pose is an important factor of face quality, which the underlying DDPM can not explicitly account for, we extend our methodology, by first constructing a horizontally flipped image x^f that we utilize alongside the original image x in the quality-score calculation step, similarly to [3]. The main intuition behind this approach is to exploit the symmetry of human faces, where large deviations from frontal pose induce large disparities between the embeddings of the original and flipped images that can be quantified during quality estimation. Thus, for

Table 1. **Summary of the characteristics of the experimental datasets.** We evaluate DifFIQA across seven diverse datasets with different quality factors and of different size.

Dataset	#Images	#IDs	#Comparisons		Main Quality Factors†‡			
			Mated	Non-mated	Pose	O-E	B-R-N	Sc
Adience [11]	19,370	2,284	20,000	20,000	M	M	L	M
CALFW [48]	12,174	4,025	3,000	3,000	M	M	L	M
CFP-FP [38]	7,000	500	3,500	3,500	H	L	L	M
CPLFW [47]	11,652	3,930	3,000	3,000	H	L	M	M
IJB-C [30]	23,124††	3,531	19,557	15,638,932	H	H	H	Lr
LFW [20]	13,233	5,749	3,000	3,000	L	L	L	M
XQLFW [27]	13,233	5,749	3,000	3,000	L	L	H	M

†O-E - Occlusion, Expression; B-R-N - Blur, Resolution, Noise; Sc - Scale.

‡L - Low; M - Medium; H - High; Lr - Large; Values estimated subjectively by the authors.

†† number of templates, each containing several images

a pair of input face images (x, x^f) and a given time step t, we construct a pair of noisy (x_t, x_t^f) and restored images $(\hat{x}, \hat{x}^f)$ and use the generated data for quality estimation.

3.5. Quality-Score Calculation

DifFIQA relies on the assumption that the embeddings of lower-quality images are more sensitive to image perturbations introduced by the forward and backward diffusion processes than higher-quality images. To quantify this sensitivity, we calculate the average cosine similarity between the embedding of the input image x and all generated noisy and restored counterparts. Additionally, since diffusion models rely on the (random) sampling from a normal distribution, we repeat the whole process n times and average the results, i.e.,

$$q_x = \frac{1}{n|\mathcal{E}|} \sum_{i=1}^{n} \sum_{e_y \in \mathcal{E}} \frac{e_x^T e_y}{\|e_x\| \cdot \|e_y\|}, \tag{7}$$

where $\mathcal{E}$ is a set of generated image embeddings, i.e, $\mathcal{E} = \{e_{x_t}, e_{\hat{x}}, e_{x^f}, e_{x_t^f}, e_{\hat{x}^f}\}$, computed with the FR model M as $e_z = M(x_z)$. In the above equation, the operator $|\cdot|$ denotes the set cardinality and $\|\cdot\|$ the L_2 norm.

3.6. Model Distillation

One of the main shortcomings of DifFIQA (and diffusion models in general) is the high computational complexity compared to other types of FIQA techniques. This complexity stems from the iterative nature of the backward diffusion process, which requires a large number of forward passes through the generative network. Since our approach repeats this process n-times, this only exacerbates the problem and adversely affects the applicability of the technique in real-world applications. To address this problem, we **distill the knowledge** encoded by DifFIQA into a regression model. Specifically, we select a pretrained CosFace FR model augmented with a (quality) regression head and fine-tune it on roughly two million quality labels extracted from the VGGFace2 [6] dataset using the proposed DifFIQA technique. Here, the labels are normalized to $[0, 1]$ and then split into train and validation sets for the training procedure. We refer to the distilled CosFace model as DifFIQA(R) hereafter, and evaluate it together with the original DifFIQA technique in the following sections.

4. Experiments and Results

4.1. Experimental Setup

Experimental setting. We analyze the performance of FaceQDiff in comparison to 10 state-of-the-art FIQA methods, i.e.: (i) the **analytical** FaceQAN [3], SER-FIQ [40], and FaceQgen [15] models, (ii) the **regression-based** FaceQnet [17], SDD-FIQA [33], PCNet [44], and LightQnet [7] techniques, and (iii) the **model-based** MagFace [31], PFE [39], and CR-FIQA [5] methods. We test all methods on 7 commonly used benchmarks with different quality charcteristics, as summarized in Table 1, i.e.: Adience [11], Cross-Age Labeled Faces in the Wild (CALFW) [48], Celebrities in Frontal-Profile in the Wild (CFP-FP) [38], Cross-Pose Labeled Faces in the Wild (CPLFW) [47], large-scale IARPA Janus Benchmark C (IJB-C) [30], Labeled Faces in the Wild (LFW) [20] and the Cross-Quality Labeled Faces in the Wild (XQLFW) [27]. Because the performance of FIQA techniques is dependent on the FR model used, we investigate how well the techniques generalize over 4 state-of-the-art models, i.e.: AdaFace[1] [25], ArcFace[2] [10], CosFace[2] [43], and CurricularFace[3] [21] - all named after their training losses. All FR models use a ResNet100 backbone, and are trained on the WebFace12M[1], MS1MV3[2], Glint360k[2], and CASIA-WebFace[3] datasets.

Evaluation methodology. Following standard evaluation methology [3, 5, 40] and taking recent insights into account [22, 37], we evaluate the performance of DifFIQA using *non-interpolated* Error-versus-Discard Characteristic (EDC) curves (often also referred to as Error-versus-Reject Characteristic or ERC curves in the literature) and the consequent pAUC (partial Area Under the Curve) values. The EDC curves measure the False Non-Match Rate (FNMR), given a predefined False Match Rate (FMR) (10^{-3} in our case), with increasing low-quality image discard (reject) rates. In other words, EDC curves measure how the performance of a given FR model improves when some percentage of the lowest quality images is discarded. Since rejecting a large percentage of all samples is not feasible/practical in real-world application scenarios, we are typically most interested in the performance at the lower discard rates. For this reason we report the pAUC values, where only the results up to a predetermined drop rate threshold are considered. Furthermore, for easier interpretation and comparison of scores over different dataset, we normalize the calculated pAUC values using the FNMR at 0% discard rate, with lower pAUC values indicating better performance.

Implementation Details. During training of the DDPM,

[1] https://github.com/mk-minchul/AdaFace

[2] https://github.com/deepinsight/insightface

[3] https://github.com/HuangYG123/CurricularFace

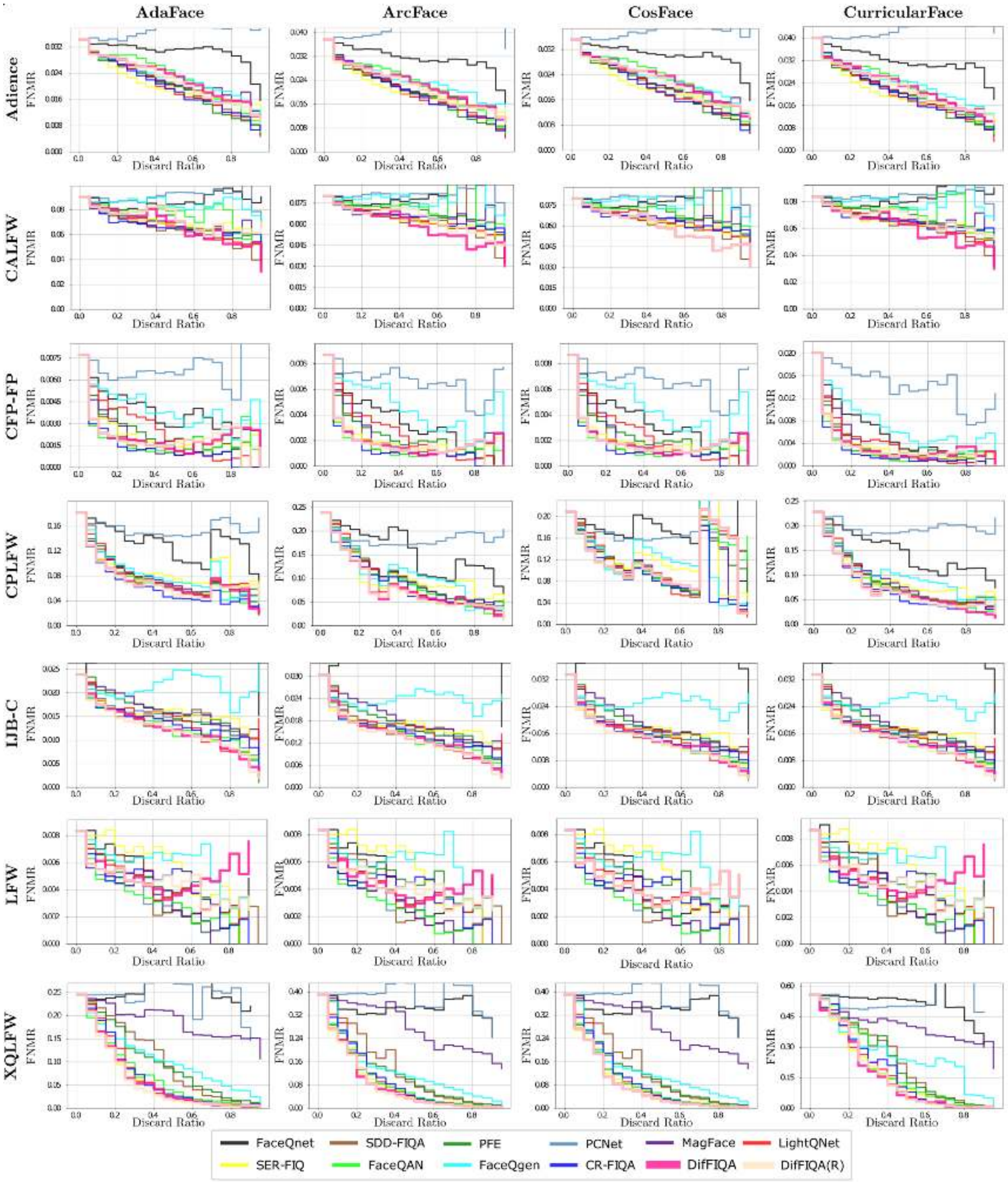

Figure 4. **Comparison to the state-of-the-art in the form of (non-interpolated) EDC curves.** Results are presented for seven diverse datasets, four FR models and in comparison to ten recent FIQA competitors. Observe how the distilled model performs comparably to the non-distilled version, especially at low discard rates. DifFIQA and DifFIQA(R) most convincingly outperform the competitors on the most chalenging IJB-C and XQLFW datasets. The figure is best viewed in color.

the maximum number of forward steps is set to $T = 1000$, yet the underlying model is trained only using up to $T' = 100$ forward diffusion steps. The value of T' does not define the number of time steps t taken at inference time, it only sets the possible upper bound. This process ensures that images produced by the forward process are only partially noisy, so the backward process is properly conditioned on the input image and learns to restore it during training. To account for the randomness introduced by the forward process, we repeat the diffusion process $n = 10$ times and average the results over all iterations, when computing the final quality score. The utilized UNet model (D) consists of four downsampling and upsampling modules, each decreasing (increasing) the dimensions of the representations by a factor of two. Training is done using the Adam optimizer, with a learning rate of $8.0e^{-5}$ in combination with

Table 2. **Comparison to the state-of-the-art.** The table reports pAUC scores at a discard rate of 0.3 and a FMR of 10^{-3}. Average results across all datasets are marked $\overline{\text{pAUC}}$. The best result for each dataset is shown in bold, the overall best result is colored green, the second best blue and the third best red.

AdaFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.963	0.938	0.717	0.887	1.256	0.884	0.977	0.946
SDD-FIQA [33]	0.839	0.871	0.500	0.688	0.782	0.825	0.842	0.764
PFE [39]	0.833	0.890	0.566	0.681	0.868	0.771	0.798	0.772
PCNet [44]	1.005	0.979	0.862	0.898	0.788	0.661	0.987	0.883
MagFace [31]	0.860	0.866	0.524	0.664	0.883	0.666	0.913	0.768
LightQNet [7]	0.847	0.894	0.641	0.684	0.797	0.777	0.704	0.763
SER-FIQ† [41]	**0.807**	0.892	0.475	0.626	0.762	0.935	n/a	0.749
FaceQAN [3]	0.890	0.919	**0.383**	0.619	0.756	**0.656**	0.654	**0.697**
CR-FIQA [5]	0.844	**0.851**	0.391	**0.588**	0.750	0.707	0.684	**0.688**
FaceQgen [15]	0.858	0.970	0.718	0.694	0.853	0.834	0.736	0.809
DifFIQA (ours)	0.864	0.900	0.416	0.608	0.761	0.719	0.627	0.699
DifFIQA(R) (ours)	0.865	0.895	0.412	0.601	**0.731**	0.708	**0.610**	**0.689**

ArcFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.943	0.955	0.693	0.878	1.224	0.884	0.899	0.925
SDD-FIQA [33]	0.783	0.901	0.491	0.734	0.720	0.808	0.774	0.744
PFE [39]	0.774	0.932	0.524	0.738	0.783	0.779	0.641	0.739
PCNet [44]	1.022	1.006	0.868	0.783	0.706	**0.623**	1.004	0.859
MagFace [31]	0.812	0.902	0.549	0.717	0.824	0.635	0.943	0.769
LightQNet [7]	0.789	0.913	0.612	0.752	0.721	0.745	0.621	0.736
SER-FIQ† [40]	**0.767**	0.903	0.416	0.656	0.671	0.935	n/a	0.724
FaceQAN [3]	0.824	0.941	0.373	0.677	0.673	0.624	0.581	**0.670**
CR-FIQA [5]	0.808	**0.891**	**0.358**	0.689	0.664	0.675	0.642	0.675
FaceQgen [15]	0.817	0.985	0.784	0.701	0.785	0.802	0.653	0.789
DifFIQA (ours)	0.805	0.900	0.399	0.647	0.675	0.695	**0.546**	**0.667**
DifFIQA(R) (ours)	0.801	0.898	0.389	**0.646**	**0.655**	0.708	0.554	**0.665**

CosFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.952	0.955	0.693	0.879	1.248	0.884	0.899	0.930
SDD-FIQA [33]	0.825	0.901	0.491	0.735	0.721	0.808	0.774	0.751
PFE [39]	0.813	0.932	0.524	0.748	0.784	0.779	0.641	0.746
PCNet [44]	1.009	1.006	0.868	0.835	0.710	**0.623**	1.004	0.865
MagFace [31]	0.852	0.902	0.549	0.724	0.821	0.635	0.943	0.775
LightQNet [7]	0.835	0.913	0.612	0.753	0.713	0.745	0.621	0.742
SER-FIQ† [40]	**0.793**	0.903	0.416	0.711	0.661	0.935	n/a	0.736
FaceQAN [3]	0.871	0.941	0.373	**0.667**	0.675	0.624	0.581	**0.676**
CR-FIQA [5]	0.835	**0.891**	**0.358**	0.681	0.664	0.675	0.642	0.678
FaceQgen [15]	0.847	0.985	0.784	0.702	0.783	0.802	0.653	0.794
DifFIQA (ours)	0.841	0.900	0.399	0.669	0.672	0.695	**0.546**	**0.675**
DifFIQA(R) (ours)	0.838	0.900	0.389	0.669	**0.644**	0.695	**0.546**	**0.669**

CurricularFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.921	0.947	0.601	0.867	1.248	0.908	0.984	0.925
SDD-FIQA [33]	0.776	0.900	0.409	0.696	0.721	0.821	0.817	0.734
PFE [39]	0.759	0.923	0.415	0.691	0.784	0.785	0.835	0.742
PCNet [44]	1.004	0.996	0.887	0.899	0.710	0.656	0.938	0.870
MagFace [31]	0.793	0.892	0.477	0.689	0.821	0.661	0.862	0.742
LightQNet [7]	0.769	0.910	0.462	0.704	0.713	0.767	0.739	0.723
SER-FIQ† [40]	**0.750**	0.883	0.389	0.625	0.661	0.942	n/a	0.708
FaceQAN [3]	0.811	0.931	0.343	0.637	0.669	**0.644**	0.835	0.696
CR-FIQA [5]	0.797	**0.877**	**0.318**	**0.615**	0.664	0.693	0.789	**0.679**
FaceQgen [15]	0.815	0.974	0.662	0.698	0.783	0.845	0.750	0.790
DifFIQA (ours)	0.806	0.884	0.384	0.624	0.672	0.711	**0.736**	**0.688**
DifFIQA(R) (ours)	0.788	0.892	0.358	0.622	**0.644**	0.724	0.768	**0.685**

†SER-FIQ was used to create XQLFW, so the results here are not reported for a fair comparison.

an Exponential Moving Average (EMA) model, with a decay rate of 0.995. The presented hyperparameters were determined through preliminary experiments on hold-out data to ensure a reasonable trade-off between training speed and reproducible performance. All experiments were conducted on a desktop PC with an Intel i9-10900KF CPU, 64 GB of RAM and an Nvidia 3090 GPU.

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

In this section, we compare DifFIQA and the distilled version, DifFIQA(R), with ten state-of-the-art competitors and analyze: (i) the *performance characteristics* of the tested techniques, and (ii) their *runtime complexity*.

Performance analysis. In Figure 4, we show the (non-interpolated) EDC curves for all tested FR models and datasets, and report the corresponding pAUC scores in Table 5. Following the suggestions in [22, 37], we chose a discard rate of 0.3, when calculating the pAUC values, but also report additional results in the supplementary material. We observe that the proposed diffusion-based FIQA techniques result in highly competitive performance across all datasets and FR models. The distilled DifFIQA(R) model, for example, leads to the lowest average $\overline{\text{pAUC}}$ score with the ArcFace and CosFace FR models, and is the runner-up with the AdaFace and CurricularFace models with $\overline{\text{pAUC}}$ scores comparable to the top performer CR-FIQA[4]. Several interesting findings can be made from the reported results, e.g.: (i) While the performance of DifFIQA and DifFIQA(R) is in general close, the distilled version has a slight edge over the original, which suggests that the distillation process infuses some additional information into the FIQA procedure through the FR-based regression model; (ii) The proposed FIQA models are particularly competitive on the difficult large-scale IJB-C dataset, where the DifFIQA(R) approach consistently outperforms all competing baseline models. A similar observation can also be made for the challenging XQLFW dataset, where the diffusion-based models are again the top performers, which speaks of the effectiveness of diffusion-based quality estimation.

Runtime complexity. In Table 3, we compare the runtime complexity of the evaluated FIQA techniques (in ms). To ensure a fair comparison, we utilize (i) the same experimental hardware for all methods (described in Section 4.1), (ii) use the official code, released by the authors for all techniques, and (iii) compute average runtimes and standard deviations over the entire XQLFW dataset. As can be seen, the original approach, DifFIQA, despite being highly competitive in terms of performance, is among the most computationally demanding due to the use of the complex diffusion processes. With around $1s$ on average per image, the runtime complexity of the model is even significantly higher than that of the FaceQAN or SER-FIQ techniques that require multiple passes through their networks to estimate quality and which are already among the slower FIQA models. However, the distillation process, allows to reduce the runtime by roughly three orders of magnitude (or by $99,9\%$), making the distilled DifFIQA(R) comparable to the faster models evaluated in this experiment.

4.3. Ablation Study

We perform several ablation studies to explore the impact of the main components of DifFIQA. Specifically, we are interested in: (A1) the impact of the flipping procedure, utilized to capture pose-related quality factors, (A2) the contribution of the forward pass (i.e., the noising step of the diffusion), and (A3) the impact of the number of forward diffusion steps t, where a larger number corresponds to higher amounts of noise in the image x_t produced by the

[4] In the supplementary material we show that with a discard rate of 0.2, DifFIQA(R) is the top performer with 3 of the 4 FR models.

Table 3. **Runtime complexity.** The reported results (in ms) were computed over the XWLFW dataset and the same experimental hardware. Note how the destillation process leads to a speed-up of more than three orders of magnitude from DifFIQA to DifFIQA(R).

FIQA model	**CR-FIQA** [5]	**SDD-FIQA** [33]	**FaceQAN** [3]	**MagFace** [31]	**SER-FIQ** [40]	**FaceQnet** [17]	**FaceQgen** [15]	**LightQnet** [7]	**PCNet** [44]	**PFE** [39]	**Ours**	
											DifFIQA	**DifFIQA(R)**
Runtime ($\mu \pm \sigma$)	0.15 ± 0.37	0.62 ± 0.36	334.13 ± 118.79	1.08 ± 0.36	112.93 ± 33.81	42.11 ± 2.14	42.11 ± 2.05	18.54 ± 18.68	17.06 ± 0.34	42.69 ± 12.26	1074.62 ± 11.45	1.24 ± 0.36

Table 4. **Results of the ablation study.** The results are reported in terms of pAUC (↓) at a FMR of 10^{-3} and a discard rate of 0.3.

Model variant	**LFW**	**CPLFW**	**CALFW**	**XQLFW**	$\overline{\text{pAUC}}$
(A1): w/o Image Flipping	0.702	0.727	0.888	0.535	0.713
(A2): w/o Forward Pass	0.730	0.684	0.897	0.531	0.710
(A3): DifFIQA ($t = 20$)	0.657	0.694	0.945	0.628	0.731
DifFIQA (complete)	0.695	0.669	0.900	0.546	**0.702**

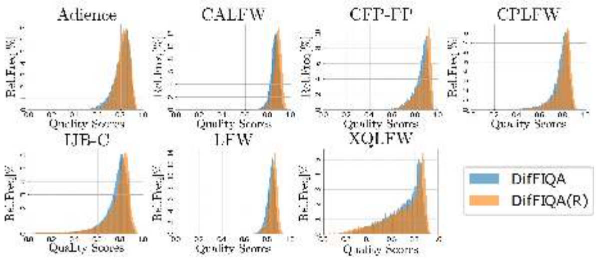

Figure 5. **Quality-score distributions.** DifFIQA by DifFIQA(R) produce very consistent distributions over all seven test datasets.

forward process. Because the ablations are only relevant for the (non-distilled) approach, we experiment solely with the DifFIQA technique and report results using the CosFace FR model and four datasets that feature a broad range of quality characteritics, i.e., LFW, CPLFW, CALFW and XQLFW.

From the results in Table 4, we observe that the exclusion of the flipping operation significantly degrades performance on the cross-pose (CPLFW) dataset, while contributing to minor improvements on CALFW and XQLFW. However, given that pose is considered one of the main factors still adversely affecting modern FR models, the flipping operation still helps with the performance across all the test datasets (see average A1 results). When removing the forward pass (in A2), we again see considerable performance drops on LFW and CPLFW, leading to lower average pAUC scores. This suggest that both (forward and backward) processes are important for good results across different datasets. Finally, we see that lesser amounts of noise and stronger conditioning on the input images leads to better results as evidenced by the A3 results with our model with 20 timesteps, instead of the 5 utilized in the complete DifFIQA approach.

4.4. Qualitative Evaluation

While the proposed DifFIQA approach has a sound theoretical basis that links the forward and backward diffusion processes to face image quality, the distilled variant abstracts this relation away and approaches the FIQA task from a pure learning perspective. To get better insight into the characteristics of both models, we investigate in this section their behavior in a qualitative manner.

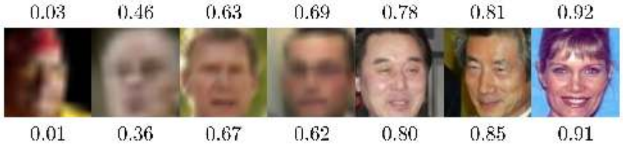

Figure 6. **Illustration of the quality scores produced by the proposed FIQA techniques.** The scores on the top shows results for DifFIQA and the scores at the bottom for DifFIQA(R). While the concrete scores differ, both models generate similar rankings.

Quality-score distributions. In Figure 5, we compare the quality-score distributions, generated by DifFIQA and DifFIQA(R) on all seven test datasets. As can be seen, the two models produce very similar distributions, with a slight preference of DifFIQA(R) towards higher quality scores.

Visual analysis. In Figure 6, we show example images from the XQLFW dataset and the corresponding quality scores, generated by the DifFIQA and DifFIQA(R) techniques. Note that both approaches produce a similar ranking but differ in the concrete quality score assigned to a given image. It is interesting to see that some blurry images with low (human-perceived) visual quality receive relatively high quality scores, as they feature frontal faces that may still be useful for recognition purposes. Additional qualitative results that illustrate the capabilties of the DifFIQA model are also shown on the right part of Figure 1.

5. Conclusion

We have presented a novel approach to face image quality assessment (FIQA), called DifFIQA, that uses denoising diffusion probabilistic models as the basis for quality estimation. Through comprehensive experiments on multiple datasets we showed that the proposed model yields highly competitive results, when benchmarked against state-of-the-art techniques from the literature and that the runtime performance can be reduced significantly if the model is distilled into a quality predictor through a regression-based procedure. As part of our future work, we plan to investigate extensions to our model, including transformer-based UNet alternatives and latent diffusion processes.

References

[1] Shahina Anwarul and Susheela Dahiya. A Comprehensive Review on Face Recognition Methods and Factors Affecting Facial Recognition Accuracy. *Proceedings of International Conference on Recent Innovations in Computing (ICRIC)*, pages 495–514, 2020.

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

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

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

[5] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A Dataset for Recognising Faces Across Pose and Age. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018.

[7] Kai Chen, Taihe Yi, and Qi Lv. LightQNet: Lightweight Deep Face Quality Assessment for Risk-Controlled Face Recognition. *Signal Processing Letters*, 28:1878–1882, 2021.

[8] Kai Chen, Taihe Yi, and Qi Lv. Fast and Reliable Probabilistic Face Embeddings Based on Constrained Data Uncertainty Estimation. *Image and Vision Computing (IVC)*, page 104429, 2022.

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A survey. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

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

[11] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and Gender Estimation of Unfiltered Faces. *Transactions on Information Forensics and Security (TIFS)*, 9(12):2170–2179, 2014.

[12] Xiufeng Gao, Stan Z Li, Rong Liu, and Peiren Zhang. Standardization of Face Image Sample Quality. In *Proceedings of the IAPR International Conference on Biometrics (ICB)*, pages 242–251. Springer, 2007.

[13] Klemen Grm, Vitomir Štruc, Anais Artiges, Matthieu Caron, and Hazım K Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biometrics*, 7(1):81–89, 2018.

[14] Olaf Henniger, Biying Fu, and Cong Chen. On the Assessment of Face Image Quality Based on Handcrafted Features. In *Proceedings of the IEEE International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2020.

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

[16] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric Quality: Review and Application to Face Recognition with FaceQnet. *arXiv preprint arXiv:2006.03298*, 2020.

[17] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. FaceQnet: Quality Assessment for Face Recognition Based on Deep Learning. In *Proceedings of the IAPR International Conference on Biometrics (ICB)*, pages 1–8, 2019.

[18] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality Assessment for Face Recognition Based on Deep Learning. In *Proceedings of the IEEE International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

[20] 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.

[21] 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 *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020.

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

[23] Marija Ivanovska and Vitomir Štruc. Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models. In *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2023.

[24] Hyung-Il Kim, Seung Ho Lee, and Yong Man Ro. Investigating Cascaded Face Quality Assessment for Practical Face Recognition System. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, pages 399–400, 2014.

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

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

[28] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical Confidence Learning for Face Recognition. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15629–15637, 2021.

[29] Zhang Lijun, Shao Xiaohu, Yang Fei, Deng Pingling, Zhou Xiangdong, and Shi Yu. Multi-Branch Face Quality Assessment for Face Recognition. In *Proceedings of the*

IEEE International Conference on Communication Technology (ICCT), pages 1659–1664, 2019.

[30] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. IARPA Janus Benchmark-C: Face Dataset and Protocol. In *Proceedings of the International Conference on Biometrics (ICB)*, pages 158–165, 2018.

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

[32] Kamal Nasrollahi and Thomas B Moeslund. Extracting a Good Quality Frontal Face Image From a Low-Resolution Video Sequence. *Transactions on Circuits and Systems for Video Technology (TCSVT)*, 21(10):1353–1362, 2011.

[33] 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 *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021.

[34] Ramachandra Raghavendra, Kiran B Raja, Bian Yang, and Christoph Busch. Automatic Face Quality Assessment From Video Using Gray Level Co-Occurrence Matrix: An Empirical Study on Automatic Border Control System. In *Proceedings of the IAPR International Conference on Pattern Recognition (ICPR)*, pages 438–443, 2014.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[36] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. Face Image Quality Assessment: A Literature Survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–49, 2022.

[37] Torsten Schlett, Christian Rathgeb, Juan Tapia, and Christoph Busch. Considerations on the Evaluation of Biometric Quality Assessment Algorithms. *arXiv preprint arXiv:2303.13294*, 2023.

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

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

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

[41] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: Unsupervised Estimation of Face Image Quality Based on Stochastic Embedding Robustness. In *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5651–5660, 2020.

[42] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *Transactions on Technology and Society (TTS)*, 3(1):16–30, 2021.

[43] 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 *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018.

[44] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing Predictive Uncertainty Estimation for Face Verification. In *British Machine Vision Conference (BMVC)*, 2020.

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

[46] Yuying Zhao and Weihong Deng. Dual Gaussian Modeling for Deep Face Embeddings. *Pattern Recognition Letters*, 161:74–81, 2022.

[47] 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.

[48] 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.

6. Supplementary Material

In the main part of the paper, we evaluated the proposed DifFIQA technique in comprehensive experiments across 7 diverse datasets, in comparison to 10 state-of-the-art (SOTA) competitors, and with 4 different face recognition models. In this supplementary material, we now show additional results using the same setup as in the main part of the paper that: (1) illustrate the performance of the model at another discard rate, (2) show the average performance of the proposed approach across all datasets and FR models and in comparison to all considered SOTA techniques for two different discard rates, and (3) provide details on the runtime complexity of the DifFIQA model. Additionally, we also discuss the limitation of the proposed FIQA models and provide information on the reproducibility of the experiments described in the main part of the paper.

6.1. Additional Results

Comparison to SOTA techniques. In Table 5, we present additional comparisons to the ten state-of-the-art techniques already considered in the main part of the paper. However, here the results are reported for a lower drop rate of 0.2. We note again that the performance of FIQA techniques is most relevant at lower drop rates, since this facilitates real-world applications, as also emphasized in [37].

From the presented results, we observe that the distilled model, DifFIQA(R) yields the lowest average pAUC scores (computed over the seven test datasets), when used with the AdaFace, ArcFace and CosFace models. With the CurricularFace model, DifFIQA(R) is the runner-up with performance close to the best performing CR-FIQA technique. It is worth noting that among the tested methods, four FIQA techniques performed significantly better than the rest across the four different FR models, i.e., CR-FIQA [5], FaceQAN [3] and the two diffusion-based models proposed in this paper, DifFIQA and DifFIQA(R). However, the distilled DifFIQA(R) technique is overall the top performer and fares particularly well on the most challenging datasets considered in the experiments, i.e., IJB-C and XQLFW.

Overall performance. To further illustrate the performance of the proposed DifFIQA and DifFIQA(R) techniques, we present in Tables 6 and 7 the average pAUC scores for two discard rates (0.2 and 0.3), computed over the seven test datasets and all four considered FR models. The reported results again support the findings already made above, i.e., FaceQAN, CR-FIQA, and our proposed techniques significantly outperform all other FIQA techniques, while DifFIQA(R) performs overall the best.

Table 5. **Comparison to the state-of-the-art.** The table reports pAUC scores at a discard rate of 0.2 and a FMR of 10^{-3}. Average results across all datasets are marked $\overline{\text{pAUC}}$. The best result for each dataset is shown in bold, the overall best result is colored green, the second-best blue and the third-best red.

AdaFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.969	0.960	0.772	0.935	1.133	0.934	0.969	0.953
SDD-FIQA [33]	0.884	0.911	0.632	0.789	0.854	0.857	0.907	0.833
PFE [39]	0.873	0.917	0.659	0.772	0.918	0.854	0.885	0.840
PCNet [44]	1.003	0.985	0.893	0.926	0.843	0.730	0.999	0.911
MagFace [31]	0.890	0.900	0.632	0.747	0.915	0.735	0.958	0.825
LightQNet [7]	0.890	0.925	0.711	0.784	0.846	0.837	0.836	0.833
SER-FIQ [40]	**0.871**	0.930	0.563	0.715	0.812	0.982	n/a	0.812
FaceQAN [3]	0.905	0.942	**0.474**	0.700	0.800	**0.721**	0.764	**0.758**
CR-FIQA [5]	0.890	**0.887**	0.504	**0.684**	0.796	0.755	0.830	**0.764**
FaceQgen [15]	0.889	0.967	0.774	0.778	0.877	0.887	0.814	0.855
DifFIQA	0.897	0.932	0.500	0.698	0.813	0.770	0.769	0.768
DifFIQA(R)	0.893	0.913	0.505	0.696	**0.796**	0.752	**0.754**	**0.758**

ArcFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.957	0.970	0.761	0.918	1.123	0.934	0.933	0.942
SDD-FIQA [33]	0.841	0.931	0.637	0.829	0.806	0.857	0.874	0.825
PFE [39]	**0.823**	0.943	0.624	0.833	0.844	0.854	0.746	0.810
PCNet [44]	1.013	0.998	0.910	0.809	0.770	**0.697**	1.003	0.886
MagFace [31]	0.852	0.925	0.683	0.809	0.867	0.712	0.961	0.830
LightQNet [7]	0.840	0.930	0.706	0.857	0.788	0.814	0.772	0.816
SER-FIQ [40]	0.840	0.934	0.508	0.797	0.732	0.982	n/a	0.798
FaceQAN [3]	0.850	0.957	**0.470**	0.771	0.731	0.699	0.710	**0.741**
CR-FIQA [5]	0.861	**0.912**	0.475	0.791	**0.724**	0.732	0.764	0.751
FaceQgen [15]	0.857	0.980	0.823	0.834	0.823	0.865	0.786	0.853
DifFIQA	0.848	0.931	0.493	**0.771**	0.743	0.759	0.696	**0.749**
DifFIQA(R)	0.840	0.920	0.484	0.772	0.732	0.752	**0.688**	**0.741**

CosFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.962	0.970	0.761	0.917	1.139	0.934	0.933	0.945
SDD-FIQA [33]	0.873	0.931	0.637	0.832	0.806	0.857	0.874	0.830
PFE [39]	**0.856**	0.943	0.624	0.837	0.848	0.854	0.746	0.816
PCNet [44]	1.005	0.998	0.910	0.861	0.776	**0.697**	1.003	0.893
MagFace [31]	0.882	0.925	0.683	0.808	0.875	0.712	0.961	0.835
LightQNet [7]	0.880	0.930	0.706	0.855	0.787	0.814	0.772	0.821
SER-FIQ [40]	0.863	0.934	0.508	0.790	0.725	0.982	n/a	0.800
FaceQAN [3]	0.890	0.957	**0.470**	0.759	0.741	0.699	0.710	**0.747**
CR-FIQA [5]	0.884	**0.912**	0.475	0.778	0.734	0.732	0.764	0.754
FaceQgen [15]	0.880	0.980	0.823	0.821	0.824	0.865	0.786	0.854
DifFIQA	0.881	0.931	0.493	**0.758**	0.738	0.759	**0.696**	**0.751**
DifFIQA(R)	0.870	0.931	0.484	0.758	**0.723**	0.759	**0.696**	**0.746**

CurricularFace - pAUC@FMR=10^{-3} (↓)

FIQA model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{\text{pAUC}}$
FaceQnet [17]	0.941	0.964	0.692	0.914	1.139	0.960	0.990	0.943
SDD-FIQA [33]	0.838	0.932	0.556	0.802	0.806	0.865	0.867	0.810
PFE [39]	**0.815**	0.937	0.539	0.793	0.848	0.863	0.900	0.814
PCNet [44]	1.000	0.993	0.931	0.938	0.776	0.732	0.971	0.906
MagFace [31]	0.841	0.921	0.624	0.779	0.875	0.736	0.901	0.811
LightQNet [7]	0.827	0.938	0.574	0.815	0.787	0.834	0.857	0.805
SER-FIQ [40]	0.832	0.926	0.493	0.747	0.725	0.986	n/a	0.784
FaceQAN [3]	0.843	0.948	0.453	0.736	0.730	**0.713**	0.908	**0.762**
CR-FIQA [5]	0.859	**0.908**	**0.428**	**0.729**	0.734	0.746	0.902	**0.758**
FaceQgen [15]	0.858	0.972	0.754	0.806	0.824	0.894	0.836	0.849
DifFIQA	0.851	0.919	0.499	0.738	0.738	0.771	0.863	0.768
DifFIQA(R)	0.832	0.922	0.467	0.740	**0.723**	0.764	0.883	**0.762**

[†] SER-FIQ was used to create XQLFW, so the results here are not reported for a fair comparison.

Runtime complexity. In the main part of the paper, we analyzed and tested all considered techniques from a runtime-performance perspective. Here, we explore the runtime complexity of DifFIQA in more detail to get better insight into the computationally most demanding steps of the approach. The whole method includes five steps: the initialization step (i), which creates all the necessary image copies and converts them into tensors, the forward diffusion step (f), the backward diffusion step (b), the image embedding step (fr), and the quality score calculation step (q). As can be seen from the reported results in Table 8, DifFIQA takes 1074ms on average to estimate the quality of a single face image. Recall, that the distilled approach requires only around 1ms for the same task. By far the most demanding part of the quality estimation procedure is the backward diffusion process, which iteratively denoises the given images, with an average time of a little more than 840ms. Even though we use only 5 iterations, we create for a single image 10 noisy copies of the original and the flipped version. All of these images are then passed through the denoising network, which accounts for the high time complexity of the backward process. The generation of image embeddings also requires some time, i.e., 66ms, as the step encapsulates

Table 6. **Average performance over all seven test datasets and four FR models at a drop rate of 0.2.** The results are reported in terms of average pAUC score at the FMR of 10^{-3}. The proposed DifFIQA(R) approach is overall the best performer. The best result is colored green, the second-best blue and the third-best red.

FaceQnet [17]	SDD-FIQA [33]	PFE [39]	PCNet [44]	MagFace [31]	LightQNet [7]	SER-FIQ [40]	FaceQAN [3]	CR-FIQA [5]	FaceQgen [15]	DifFIQA	DifFIQA(R)
0.9458	0.8244	0.8197	0.8989	0.8253	0.8183	0.7985	**0.7519**	**0.7567**	0.8527	0.7591	**0.7518**

Table 7. **Average performance over all seven test datasets and four FR models at a drop rate of 0.3.** The results are reported in terms of average pAUC score at the FMR of 10^{-3}. The proposed DifFIQA(R) approach is overall the best performer. The best result is colored green, the second-best blue and the third-best red.

FaceQnet [17]	SDD-FIQA [33]	PFE [39]	PCNet [44]	MagFace [31]	LightQNet [7]	SER-FIQ [40]	FaceQAN [3]	CR-FIQA [5]	FaceQgen [15]	DifFIQA	DifFIQA(R)
0.9315	0.7483	0.7497	0.8691	0.7635	0.7412	0.7292	0.6847	**0.6800**	0.7954	**0.6822**	**0.6768**

Table 8. **Detailed analysis of the runtime performance of DifFIQA in ms.** The reported results were computed over the entire XQLFW dataset and for each component of the model separately. For DifFIQA the times are presented separately for the initialization t_i, the forward process t_f, the backward process t_b, embedding of the images t_{fr}, and the quality calculation t_q steps. The symbol Σ denotes the overall runtime.

Model component runtime	t_i	t_f	t_b	t_{fr}	t_q	Σ
Runtime in ms ($\mu \pm \sigma$)	0.166 ± 0.006	0.192 ± 0.010	842.041 ± 9.068	66.224 ± 0.689	166.335 ± 1.750	1074.627 ± 11.458

the collection of all starting, noisy and reconstructed images into a single tensor as well as the forward pass through the FR model. In total, the image embedding steps need to produce embeddings for 60 images, all constructed from the given input sample. The score computation also takes close to 170ms, because it includes the calculation of five separate cosine similarities for all image copies, calculation of the average value over all copies and the data transfer from VRAM to RAM.

6.2. Limitations

The proposed DDPM-based DifFIQA technique ensure highly competitive FIQA performance, but also has some **limitations**. One obvious limitation is the computational complexity that affects the model's runtime performance, as emphasized throughout the paper. While this can be addressed through a distillation procedure, the distillation process removes the relation between the (noising and denoising) tasks and image quality, and consequently impacts the interpretability of the results. From a conceptual point of view, the nosing and denoising steps probe the quality of the facial images by (in a sense) first obscuring important facial features and then measuring the ability to restore the obscured features through denoising. Such restoration-based solutions may depend, to a significant degree, on the restoration model utilized, which in our case is a CNN-based UNet that implements the denoising diffusion. While such models are known to be able to capture local image characteristics very well, they may be less capable in capturing key global image properties, and we plan to explore transformer-based models in our future work to further improve on this limitation.

6.3. Reproduciblity

We would like to note that all of our experiments are fully reproducible. Most of the models used for the implementation and testing of DifFIQA and DifFIQA(R) are publicly available from the official repositories, while all others can be obtained by request from the authors, i.e.:

- AdaFace:
  https://github.com/mk-minchul/AdaFace
- ArcFace:
  https://github.com/deepinsight/insightface
- CosFace:
  https://github.com/deepinsight/insightface
- CurricularFace:
  https://github.com/HuangYG123/CurricularFace
- FaceQnet:
  https://github.com/javier-hernandezo/FaceQnet
- SDD-FIQA:
  https://github.com/Tencent/TFace/tree/quality
- PFE:
  https://github.com/seasonSH/Probabilistic-Face-Embeddings
- PCNet:
  Requested from authors
- MagFace:
  https://github.com/IrvingMeng/MagFace
- LightQNet:
  https://github.com/KaenChan/lightqnet

- SER-FIQ:
  https://github.com/pterhoer/FaceImageQuality
- FaceQAN:
  https://github.com/LSIbabnikz/FaceQAN
- FaceQgen:
  https://github.com/javier-hernandezo/FaceQgen
- CR-FIQA:
  https://github.com/fdbtrs/CR-FIQA
- Diffusion models:
  https://github.com/lucidrains/denoising-diffusion-pytorch

Additionally, we also plan to publicly release the DifFIQA source code, including all training and testing scripts, model design and learned weights, once the review procedure is completed.