eDifFIQA: Towards Efficient Face Image Quality Assessment based on Denoising Diffusion Probabilistic Models

Žiga Babnik, *Member, IEEE,* Peter Peer, *Senior Member, IEEE,* and Vitomir Štruc, *Senior Member, IEEE*

Abstract—State-of-the-art Face Recognition (FR) models perform well in constrained scenarios, but frequently fail in difficult real-world scenarios, when no quality guarantees can be made for face samples. For this reason, Face Image Quality Assessment (FIQA) techniques are often used by FR systems, to provide quality estimates of captured face samples. The quality estimate provided by FIQA techniques can be used by the FR system to reject samples of low-quality, in turn improving the performance of the system and reducing the number of critical false-match errors. However, despite steady improvements, ensuring a good trade-off between the performance and computational complexity of FIQA methods across diverse face samples remains challenging. In this paper, we present DifFIQA, a powerful unsupervised approach for quality assessment based on the popular denoising diffusion probabilistic models (DDPMs) and the extended (eDifFIQA) approach. The main idea of the base DifFIQA 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 of the iterative nature of DDPMs the base DifFIQA approach is extremely computationally expensive. Using eDifFIQA we are able to improve on both the performance and computational complexity of the base DifFIQA approach, by employing label optimized knowledge distillation. In this process, quality information inferred by DifFIQA is distilled into a quality-regression model. During the distillation process we use an additional source of quality information hidden in the relative position of the embedding to further improve the predictive capabilities of the underlying regression model. By choosing different feature extraction backbone models as the basis for the quality-regression eDifFIQA model, we are able to control the trade-off between the predictive capabilities and computational complexity of the final model. We evaluate three eDifFIQA variants of varying sizes in comprehensive experiments on 7 diverse datasets containing static-images and a separate video-based dataset, with 4 target CNN-based FR models and 2 target Transformer-based FR models and against 10 state-of-the-art FIQA techniques, as well as against the initial DifFIQA baseline and a simple regression-based predictor DifFIQA(R), distilled from DifFIQA without any additional optimization. The results show that the proposed label optimized knowledge distillation improves on the performance and computationally complexity of the base DifFIQA approach, and is able to achieve state-of-the-art performance in several distinct experimental scenarios. Furthermore, we also show that the distilled model can be used directly for face recognition and leads to highly competitive results.

Index Terms—Computer Vision, Face Recognition, Face Image Quality Assessment, Denoising Diffusion Probabilistic Models, Knowledge Distillation, Label Optimization

1 Introduction

The performance of Face Recognition (FR) system has improved significantly over the recent years, with state-of-the-art models achieving near-perfect results on various benchmarks with high- to medium-quality images. However, this kind of performance does not always carry-over to more challenging real-world scenarios, such as surveillance or mobile applications [1], [2], [3], where the image quality cannot be controlled, and the recognition models are often confronted with lower quality images. Such low-quality samples have an adverse effect on the performance of FR models and can cause catastrophic false-match errors, leading to privacy breaches or even monetary loss. A common solution to such challenges, is to estimate the quality of the input face images and reject or request recapture of those below a given quality threshold. Through this procedure, the stability, and performance of FR models can typically be significantly improved [4], [5], [6].

• *Žiga Babnik, Vitomir Štruc and Peter Peer are with the Faculty of Electrical Engineering and the Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia (email: ziga.babnik@fe.uni-lj.si)*

Face Image Quality Assessment (FIQA) approaches are designed to provide FR systems with estimates of the (biometric) quality of the input face images. Here, the term quality is commonly defined as the *utility* of the given sample for face recognition, as also described in the ISO/IEC 29794-1 standard [7]. In other words, FIQA techniques do not focus explicitly on assessing sample quality from the standpoint of human-perception. Instead, they aim to quantify quality "through the eyes" of a FR model and capture all image characteristics that can in some way impact the face-recognition performance. Since visual image characteristics, such as noise, lighting, and occlusions are known to negatively affect FR models, such human-perceived quality factors often exhibit a significant correlation with the utility of the images for face recognition. As such, Face Image Quality (FIQ) can also be seen as an extension of the perceivable (visual) quality that encodes additional FR model specific/relevant information.

While general-purpose Image Quality Assessment (IQA) techniques can be used for the assessment of face-image quality as well [8], [9], their performance is often unsatisfactory and far behind the current state-of-the-art (SOTA) in the FIQA area. Modern FIQA techniques, on the other hand, have been shown to perform well across facial images with diverse quality characteristics and in conjunction with a wide variety of conceptually distinct face recognition models. Several different groups of techniques have been proposed in the literature over the years. The largest of these groups consists of *supervised FIQA techniques* that aim to learn quality predictors by regressing to so-called pseudo-quality labels [10], [11], [12], [13]. Here, the pseudo-quality labels are defined in various ways, ranging from human annotations to labels derived through face recognition experiments. Another group of techniques focuses on *unsupervised face image quality assessment.* Techniques from this group commonly quantify specific image characteristics known to affect face recognition systems without relying on explicit supervision [5], [14], [15], [16], [17]. The last noteworthy group of techniques combines face recognition and face image quality assessment into one coherent task [18], [19], [20], [21], [22]. The main goal of these *quality-aware FR models* is not necessarily quality assessment, instead quality estimates of the input images are generated as a side product of the models' design. Nonetheless, contemporary quality-aware models have been demonstrated to yield state-of-the-art performance for the FIQA task as well. While considerable progress has been made in face image quality assessment over recent years, designing a FIQA model that ensures a good trade-off between performance, space and runtime efficiency, and that generalizes well across different datasets and FR models remains challenging.

To address these challenges, we introduce in this paper a novel FIQA technique, named DifFIQA (**Dif**fusion-based **F**ace **I**mage **Q**uality **A**ssessment), which exploits the capabilities of modern Denoising Diffusion Probabilistic Models (DDPMs) for quality assessment and is based on the following two key insights [5]:

- **Perturbation robustness:** Images of higher quality have stable representations in the embedding space of the given FR model and are less effected by the perturbations introduced by the forward diffusion process compared to lower quality images.
- **Reconstruction quality:** Higher-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 the above observations, DifFIQA analyzes the embedding stability of the given face images by perturbing them through the forward as well as backward diffusion process and then quantifies the results for quality estimation. While a preliminary version of the DifFIQA approach was presented in [5], we extend the initial approach in this paper by optimizing the (space and runtime) efficiency of the model, while also further improving performance. Specifically, we propose the **E**xtended DifFIQA model (eDifFIQA) that relies on a novel knowledge distillation scheme that distills the quality information extracted by DifFIQA into a quality-regression model that allows for quality prediction with a single forward pass through the model, instead of relying on the computationally heavy diffusion processes utilized by DifFIQA. Because an arbitrary backbone can be used for the regressor, the complexity of eDifFIQA can be controlled and adjusted towards the **desired performance-complexity trade-off**. Moreover, we infuse additional knowledge into the distilled model, by accounting for the following aspect of face-image-quality, often exploited by contemporary FIQA techniques [19], [21], i.e.:

- **Relative position in the embedding space:** The embeddings of high-quality face images are more likely to be located close to the centroids of the corresponding identities (or class-centers), whereas low-quality images typically fall into an area of the embedding space that is further away from the centroid of the given identity. During the distillation process, this insight can be used as an additional source of information when estimating the quality of training samples.

By accounting for the relative embedding-space position of the face images during the distillation process, we are able to derive a powerful lightweight quality predictor that inherits the well-defined theoretical motivation of the teacher DifFIQA procedure, while exhibiting highly desirable runtime characteristics.

We evaluate the proposed DifFIQA technique as well the extended distilled version, eDifFIQA, through comprehensive experiments on eight standard datasets, in comparison to ten state-of-the-art FIQA techniques and across six of the strongest face recognition models currently available. Our experimental results show that both DifFIQA and eDifFIQA lead to highly competitive performance, while the latter also allows for (space and runtime) efficient quality predictions. In summary, we make the following main contributions in this paper:

- We propose a novel technique, named DifFIQA, which leverages the generative capabilities of modern DDPMs, for the task of face image quality assessment. By exploring the perturbation robustness and reconstruction quality of the facial images, DifFIQA is able to make use of both, the forward (noising) and the backward (reconstruction) diffusion step, and accurately assess the quality of the given input sample.
- We propose a dedicated optimization-based knowledge distillation pipeline, named eDifFIQA (extended DiFIQA), which improves on the performance and run-time capabilities of DifFIQA, by using an additional source of quality information in the form of the relative position of samples in the embedding space.

2 Related Work

In this section, we review relevant prior work with the goal of providing context and the necessary background for the contributions presented in this paper. For a more comprehensive coverage of the reviewed areas, the reader is referred to some of the excellent surveys available in the literature, e.g., [4], [23].

2.1 Face Image Quality Assessment

Face Image Quality Assessment (FIQA) techniques aim at estimating the (biometric) quality of the input face images. While different definitions of quality can be found in the literature [4], the majority of modern solutions in this area adopt the notion of *utility* or *fitness* of the facial images for the recognition task as the synonym for quality. As such, face image quality is most often represented in the form of a single scalar score, where a lower value reflects a lower quality. Because these scalar values encode the overall quality characteristics of the images and not specific aspects (e.g., sharpness, occlusion level, frontalness, etc.), they are also often referred to as **unified quality scores**. Based on how the quality scores are estimated, existing techniques can in general be further partitioned into: (i) unsupervised (analytical) and (ii) supervised (regression) methods.

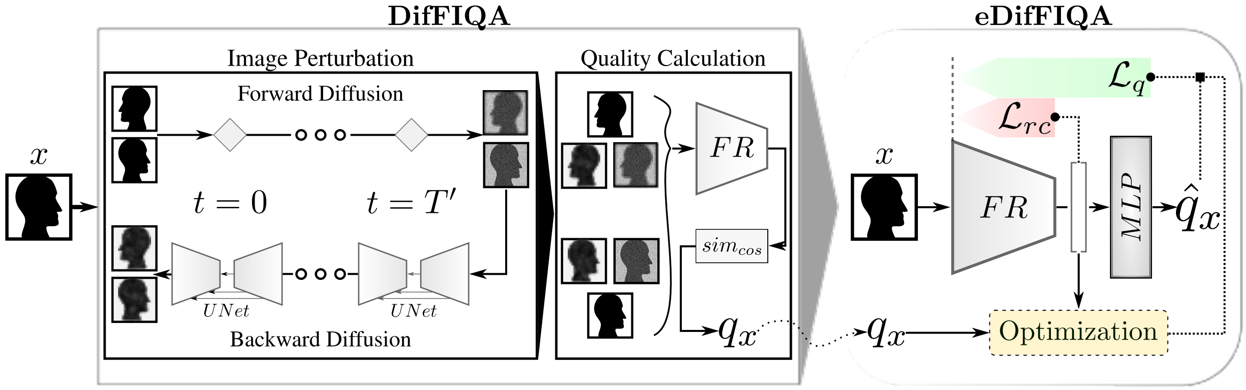


Fig. 1. **Overview of the DifFIQA and eDifFIQA.** The base approach of DifFIQA, consists of two main parts: the *Diffusion Process* and the *Quality-Score Calculation.* The diffusion process uses a custom UNet model, to generate noisy and reconstructed images using the forward and backward diffusion processes, respectively. To capture the effect of face pose on the quality estimation procedure, the process is repeated with a horizontally flipped image. The Quality Score Calculation part estimates the quality of samples by producing and comparing the embeddings of the original sample and the images generated by the diffusion part. To improve on the performance and computational complexity of the base DifFIQA approach, eDifFIQA employs knowledge distillation and label optimization. Here, the quality label q_x of a given training sample x produced by DifFIQA, is first optimized using additional quality information derived from the relative position of the sample within the embedding space of the feature extraction FR model. The quality information is then distilled into the model, consisting of a feature extraction FR backbone and an MLP quality-regression head, by employing a representation consistency $\mathcal{L}_{rc}$ and quality $\mathcal{L}_q$ loss, both of which help to improve the predictive capabilities of the final trained quality-regression model.

Unsupervised methods try to estimate sample quality by observing the characteristics of the sample either in the image space or in the latent/embedding space of a chosen recognition model. Because of this, such methods can be viewed as face-specific general purpose Image Quality Assessment (IQA) techniques [24], [25], [26]. Earlier methods from this group typically aim at assessing the quality by estimating face- and image-specific characteristics of the input images, such as illumination, pose, texture, occlusions, etc. [16], [17], [27], [28]. However, such techniques do not achieve competitive results when compared to modern FIQA methods. Conversely, state-of-the-art unsupervised FIQA methods focus on predicting unified quality scores and typically exploit the fact that the stability of image representations in the latent/embedding space is highly correlated with the quality of the input samples. To probe the stability of the computed representations, perturbations of the input samples or intermediate model representations are typically induced and then quantified to measure quality. One of the first methods using this approach, called SER-FIQ [14], for example, relies on the use of dropout layers to generate perturbations that can be evaluated for quality prediction. Another, more recent method, named FaceQAN [15], uses adversarial methods to generate perturbed samples, and consequently, to estimate quality scores. While recent unsupervised techniques, have shown great promise for the FIQA task, they are often computationally heavier than representatives from the group of supervised methods and due to their (recent) reliance on data perturbations may capture only a partial view of the overall sample quality.

Supervised methods most commonly train quality-regression models using pseudo quality labels as the targets for the regression task. The main differences among existing solutions, therefore, come from the pseudo-quality-label generation process. Earlier works [29], for example, used human annotations to generate quality labels. However, such approaches do not achieve competitive performance from today's perspective, as (human) perceivable quality does not fully capture all factors and image characteristics relevant for *the utility of face images for recognition* – the common definition of face image quality. One of the first approaches not relying on human annotations is FaceQnet [10], [30]. The model computes quality labels by comparing the latent representations (or embeddings) of all images of a given subject to the subject's highest quality image. In this case, the highest quality image is determined by using third-party software. Similar to FaceQnet, most newer techniques rely on comparison scores to generate quality labels. PCNet [11], for instance, uses many mated image pairs (pair of distinct images of the same individual), while SDD-FIQA [13] uses both, non-mated (pair of distinct images of different individuals) and mated pairs of images for generation of pseudo quality labels. Notably, LightQNet [12] also focuses on minimizing the complexity of the regression model by employing a custom quality loss combined with a lightweight network architecture. In general, supervised FIQA technique are often computationally efficient and typically require only a single forward pass through the prediction model to generate a quality scores, but are strongly dependent on the label generation process to ensure competitive performance.

2.2 Quality-Aware Face Recognition

Quality-Aware Face Recognition models are related to FIQA techniques in the sense that they incorporate quality information in the training process of face recognition models, which then commonly acts, as a regularizer for the learning procedure. As a result of this setup, such models often produce quality predictions in addition to facial embeddings during runtime and exhibit improved reliability and performance in challenging real-world scenarios [19], [22]. To include quality in the learning process, techniques from this group employ custom loss functions, which encompass both, identity separability and sample quality information. The PFE approach from [18], for instance, learns to predict the mean and variance representation of the input images, where the mean represents the sample embedding and the variance the uncertainty of the embedding in the latent space. The sample quality can then be trivially obtained from the uncertainty vector. The more recent technique, MagFace [19], extends the ArcFace [31] loss with a magnitude-aware angular margin term that enables the model to distinguish between samples of differing qualities using the magnitude of the sample embedding. Another excellent technique, named CR-FIQA [21], infers the quality by observing the ratio between the distances of the sample to the positive class center and the nearest negative class center. Quality scores produced by quality-aware face recognition techniques (or, in other words, *model-based FIQA* methods) are highly competitive due to the linkage between face recognition models and the quality estimation process, but may still lead to sub-optimal results if the quality/utility needs to be estimated for FR models that are conceptually very different from their own backbone architecture.

2.3 Our Contributions

The first contribution of this paper, the DifFIQA technique, can be seen as an unsupervised technique that leverages the generative capabilities of modern DDPMs in combination with a targeted FR model. Thus, DifFIQA is in general designed towards a specific FR model, but the quality scores (as we show later) also generalize well across a wide variety of models, including transformer based FR models that have so far not been studied widely in the context of FIQA. From a conceptual point of view, DifFIQA is most closely related to FaceQgen [32], where a generative (GAN-based) model is used to improve the quality of input samples, while the discriminative model tries to distinguish between genuine and restored (high-quality) images. However, different from FaceQgen, DifFIQA analyzes the results from the forward (i.e., noising/degradation) as well as backward (denoising/restoration) diffusion processes for quality estimation. This dual approach to quality estimation leads to highly competitive FIQA results, as we demonstrate in the experimental section.

The second contribution of this work, the distilled eDiFIQA model, on the other hand, can be seen as a supervised approach that learns to predict quality from a teacher, while incorporating ideas from quality-aware face recognition models, such as CR-FIQA [21], [33]. Due to the feed-forward nature of eDiFIQA, the model is computationally efficient, but also ensures a highly desirable trade-off between performance, generalization capabilities and space/runtime complexity. The distillation process in this work is applied to DifFIQA, but is otherwise general and applicable to any existing FIQA technique.

3 Methodology

The stability of the image representations in the embedding space of modern FR models correlates heavily with the quality of the input face images, as demonstrated by the success of various recent FIQA techniques [14], [15]. One way to investigate this stability is by inducing 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. One of our main contributions, **the DifFIQA technique**, takes advantage of this insight, as shown in Fig. 1, and adopts a custom DDPM model for the generation of noisy and reconstructed images. The generated images are then passed through a selected FR model to explore the impact of the perturbations on the variability of the embedding corresponding to the input image. While this process is typically effective, it is also computationally expensive. We, therefore, introduce a second contribution, the **eDifFIQA technique**, that distills the knowledge from DifFIQA into a computationally efficient feed-forward model for the FIQA task.

3.1 Preliminaries

To make the paper self-contained, we briefly present the main concepts 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 [34].

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 a backward (denoising) process [34], [35]. 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 $\{0, 1, \dots, T\}$. 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 [36], [37], 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** $\mathcal{B}_d$, on the other hand, attempts to iteratively denoise the generated noisy 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 $t = T, \dots, 0$, $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, 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 in our case is an unconditional UNet model.

3.2 Overview of DifFIQA and eDifFIQA

Given an input face image x, the goal of DifFIQA is to predict the quality score $q_x \in \mathbb{R}$ by quantifying the impact 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 key components, dedicated to: (i) *image perturbation* and (ii) *quality-score calculation*. The image perturbation component relies on the forward diffusion process to create a noisy image x_t from the input sample x and the backward process to generate the restored (denoised) image

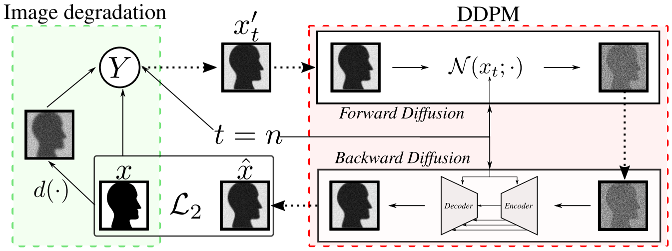


Fig. 2. **Presentation of the extended DDPM learning scheme.** Given a training sample x and a time step t, the proposed approach generates a time step dependent degraded image x'_t, by combining the original image with a degraded image using the function Y. The image x'_t is then used to generate a noisy sample using the standard forward diffusion approach. A UNet model is then trained to reconstruct the input sample in the backward process.

$\hat{x}$. 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 image 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. To make DifFIQA applicable to real-world scenarios, where computing resources are typically scarce, we distill the model into eDifFIQA. As can be seen from Fig. 1, the eDifFIQA model consists of a pretrained FR model M_r and a quality-regression head MLP that is trained using pseudo-quality labels produced by the original DifFIQA approach. However, to further improve performance, these pseudo quality labels are improved during training by incorporating information about the relative position of the training samples within the embedding space of the FR model M. During eDifFIQA training, a quality loss term is used to ensure accurate quality label predictions, while a representation-consistency loss is used to prevent large drifts of the regression backbone M_r, in turn improving the accuracy of the quality predictions.

3.3 Extended DDPM Training

The DDPM model used by DifFIQA, is based on the UNet architecture [38] and extends the standard DDPM training paradigm by: (i) incorporating time-dependent image degradations, and by (ii) limiting the noise added by the forward diffusion step. Details on the training process and our two extensions are given below.

Time-dependent image degradations. The training procedure starts by first degrading the input sample $x_0 = x$ using a degradation function $d(\cdot)$, such that $x' = d(x)$. Here, the BSRGAN [39] model is used to model the degradation process. The degraded sample is then combined with the original input sample to produce a time-dependent degraded image x'_t according to:

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

where $\ddot{\alpha}_t$ is a time-dependent variable, calculated as $\sin(\frac{t}{T} \cdot \frac{\pi}{2})$, that monotonically increases on the interval $t \in [0, T]$, such that $\ddot{\alpha}_{t=0} = 0$ and $\ddot{\alpha}_{t=T} = 1$, and $Y(\cdot)$ is the time-dependent degradation process. 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. The training sample x'_t is then used in the standard DDPM training scheme using the forward diffusion step to apply noise to the sample and the backward step (Eq. (2)) to reconstruct the input sample x, as shown in Fig. 2. Note that the (time-dependent) degradations allow the model to learn to gradually reverse the degradations and, in turn, to construct higher quality images during the backward diffusion process.

Noise limit. Diffusion models are often trained to reconstruct images from pure noise, guiding the generative (de-noising) process via conditioning on the chosen image (or text). 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. We, therefore, decide to limit the time steps on which the DDPM model is trained on $t \in [1, T']$, where $T' < T$, which in turn decreases the amount of noise in x'_t produced by the forward diffusion process and leads to proper conditioning on the given input image x. The extended training procedure then minimizes Eq. (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 model. Because head pose is an important factor of face quality, which the underlying DDPM cannot 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 [15]. 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 the given input face sample x, its flipped version x^f and a given time step t, we generate pairs of noisy (x_t, x^f_t) and restored images $(\hat{x}_t, \hat{x}^f_t)$ and use the generated data for quality-score calculation.

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.,

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

where $\mathcal{E}$ is a set of generated image embeddings, i.e, $\mathcal{E} = \{e_{x_t}, e_{\hat{x}}, e_{x^f}, e_{x^f_t}, e_{\hat{x}^f}\}$, computed with the FR model M as $e_z = M(x_z)$. In the above equation, the operator $|\cdot|$ denotes the set cardinality, $\|\cdot\|$ the L_2 norm nad q_x stands for the unified quality score of the input sample x, produced by DifFIQA.

3.6 Label Optimization & Knowledge 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 numerous forward passes through the generative network. Since DifFIQA repeats this process n-times, this only exacerbates the

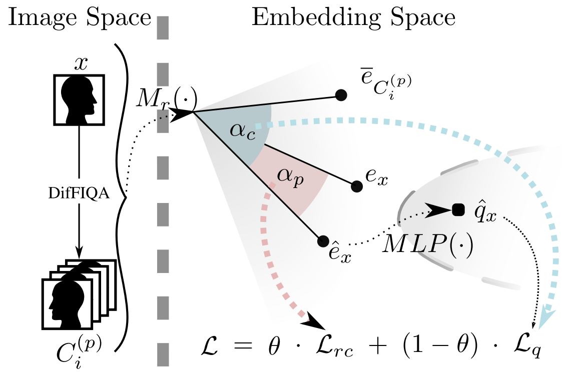


Fig. 3. **Overview of the distillation process, with added label optimization.** Prior to training, we first compute the embeddings e_x of all training samples x and the mean embedding $\overline{e}_{C_i}$ of the highest quality images for each identity C_i.

TABLE 1
Summary of the characteristics of the experimental datasets. We evaluate eDifFIQA across eight diverse datasets with different sizes and use-cases.

Dataset	#Images	#IDs	#Comparisons		Use Case
			Mated	Non-mated	
LFW [40]	13,233	5,749	3,000	3,000	General
Adience [41]	19,370	2,284	20,000	20,000	General
IJB-C [42]	23,124††	3,531	19,557	15,638,932	General
CALFW [43]	12,174	4,025	3,000	3,000	Cross-Age
CFP-FP [44]	7,000	500	3,500	3,500	Cross-Pose
CPLFW [45]	11,652	3,930	3,000	3,000	Cross-Pose
XQLFW [46]	13,233	5,749	3,000	3,000	Cross-Quality
YouTubeFaces [47]	3,425††	1,595	3,024	5,860,576	Cross-Video

†† number of templates/videos, each containing several images

problem and adversely affects the applicability of the technique in real-world applications. To address this issue, we **distill the knowledge** encoded by DifFIQA into a regression model, while simultaneously optimizing (improving) the initial quality labels produced by DifFIQA, by infusing additional knowledge into the distillation process, as illustrated in Fig. 3. Here, the additional knowledge comes in the form of the *relative position* of the embedding of the input sample x in comparison to the centroid of the identity that x corresponds to.

Formally, this process can be described as follows. Consider a dataset of N training images $\{x_i\}_{i=0}^N$ and let these N images correspond to K distinct identities $\{C_1, C_2, \ldots, C_K\}$. For the distillation process, we first compute the corresponding embeddings $\{e_{x_i}\}_{i=0}^N$ using the targeted FR model M, such that $e_x = M(x)$, and then estimate mean representations (class-centers/centroids) C of all K identities present among the training samples. To ensure these centroids are indicative of good quality samples, we make use of the quality labels produced by DifFIQA and only use p-percent of the highest quality images in the dataset. For each identity C_i, where $i = \{1, 2, \ldots, K\}$, we then construct the mean representation, as follows:

$$\overline{e}_{C_i} = \frac{1}{|C_i^{(p)}|} \sum_{x_j \in C_i^{(p)}} e_{x_j}^{(p)}, \tag{8}$$

where $e_{x_j}^{(p)}$ is the embedding of a sample from class C_i, and the superscript $^{(p)}$ indicates that the samples correspond to top p-percent of images in the training set. $|\cdot|$ again denotes the cardinality operator.

Once all embeddings and average representations for all K classes are computed, we distil the quality information from DifFIQA into our eDifFIQA quality-regressor. The model combines a pretrained backbone FR (for feature extraction) M_r with an additional quality-regression head, implemented through an MLP, for quality prediction. During training, we make use of a *representation consistency* $\mathcal{L}_{rc}$ and a *quality* $\mathcal{L}_q$ loss. The representation-consistency loss

$$\mathcal{L}_{rc} = 1 - \frac{\hat{e}_x^T \cdot e_x}{\|\hat{e}_x\| \cdot \|e_x\|}, \tag{9}$$

encourages the model M_r to produce consistent representations as the targeted FR model M, i.e., $\hat{e}_x = M_r(x)$, and in a sense adapts eDifFIQA to the characteristics of M. The quality loss, on the other hand, is defined as

$$\mathcal{L}_q = \|\hat{q}_x - q_x^o\|, \tag{10}$$

where $\hat{q}_x$ is the quality score predicted by the MLP and q_x^o the optimized quality label. The quality loss serves as the main component of the distillation process and ensures that the DifFIQA generated pseudo-quality labels are transferred to the regressor, while also considering the relative position in the embedding space. The optimization during the training is, hence, done using

$$q_x^o = (q_x - \epsilon \cdot \frac{\hat{e}_x^T \cdot \overline{e}_{C_i^{(p)}}}{\|\hat{e}_x\| \cdot \|\overline{e}_{C_i^{(p)}}\|}), \tag{11}$$

where q_x is the DifFIQA pseudo-quality label, $\overline{e}_{C_i}$ is the representation of the class-center of C_i that the input sample corresponds to, and ϵ a hyperparameter of the eDifFIQA technique that balances the impact of the pseudo-quality labels and the impact of the relative-embedding space position. In practice, the above equations suggests that samples further away from the class-centers will be punished more heavily by the optimization step and their pseudo-quality labels will be reduced by a larger amount and vice versa. The final loss term is calculated as:

$$\mathcal{L} = \theta \cdot \mathcal{L}_{rc} + (1-\theta) \cdot \mathcal{L}_q, \tag{12}$$

where θ is a hyperparameter that determines how much weight is put on either the representation-consistency or quality loss.

4 Experimental Setup

In this section, we present the experimental setup used to evaluate the proposed DifFIQA and eDifFIQA technique. Specifically, we discuss the state-of-the-art (SOTA) competitors and datasets utilized for the experiments, the evaluation methodology as well as hyperparameter settings and experimental hardware adopted for the evaluation.

4.1 Experimental Setting

We analyze the performance of DifFIQA and eDifFIQA in comparison to 10 state-of-the-art quality assessment methods: (i) the **unsupervised** FaceQAN [15], SER-FIQ [14], and FaceQgen [32] methods, (ii) the **supervised** FaceQnet [10], SDD-FIQA [13], PCNet [11], and LightQnet [12] techniques, and (iii) the **model-based** MagFace [19], PFE [18], and CR-FIQA [21] methods. We additionally implemented a simplified distilled model, by simply regressing to the DifFIQA generated pseudo-quality

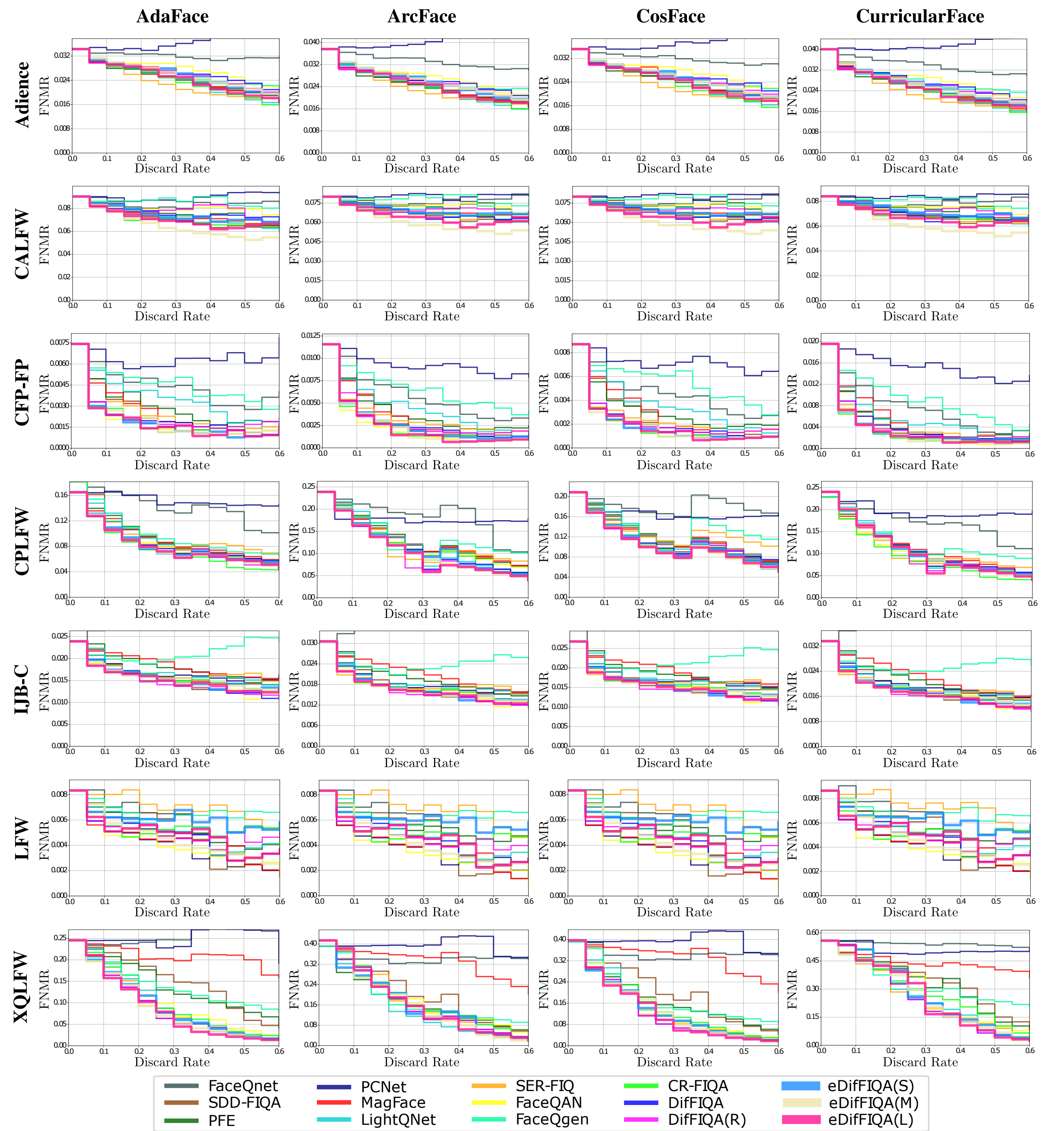


Fig. 4. **Comparison to the state-of-the-art with (non-interpolated) EDC curves and CNN-based face recognition models.** The results are presented for 7 datasets, 4 FR models and with 10 recent FIQA competitors. DifFIQA and eDifFIQA perform close-to or better than all other methods included in the experiments in all configurations, i.e., with a large (L), medium (M) and small (S) regression backbone.

labels, without any label optimization. We refer to this models as DifFIQA(R), similarly as in [5], and use it to study the impact of the label optimization process throughout the experiments. We test all methods on 8 commonly used benchmarks with vastly different quality characteristics, as summarized in Table 1, i.e.: Adience [41], Cross-Age Labeled Faces in the Wild (CALFW) [43], Celebrities in Frontal-Profile in the Wild (CFP-FP) [44], Cross-Pose Labeled Faces in the Wild (CPLFW) [45], large-scale IARPA Janus Benchmark-C (IJB-C) [42], Labeled Faces in the Wild (LFW) [40], Cross-Quality Labeled Faces in the Wild (XQLFW) [46] and YouTubeFaces [47]. 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 CNN-based models, i.e.: AdaFace[1] [22], ArcFace[2] [31],

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

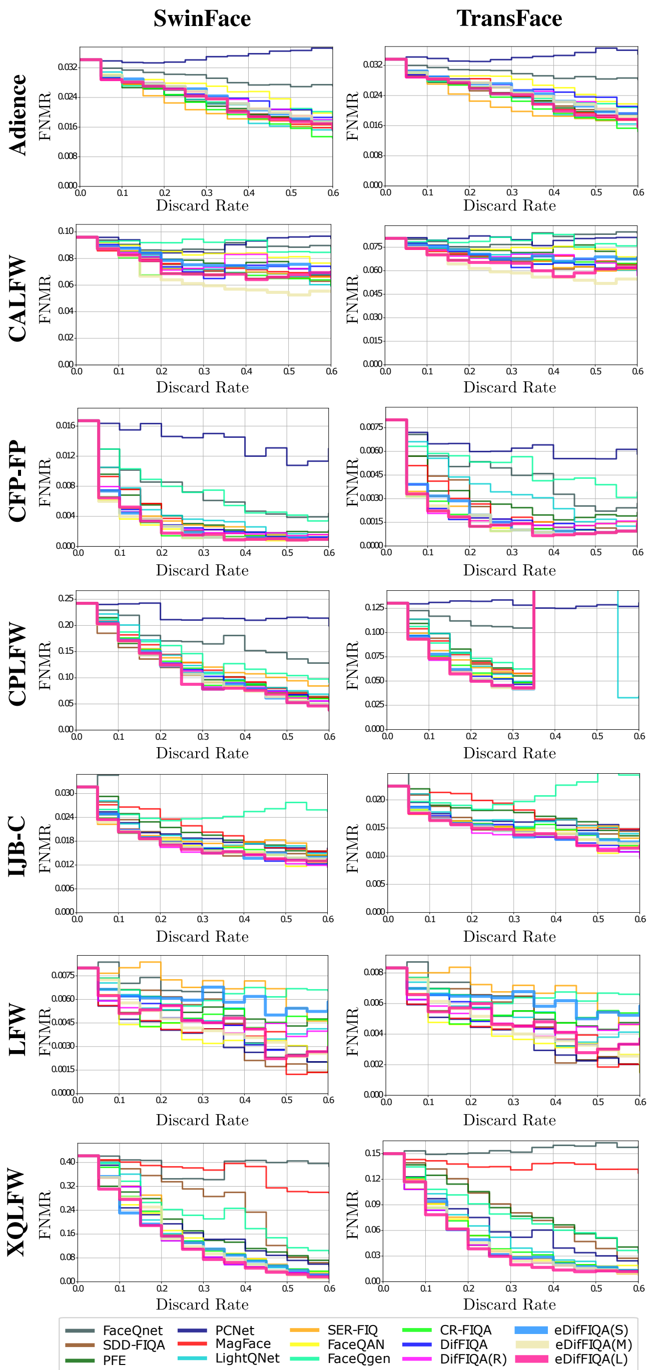


Fig. 5. **Comparison to the state-of-the-art with (non-interpolated) EDC curves and Transformer-based face recognition models.** Results are presented for 7 datasets, 2 FR models and with 12 recent FIQA competitors. DifFIQA and eDifFIQA perform close-to or better than all other methods included in the experiments.

CosFace[2] [48], and CurricularFace[3] [49] and 2 state-of-the-art Transformer-based models i.e.: SwinFace[4] [50], and TransFace[5] [51]. The CNN-based models use a ResNet100 backbone, while the Transformer-based models use ViT[5] and SwiN[4] backbones. The models are trained on the WebFace12M[1], MS1MV3[2,5,4], Glint360k[2,5], and CASIA-WebFace[3] datasets. The empirical validation with 12 SOTA competitors, over 8 datasets and with 6 face recognition models allows for a comprehensive evaluation of the proposed eDifFIQA techniques and an in-depth analysis of its main characteristics.

[3] https://github.com/HuangYG123/CurricularFace
[4] https://github.com/lxq1000/SwinFace
[5] https://github.com/DanJun6737/TransFace

4.2 Experimental Methodology

Following standard evaluation methodology [7], [14], [15], [21], [52], we use *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 [7], [52] for the evaluation. 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. In real-world applications it is often not feasible/practical to reject a large percentage of all samples and we are, therefore, 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 discard rate threshold (0.2 and 0.3) 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.

4.3 Implementation Details

The baseline DifFIQA technique uses a maximum of $T = 1000$ forward diffusion steps, but is trained only on the first $T' = 100$ to ensure that the image generated by the forward diffusion process is only partially noisy. This guarantees that both the noisy and reconstructed images are properly conditioned on the identity of the initial input image. The utilized UNet model consists of four downsampling and upsampling modules, each decreasing (increasing) the dimensions of the representations by a factor of two. Training of the UNet model is done using the Adam optimizer, with a learning rate of $8.0e^{-5}$ in combination with an Exponential Moving Average (EMA) model, with a decay rate of 0.995. During inference time, we use a much smaller number of forward steps $t = 5$, to generate noisy and reconstructed images. For the extended approach, we present three separate models at three different scales. The largest (L) uses the ResNet100 architecture (**eDifFIQA(L)**), the medium-sized (M) uses the ResNet50 architecture (**eDifFIQA(M)**), and the smallest (S) uses the ResNet18 architecture (**eDifFIQA(S)**). All versions use a pretrained FR model trained using the CosFace loss function combined with a quality regression MLP with 1024 hidden neurons. The quality-regression model is trained using the Adam optimizer with a learning rate of $1e^{-3}$ for the MLP head and a learning rate of $1e^{-4}$ for the feature extraction FR backbone, on roughly two million images from the VGGFace2 [53] dataset. We set p to 20%, meaning only the top twenty percent of all images are used for the calculation of the mean representations of each identity. We use a value of 0.5 for the hyperparameters ϵ and θ. The hyperparameter ϵ controls how drastically the quality labels change due to optimization, while the hyperparameter θ controls the balance between the representation consistency and quality loss terms. The values for both of the parameters were determined by a small sensitivity analysis, presented later in Section 5.5. All

TABLE 2
Comparison to the state-of-the-art using CNN-based face recognition models. 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. The best result among the baselines and proposed methods is marked with *.

AdaFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.963	0.938	0.738	0.887	1.256	0.884	0.977	0.949
SDD-FIQA [13]	0.839	0.871	0.513	0.688	0.782	0.825	0.842	0.766
PFE [18]	0.833	0.890	0.581	0.681	0.868	0.771	0.798	0.775
PCNet [11]	1.005	0.979	0.851	0.898	0.788	0.661	0.987	0.881
MagFace [19]	0.860	0.866	0.537	0.664	0.883	0.666	0.913	0.770
LightQNet [12]	0.847	0.894	0.659	0.684	0.797	0.777	0.704	0.766
SER-FIQ [14]	**0.807**	0.892	0.486	0.626	0.762	0.935	$0.654^\dagger$	0.738
FaceQAN [15]	0.890	0.919	**0.392**	0.619	0.756	**0.656**	0.654	0.698
CR-FIQA [21]	0.844	0.851	0.404	**0.588**	0.750	0.707	0.685	0.690
FaceQgen [32]	0.858	0.970	0.739	0.694	0.853	0.834	0.736	0.812
DifFIQA [5]	0.864	0.905	0.426	0.656	0.761	0.730	0.627	0.710
DifFIQA(R) [5]	0.865	0.895	0.421	0.646	$\mathbf{0.731}^*$	0.708^*	$\mathbf{0.610}^*$	0.697
eDifFIQA(S)	0.868	0.892	0.403	0.651	0.760	0.786	0.647	0.716
eDifFIQA(M)	0.871	$\mathbf{0.834}^*$	0.414	0.645^*	0.747	0.728	0.620	0.694
eDifFIQA(L)	0.849^*	0.854	0.399^*	0.646	0.743	0.713	0.629	0.690

ArcFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.943	0.955	0.702	0.878	1.224	0.884	0.899	0.926
SDD-FIQA [13]	0.783	0.901	0.497	0.734	0.720	0.808	0.774	0.745
PFE [18]	0.774	0.932	0.522	0.738	0.783	0.779	0.641	0.738
PCNet [11]	1.022	1.006	0.863	0.783	0.706	**0.623**	1.004	0.858
MagFace [19]	0.812	0.902	0.500	0.717	0.824	0.635	0.943	0.762
LightQNet [12]	0.789	0.913	0.582	0.752	0.721	0.745	**0.621**	0.732
SER-FIQ [14]	**0.767**	0.903	0.446	0.656	0.671	0.935	$0.676^\dagger$	0.722
FaceQAN [15]	0.824	0.941	**0.347**	0.677	0.673	0.624	0.667	0.679
CR-FIQA [21]	0.808	0.891	0.369	0.689	0.664	0.675	0.680	0.682
FaceQgen [32]	0.817	0.985	0.724	0.701	0.785	0.802	0.653	0.781
DifFIQA [5]	0.805	0.918	0.402	0.674	0.675	0.714	0.652	0.691
DifFIQA(R) [5]	0.801	0.898	0.408	$\mathbf{0.646}^*$	$\mathbf{0.655}^*$	0.708	0.653	0.681
eDifFIQA(S)	0.822	0.916	0.402	0.679	0.679	0.786	0.634^*	0.703
eDifFIQA(M)	0.825	$\mathbf{0.846}^*$	0.386	0.669	0.663	0.712	0.671	0.682
eDifFIQA(L)	0.793^*	0.872	0.376^*	0.667	0.663	0.705^*	0.672	0.678

CosFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.952	0.955	0.690	0.879	1.317	0.884	0.899	0.939
SDD-FIQA [13]	0.825	0.901	0.491	0.735	0.741	0.808	0.774	0.753
PFE [18]	0.813	0.932	0.524	0.748	0.808	0.779	0.641	0.749
PCNet [11]	1.009	1.006	0.868	0.835	0.729	**0.623**	1.004	0.868
MagFace [19]	0.852	0.902	0.549	0.724	0.836	0.635	0.943	0.777
LightQNet [12]	0.835	0.913	0.612	0.753	0.750	0.745	0.621	0.747
SER-FIQ [14]	**0.793**	0.903	0.426	0.711	0.699	0.935	0.567	0.719
FaceQAN [15]	0.871	0.941	0.373	0.667	0.702	0.624	0.581	0.680
CR-FIQA [21]	0.835	0.891	**0.361**	0.681	0.696	0.675	0.631	0.681
FaceQgen [32]	0.847	0.985	0.784	0.702	0.794	0.802	0.653	0.795
DifFIQA [5]	0.841	0.918	0.402	0.671	0.707	0.714	0.561	0.688
DifFIQA(R) [5]	0.838	0.898	0.389	0.660	$\mathbf{0.677}^*$	0.708	$\mathbf{0.556}^*$	0.675
eDifFIQA(S)	0.854	0.916	0.364^*	0.669	0.706	0.786	0.580	0.696
eDifFIQA(M)	0.851	$\mathbf{0.846}^*$	0.373	0.657	0.689	0.712	0.608	0.677
eDifFIQA(L)	0.828^*	0.872	0.376	$\mathbf{0.653}^*$	0.690	0.705^*	0.577	0.672

CurricularFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.921	0.947	0.602	0.867	1.248	0.908	0.984	0.925
SDD-FIQA [13]	0.776	0.900	0.409	0.696	0.721	0.821	0.817	0.734
PFE [18]	0.759	0.923	0.420	0.691	0.784	0.785	0.835	0.742
PCNet [11]	1.004	0.996	0.880	0.899	0.710	0.656	0.938	0.869
MagFace [19]	0.793	0.892	0.472	0.689	0.821	0.661	0.862	0.741
LightQNet [12]	0.769	0.910	0.463	0.704	0.713	0.767	0.739	0.724
SER-FIQ [14]	**0.750**	0.883	0.383	0.625	0.661	0.942	$\mathbf{0.721}^\dagger$	0.709
FaceQAN [15]	0.811	0.931	0.342	0.637	0.669	**0.644**	0.835	0.696
CR-FIQA [21]	0.797	0.877	**0.313**	**0.615**	0.664	0.693	0.789	0.678
FaceQgen [32]	0.815	0.974	0.661	0.698	0.783	0.845	0.750	0.790
DifFIQA [5]	0.806	0.905	0.384	0.669	0.672	0.730	0.761^*	0.704
DifFIQA(R) [5]	0.788	0.892	0.358	0.650^*	$\mathbf{0.644}^*$	0.724	0.785	0.692
eDifFIQA(S)	0.802	0.919	0.330	0.664	0.671	0.806	0.803	0.714
eDifFIQA(M)	0.799	$\mathbf{0.835}^*$	0.326^*	0.664	0.650	0.741	0.775	0.684
eDifFIQA(L)	0.773^*	0.869	0.330	0.664	0.652	0.721^*	0.804	0.688

†-SER-FIQ was used to create XQLFW.

TABLE 3
Comparison to the state-of-the-art using Transformer-based face recognition models. 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. The best result among the baselines and proposed methods is marked with *.

SwinFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.919	0.937	0.640	0.833	1.241	0.914	0.927	0.916
SDD-FIQA [13]	0.811	0.862	0.448	**0.650**	0.728	0.827	0.872	0.743
PFE [18]	0.800	0.877	0.456	0.688	0.792	0.805	0.672	0.727
PCNet [11]	0.987	0.935	0.937	0.955	0.729	**0.642**	0.649	0.833
MagFace [19]	0.828	0.854	0.450	0.735	0.827	0.654	0.943	0.756
LightQNet [12]	0.819	0.860	0.475	0.759	0.732	0.769	0.666	0.726
SER-FIQ [14]	**0.778**	0.862	0.393	0.720	0.691	0.960	$0.657^\dagger$	0.723
FaceQAN [15]	0.863	0.921	**0.335**	0.685	0.692	0.643	0.623	0.680
CR-FIQA [21]	0.807	0.813	0.335	0.708	0.675	0.696	0.640	0.668
FaceQgen [32]	0.819	0.972	0.639	0.769	0.805	0.821	0.745	0.796
DifFIQA [5]	0.833	0.873	0.380	0.689	0.694	0.737	0.604	0.687
DifFIQA(R) [5]	0.836	0.859	0.378	0.689	$\mathbf{0.664}^*$	0.731	$\mathbf{0.567}^*$	0.675
eDifFIQA(S)	0.847	0.886	0.365	0.689	0.684	0.812	0.585	0.696
eDifFIQA(M)	0.842	$\mathbf{0.794}^*$	0.341^*	0.680	0.672	0.735	0.618	0.669
eDifFIQA(L)	0.823^*	0.836	0.349	0.671^*	0.669	0.727^*	0.576	0.664

TransFace - pAUC@FMR=10^{-3}(↓)								
FIQA Model	Adience	CALFW	CFP-FP	CPLFW	IJB-C	LFW	XQLFW	$\overline{pAUC}$
FaceQnet [10]	0.938	0.964	0.753	0.889	1.291	0.898	1.007	0.963
SDD-FIQA [13]	0.837	0.905	0.545	0.694	0.809	0.840	0.819	0.778
PFE [18]	0.834	0.930	0.546	0.711	0.879	0.810	0.798	0.787
PCNet [11]	0.998	1.003	0.837	1.010	0.804	**0.661**	0.654	0.853
MagFace [19]	0.867	0.898	0.512	0.682	0.930	0.673	0.935	0.786
LightQNet [12]	0.847	0.916	0.665	0.697	0.817	0.784	0.634	0.766
SER-FIQ [14]	**0.784**	0.901	0.430	0.647	0.773	0.935	$0.585^\dagger$	0.722
FaceQAN [15]	0.890	0.942	0.377	0.599	0.774	0.663	0.558	0.686
CR-FIQA [21]	0.832	0.887	0.379	0.614	0.765	0.714	0.580	0.681
FaceQgen [32]	0.858	0.985	0.761	0.699	0.877	0.834	0.739	0.822
DifFIQA [5]	0.863	0.917	0.405	0.597	0.782	0.736	0.529	0.690
DifFIQA(R) [5]	0.856	0.901	0.394	0.579	$\mathbf{0.744}^*$	0.722^*	$\mathbf{0.523}^*$	0.674
eDifFIQA(S)	0.867	0.917	0.446	0.591	0.770	0.825	0.558	0.711
eDifFIQA(M)	0.861	$\mathbf{0.849}^*$	0.384	0.577	0.759	0.750	0.553	0.676
eDifFIQA(L)	0.836^*	0.872	$\mathbf{0.374}^*$	$\mathbf{0.575}^*$	0.755	0.735	0.528	0.668

†SER-FIQ was used to create XQLFW.

experiments were conducted on a desktop PC with an Intel i9-10900KF CPU, 64 GB of RAM and an Nvidia 3090 GPU.

5 Results and Discussions

We evaluate the proposed FIQA models in comprehensive experiments that: (i) compare DifFIQA and eDifFIQA to a broad range of conceptually distinct state-of-the-art FIQA techniques, (ii) investigate the space and runtime complexity of the proposed approached, (iii) explore the applicability of the methods for predicting the quality of video frames instead of static images, (iv) analyze the robustness of the models to face misalignment, (v) ablate various model components to study their impact, and (vi) qualitatively evaluate different aspects of the proposed DifFIQA and eDifFIQA techniques.

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

Prior studies on face image quality assessment focused primarily on exploring the performance of FIQA methods with CNN-based face recognition techniques. However, Transformer-based models have recently been shown to lead to highly competitive performance in face recognition as well, but the behavior of such models with modern FIQA solutions has, to the best of our knowledge, not been studied in the literature before. Here, we therefore split our analysis into two parts and separately discuss results for CNN and Transformer-based face recognition models. For this series of experiment, we use the seven datasets that contain static images.

Analysis on CNN-Based Face Recognition Models. In Figure 4, we show the EDC curves of all CNN-based models and benchmark datasets, while the corresponding pAUC scores, calculated at a discard rate of 30%, are shown in Table 2. Additional results for a discard ratio of 20% are included in the supplementary material. From the reported results we make two key observations: (i) the proposed label-optimized knowledge distillation improves on the results over the two baseline approaches, i.e., DifFIQA and DifFIQA(R), and (ii) one of the tested eDifFIQA techniques consistently performs the best across all the tested state-of-the-art methods, achieving the best or second-best (average) $\overline{pAUC}$ result

TABLE 4
Space and Runtime complexity of the diffusion-based FIQA methods. The reported results (in ms) were computed over the XQLFW dataset and the same experimental hardware. The lower row shows the parameter count ($\times 10^6$)

FIQA model	Baseline methods [5]		eDifFIQA variants		
	DifFIQA	DifFIQA(R)	eDifFIQA(S)	eDifFIQA(M)	eDifFIQA(L)
Runtime ($\mu \pm \sigma$) [in ms]	1074.62 ± 11.45	1.24 ± 0.36	0.36 ± 0.77	0.63 ± 0.75	1.09 ± 0.77
# Parameters (10^6)	$142.87 + 65.15^{\dagger}$	65.15	24.55	44.11	65.68

†The sum refers to the number of parameters of the generative UNet model and the recognition model.

with all targeted CNN-based models. The proposed eDifFIQA in general outperforms the baseline diffusion-based FIQA methods on all datasets and with all face recognition model, except on the IJB-C and XQLFW benchmarks, where we observe a minute decrease in performance. Nonetheless, the proposed models still convincingly outperform all other competitors on this two datasets as well. When comparing results across different face recognition models, we see a slightly weaker (but still competitive) average performance in comparison to other FIQA techniques with the CurricularFace model. Interestingly, this is also the worst performing FR model according to our verification experiments (shown in the supplementary material) and this fact likely also contributes to the observed results.

Analysis on Transformer-Based Models. In Figure 5, we show the EDC curves on the two Transformer-based face recognition models and all benchmark datasets, while the corresponding pAUC scores calculated at a discard rate of 30% are shown in Table 3. Additional results for a discard ratio of 20% are included in the supplementary material. The results convey a similar story as the results reported for the CNN-based models. The proposed eDifFIQA models achieve the best and third best results on both Transformer models, outperforming all other state-of-the-art FIQA techniques. Among the different eDifFIQA models, the largest model performs the best, followed in order by the medium and then the small model. These results are in line with expectation, as the larger backbones have more capacity to capture the various quality characteristics of the different input face images. However, the medium-sized and small models still lead to competitive performance. Additionally, we also observe that the results improve when compared to the baselines (DifFIQA and DifFIQA(R)), however there once again appears to be slight decrease on the IJB-C and XQLFW benchmark datasets, where the baseline diffussion-based FIQA models have a slight edge.

5.2 Space & Runtime Complexity Analysis

Next, we analyze the space and runtime complexity of the proposed eDifFIQA approach from two aspects: (i) in comparison to the baseline difussion-based DifFIQA and DifFIQA(R) techniques, and (ii) in comparison to the state-of-the-art competitors.

Analysis of Diffusion-based FIQA methods. In Table 4, we report the time (in ms) needed to process a single input image and estimate the quality score on our experimental hardware and the parameter count of the models involved in the computations. The runtime for this experiment was estimated over the XQLFW dataset. As expected, we observe that the two models of comparable size, i.e., DifFIQA(R) and eDifFIQA(L), exhibit similar runtimes, while the lowest runtime is achieved with eDifFIQA(S), which decreased the runtime of the baseline DifFIQA by at least three orders of magnitude. While a speed-up of a factor of close to $1000\times$ over the initial DifFIQA approach is already observed

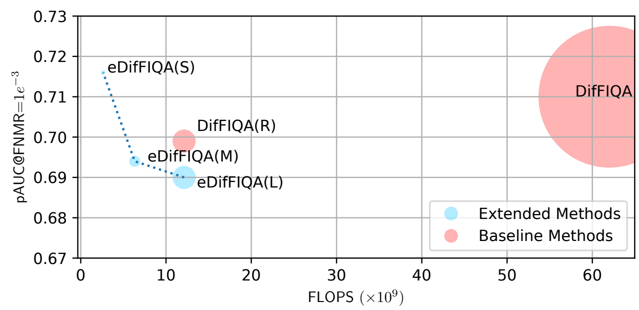


Fig. 6. **Complexity analysis between the baseline methods and eDifFIQA.** We compare the trade-off between performance and complexity of the baseline and extended methods. Performance is measured using the average pAUC values over all datasets when using the AdaFace FR model, while the complexity is measured using FLOPS.

with the larger distilled models, DifFIQA(R) and eDifFIQA(L), the change to a smaller regression backbone further improves on this result by a factor of close to $3\times$.

To get better insight into what this reduced runtime means with respect to performance, we show in Figure 6 the trade-off between the model complexity (measured using FLOPS) and performance measured using pAUC, for both of the baseline methods and the three versions of eDifFIQA. The pAUC presented was calculated on the AdaFace model, as the average pAUC of all used datasets. Note, that lower pAUC values represent better performance. From the presented results, we can observe an increase in performance when using the extended pipeline. Looking at models of similar complexity such as DifFIQA(R) and eDifFIQA(L) we also observe a noticeable increase in performance of 0.1. Interestingly, the least complex eDifFIQA(S) shows only a small decrease (less than 0.1) in performance when compared to the baseline DifFIQA approach, yet is notably less complex. These results and observations showcase the strength of our proposed label-optimized knowledge distillation.

Comparison to the State-Of-The-Art. Table 5 shows the runtime complexity of the competing FIQA methods. We note at this point that the runtime analysis is only a rough approximation of the actual runtime of methods and is highly dependent on hardware and software specifications of the target machine. For this reason, we compare only the order of magnitudes of the actual runtimes. We see that most methods achieve a runtime between $1 - 10ms$, with two notable exceptions i.e.: SER-FIQ and FaceQAN, which use several passes through a targeted FR model to estimate quality and therefore require a longer execution time. Comparing these results to the runtime assessment of DifFIQA and eDifFIQA, it is clear that the baseline DifFIQA has by far the longest runtime, due to the complexity of the generative model and the use of several runs to compute the final score, while our smallest presented model, eDifFIQA(S), is among the fastest methods. The medium-

TABLE 5
Runtime complexity of state-of-the-art methods. The reported results (in ms) were computed over the XQLFW dataset and the same experimental hardware.

FIQA model	FaceQnet [10]	SDD-FIQA [13]	PFE [18]	PCNet [11]	MagFace [19]	LightQNet [12]	SER-FIQ [14]	FaceQAN [15]	CR-FIQA [21]	FaceQgen [32]
Runtime ($\mu \pm \sigma$)	1.95 ± 0.90	0.62 ± 0.36	5.67 ± 0.79	2.16 ± 0.74	1.08 ± 0.36	1.51 ± 0.99	112.93 ± 33.81	334.13 ± 118.79	1.10 ± 0.74	1.97 ± 0.28

sized model, eDifFIQA(M), also leads to a runtime performance that is on par with the fastest of the tested competitors.

5.3 Video Recognition Experiments

The vast majority of FIQA techniques presented in the literature are predominantly targeting static images, while video data is only rarely considered. In the next series of experiments, we therefore simulate real-world scenarios in which the system is given several frames (or a video) of an individual with the task of first ranking the frames based on quality and then performing face recognition. Based on this task, we design two distinct FIQA application scenarios for the experiments, i.e.,

- **Best-Frame Scenario:** In the best-frame scenario, each given video is subjected to the selected FIQA techniques and the goal is to identify the highest quality frame in the given sequence. Performance evaluation (i.e., verification experiments) are then conducted with the highest quality frame.
- **Quality-Weighted Scenario:** In the quality-weighted scenario each frame in the video sequence is again first assigned a quality score using the given FIQA approach, but then a weighted representation is constructed for each video by utilizing the computed quality scores as weights for the combined representation.

Below we evaluate all considered FIQA techniques in the two outlined scenarios.

Analysis of Best-Frame Experiments. In Table 6, we show the results of the best-frame experiments and report the True Acceptance Rate calculated at two separate False Acceptance Rates, i.e.: $1e^{-3}$ and $1e^{-4}$ for all CNN-based and Transformer-based FR models. We observe that the baseline DifFIQA and extended eDifFIQA models outperform all other methods in most of the experiments. Interestingly, when comparing results between the baseline and eDifFIQA methods, we notice a clear divide when using different FR models. For the ArcFace and CosFace model the DifFIQA(R) method seems to slightly outperform the eDifFIQA techniques, while for all other FR models the eDifFIQA clearly outperforms the baseline approaches.

Analysis of Quality-Weighted Experiments. Table 7 presents the results of the quality-weighted experiment. Note, that this scenario is similar to the IJB-C Mixed Verification protocol, for which we show results in the supplementary material. As with the best-frame scenario, we report the True Acceptance Rate calculated at two separate False Acceptance Rates, i.e.: $1e^{-3}$ and $1e^{-4}$ for all CNN-based and Transformer-based FR models. It is interesting to observe that the performance of the two baselines diffusion models, DifFIQA and DifFIQA(R), and all eDifFIQA-model variants is still the highest among all tested techniques on average, but the performance gap is considerable smaller than in the case of the best-frame experiments, described above. Another interesting observation is that the smallest of the three proposed models in some-cases performs better than the two larger models, such as when using the AdaFace model, suggesting that for quality-based template generation, lightweight models may provide a good trade-off between performance and complexity.

5.4 Alignment Sensitivity Experiments

Modern FR models still require proper alignment of the input face images to ensure the best possible performance. Here, the alignment process is commonly done by matching the key-points of the eyes, nose, and mouth to predetermined points on the images. The accuracy of alignment is heavily dependent on the capabilities of the underlying face and key-point detection method and since FR models are trained on images aligned using a specific face (key-point) detector, their performance often decreases when a different detector is used. Because FIQA techniques are closely related to the FR models, they face similar issues and, among other, are also expected to be sensitive to miss-alignment.

To evaluate the sensitivity of FIQA techniques to the alignment of the input face samples, we perform experiments on images aligned using four different face (key-point) detectors i.e.: RetinaFace [54] (ResNet100 & MobileNet), MTCNN [55] and DLib [56]. We run the standard evaluation protocol for FIQA techniques over images aligned with the help of each of the three detector separately and report the average pAUC values, calculated at a discard rate of 30%, over all experiments. The results are shown in Table 8 for CNN-based FR models and in Table 9 for Transformer-based FR models. Additional results for a discard rate of 20% and all FIQA techniques are presented in the supplementary material.

From the reported results, we observe that the behavior of the tested techniques closely matched the behavior observed during the state-of-the-art comparison in Section 5.1. The eDifFIQA methods outperform all other FIQA techniques on most of the tested FR models, only falling behind the excellent CR-FIQA approach with the ArcFace and CurricularFace models. When comparing eDifFIQA to the initial DifFIQA and DifFIQA(R) approaches, we see that the extended approach achieves better results overall, while only occasionally performing worse, such as on the Adience and XQLFW datasets. Interestingly, LightQNet seems to be the most affected by miss-alignment as its performance drops significantly, when compared to the results, with proper alignment in Section 5.1.

5.5 Ablation study

In this section, we explore how the values of the two hyperparameters (ϵ, θ), included in the extended approach, influence the performance of the eDifFIQA models. To limit the scale of our analysis, we investigate the effects of each hyperparameter independently of one another. The hyperparameter ϵ controls the degree of change of the quality labels during optimization, we perform experiments with four different values of, ϵ i.e.: $0.1, 0.2, 0.5, 1.0$. The hyperparameter θ, on the other hand, represents the weight factor when combining the representation consistency and quality loss terms. The representation consistency term is weighted using

TABLE 6
Best-Frame video experiment. The table reports the True Acceptance Rate (TAR) at two different False Acceptance Rates of $1e^{-3}$ and $1e^{-4}$. The overall best result is colored green, the second best blue and the third best red. The best result among the baselines and proposed methods is marked with $*$.

FIQA Model	CNN-Based Models - TAR@FAR(↑)								Transformer-Based Models - TAR@FAR(↑)			
	AdaFace		ArcFace		CosFace		CurricularFace		SwinFace		TransFace	
	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$
FaceQnet [10]	85.020	81.151	83.862	79.101	84.987	81.382	83.366	77.811	83.929	80.093	86.310	84.061
SDD-FIQA [13]	88.393	84.689	87.202	82.771	88.558	84.888	86.640	81.779	87.202	83.829	89.451	87.202
PFE [18]	88.856	85.417	87.500	83.366	89.021	85.615	86.839	82.705	87.765	84.226	89.583	87.632
PCNet [11]	88.690	85.681	87.566	83.532	88.591	85.714	87.070	82.407	88.459	84.358	89.749	87.963
MagFace [19]	87.798	84.656	86.409	82.407	87.731	84.656	85.615	81.085	86.310	82.870	88.492	86.772
LightQNet [12]	89.054	85.813	87.731	83.962	88.955	86.045	87.302	83.399	87.798	83.796	89.980	88.228
SER-FIQ [14]	88.161	85.020	87.103	82.606	88.393	84.888	86.276	81.548	87.434	82.804	89.352	87.401
FaceQAN [15]	88.657	86.111	87.599	83.565	88.724	85.847	87.136	83.466	87.897	84.425	89.451	87.599
CR-FIQA [21]	88.624	85.483	87.698	84.160	88.558	85.747	87.235	82.804	87.798	84.094	89.517	87.335
FaceQgen [32]	86.343	82.374	85.119	80.589	87.136	83.300	84.392	79.001	85.615	80.721	88.062	85.648
DifFIQA [5]	88.690	85.813	87.368	84.028	88.724	86.243	87.302	82.970	87.897	84.259	89.385	87.864
DifFIQA(R) [5]	89.187	85.780	**88.558***	**84.590***	**89.385***	**86.442***	87.599	83.896	88.492	85.119	90.013	88.327
eDifFIQA(S)	**89.484***	**86.409***	88.128	84.028	89.087	85.946	**87.963***	83.796	88.459	84.788	**90.410***	88.624
eDifFIQA(M)	89.187	85.946	87.765	84.458	89.220	86.078	87.632	83.499	88.360	84.854	89.881	88.228
eDifFIQA(L)	89.120	86.210	88.261	84.292	89.021	86.442	87.864	**84.259***	**88.757***	**85.218***	89.980	**88.724***

TABLE 7
Quality-Weighted video experiment. The table reports the True Acceptance Rate (TAR) at two different False Acceptance Rates of $1e^{-3}$ and $1e^{-4}$. The overall best result is colored green, the second best blue and the third best red. The best result among the baselines and proposed methods is marked with $*$.

FIQA Model	CNN-Based Models - TAR@FAR(↑)								Transformer-Based Models - TAR@FAR(↑)			
	AdaFace		ArcFace		CosFace		CurricularFace		SwinFace		TransFace	
	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$	FAR=$1e^{-3}$	FAR=$1e^{-4}$
FaceQnet [10]	91.071	88.955	90.212	88.161	90.675	88.790	89.782	87.169	90.443	87.864	91.534	90.013
SDD-FIQA [13]	91.468	88.988	90.476	88.558	90.807	89.087	90.046	87.302	90.873	88.194	91.667	90.179
PFE [18]	91.534	89.054	90.575	88.525	90.939	88.922	90.079	87.169	90.906	88.228	91.700	90.245
PCNet [11]	91.567	89.021	**90.608**	**88.624**	90.939	89.021	90.112	87.202	**90.939**	88.360	91.733	90.410
MagFace [19]	91.369	88.889	90.476	88.360	90.807	88.955	90.013	87.037	90.774	87.996	91.667	90.212
LightQNet [12]	88.657	85.780	87.599	85.185	88.294	85.847	87.235	83.730	87.831	85.185	88.889	87.401
SER-FIQ [14]	91.567	89.087	90.509	88.459	**91.038**	89.021	**90.212**	87.632	90.774	88.360	**91.799**	**90.542**
FaceQAN [15]	91.501	89.087	90.476	88.558	90.906	89.187	90.212	87.533	90.807	88.459	91.733	90.245
CR-FIQA [21]	91.336	88.955	90.344	88.492	90.807	88.955	90.013	87.169	90.741	88.228	91.634	90.212
FaceQgen [32]	91.336	88.889	90.344	88.360	90.840	88.922	89.947	87.070	90.675	88.029	91.534	90.146
DifFIQA [5]	91.336	89.054	90.542*	88.558	90.840	88.988	90.046	87.235	90.741	88.327	91.634	90.179
DifFIQA(R) [5]	91.336	89.021	90.509	88.492	90.807	89.054	90.046	87.335	90.774	88.327	91.634	90.245
eDifFIQA(S)	**91.601***	**89.220***	90.476	88.558	90.972	**89.220***	90.212*	**87.731***	90.840*	**88.591**	91.766*	90.344*
eDifFIQA(M)	91.501	89.153	90.542	88.525	91.038*	89.153	90.179	87.698	90.873	88.492*	91.733	90.344
eDifFIQA(L)	91.601	89.153	90.509	88.624*	90.939	89.087	90.212	87.500	90.873	88.426	91.667	90.245

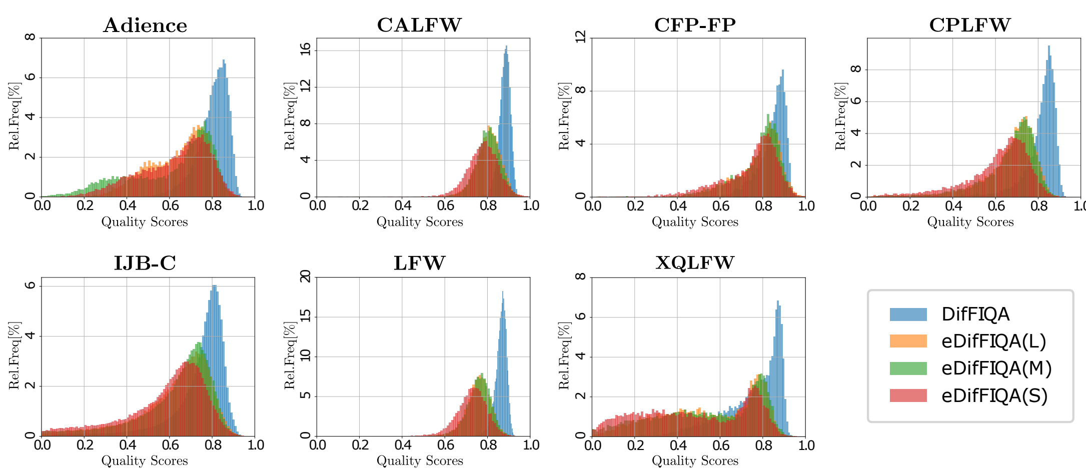


Fig. 7. **Quality-score distributions.** Comparison of the distributions of quality scores produced by the proposed eDifFIQA models in comparison with the baseline of DifFIQA.

θ, while the quality term is weighted using $(1-\theta)$. We show results for three different values of, θ i.e.: 0.1, 0.5 and 0.9. The experiments are performed using the AdaFace FR model over all benchmark datasets and the results are reported in Table 10.

As can be seen from the results, both hyperparameters have a clear impact on the performance of the final eDifFIQA model. The best performance is achieved when θ and ϵ equal 0.5, with minor

TABLE 8
Alignment sensitivity experiments using CNN-based FR models. 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. The best result among the baselines and proposed methods is marked with *.

AdaFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	1.023	0.992	0.970	0.970	0.987	1.004	0.991
SDD-FIQA [13]	0.722	0.824	0.480	0.731	0.641	0.796	0.699
PFE [18]	0.733	0.838	0.482	0.777	0.603	0.748	0.697
PCNet [11]	0.918	0.962	0.855	0.951	0.968	0.969	0.937
MagFace [19]	0.747	0.816	0.573	0.722	0.730	0.866	0.742
LightQNet [12]	1.015	0.965	1.033	1.018	1.041	0.987	1.010
SER-FIQ [14]	0.689	0.866	0.357	0.685	0.704	0.576†	0.646
FaceQAN [15]	0.748	0.899	0.442	0.681	0.593	0.589	0.659
CR-FIQA [21]	0.719	0.795	0.340	**0.647**	0.562	0.595	0.610
FaceQgen [32]	1.014	0.959	1.020	0.970	0.959	0.986	0.985
DifFIQA [5]	0.751	0.893	0.425	0.756	0.725	0.619	0.695
DifFIQA(R) [5]	**0.701***	0.890	0.346	0.654	0.609	**0.536***	0.623
eDifFIQA(S)	0.727	0.856	0.356	0.665	0.599	0.611	0.636
eDifFIQA(M)	0.741	**0.773***	0.341	0.652	0.559	0.583	0.608
eDifFIQA(L)	0.713	0.822	**0.334***	0.645 *	**0.525***	0.539	0.596
ArcFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	1.006	0.956	0.968	0.973	1.028	1.004	0.989
SDD-FIQA [13]	0.650	0.778	0.485	0.739	0.663	0.807	0.687
PFE [18]	0.655	0.772	0.486	0.748	0.619	0.721	0.667
PCNet [11]	0.905	0.937	0.887	0.941	0.953	0.976	0.933
MagFace [19]	0.680	0.736	0.523	0.700	0.696	0.887	0.704
LightQNet [12]	0.985	1.035	1.040	1.122	1.016	0.991	1.032
SER-FIQ [14]	0.635	0.843	0.353	0.672	0.686	0.633†	0.637
FaceQAN [15]	0.681	0.774	0.482	0.674	0.584	0.651	0.641
CR-FIQA [21]	0.663	**0.673**	0.349	**0.643**	0.553	0.632	0.585
FaceQgen [32]	0.999	0.969	1.005	1.011	0.949	0.964	0.983
DifFIQA [5]	0.661	0.808	0.514	0.808	0.693	0.742	0.704
DifFIQA(R) [5]	**0.632***	0.830	**0.335**	0.632 *	0.611	**0.597***	0.606
eDifFIQA(S)	0.665	0.773	0.361	0.664	0.618	0.647	0.621
eDifFIQA(M)	0.683	0.694 *	0.333	0.651	0.578	0.630	0.595
eDifFIQA(L)	0.644	0.750	0.330 *	0.645	**0.535***	0.616	0.587
CosFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	1.017	0.996	0.962	0.917	1.016	0.991	0.983
SDD-FIQA [13]	0.690	0.817	0.465	0.682	0.666	0.709	0.672
PFE [18]	0.702	0.829	0.471	0.754	0.653	0.658	0.678
PCNet [11]	0.921	0.973	0.875	0.863	0.969	0.952	0.925
MagFace [19]	0.723	0.812	0.580	0.664	0.741	0.814	0.723
LightQNet [12]	0.999	0.994	1.061	1.181	1.041	0.980	1.043
SER-FIQ [14]	0.666	0.865	0.327	0.629	0.724	0.522†	0.622
FaceQAN [15]	0.726	0.889	0.425	0.663	0.616	0.530	0.642
CR-FIQA [21]	0.690	0.781	0.317	**0.608**	0.590	0.547	0.589
FaceQgen [32]	1.009	0.986	1.004	1.017	0.961	0.947	0.987
DifFIQA [5]	0.717	0.881	0.397	0.738	0.757	0.625	0.686
DifFIQA(R) [5]	**0.675***	0.859	0.317	0.639	0.639	**0.494***	0.604
eDifFIQA(S)	0.695	0.844	0.316	0.633	0.630	0.546	0.611
eDifFIQA(M)	0.710	**0.758***	0.310	0.625 *	0.583	0.559	0.591
eDifFIQA(L)	0.683	0.795	**0.308***	0.635	**0.561***	0.515	0.583
CurricularFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	0.996	0.956	0.976	0.973	1.028	1.001	0.988
SDD-FIQA [13]	0.652	0.797	0.467	0.700	0.632	0.755	0.667
PFE [18]	0.651	0.808	0.450	0.718	0.584	0.785	0.666
PCNet [11]	0.905	0.965	0.904	0.953	0.945	0.965	0.940
MagFace [19]	0.677	0.785	0.518	0.689	0.693	0.865	0.705
LightQNet [12]	0.980	1.037	1.062	1.050	0.991	0.986	1.017
SER-FIQ [14]	0.633	0.841	0.365	0.655	0.677	0.665†	0.639
FaceQAN [15]	0.676	0.905	0.461	0.680	0.563	0.669	0.659
CR-FIQA [21]	0.660	**0.753**	0.333	**0.631**	0.541	0.693	0.602
FaceQgen [32]	0.989	0.951	1.004	1.009	0.961	0.969	0.981
DifFIQA [5]	0.670	0.898	0.496	0.805	0.700	0.760	0.721
DifFIQA(R) [5]	**0.632***	0.900	0.350	0.650	0.576	0.683	0.632
eDifFIQA(S)	0.658	0.827	0.358	0.651	0.579	0.704	0.630
eDifFIQA(M)	0.675	0.765 *	0.338	0.652	0.527	**0.658***	0.603
eDifFIQA(L)	0.639	0.811	**0.331***	0.636 *	**0.509***	0.695	0.603

†SER-FIQ was used to create XQLFW.

TABLE 9
Alignment sensitivity experiments using Transformer-based FR models. 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. The best result among the baselines and proposed methods is marked with *.

SwinFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	0.997	0.885	0.955	0.996	1.015	0.972	0.970
SDD-FIQA [13]	0.675	0.795	0.482	0.757	0.659	0.797	0.694
PFE [18]	0.687	0.794	0.487	0.761	0.590	0.688	0.668
PCNet [11]	0.901	0.952	0.904	0.928	0.963	0.979	0.938
MagFace [19]	0.699	0.761	0.538	0.770	0.676	0.868	0.719
LightQNet [12]	0.985	0.995	1.006	1.004	0.987	0.993	0.995
SER-FIQ [14]	**0.647**	0.838	0.373	0.732	0.689	0.635†	0.652
FaceQAN [15]	0.722	0.822	0.465	0.727	0.561	0.605	0.650
CR-FIQA [21]	0.678	0.719	0.361	0.714	**0.516**	0.598	0.598
FaceQgen [32]	0.999	0.938	1.014	0.978	0.958	0.988	0.979
DifFIQA [5]	0.663 *	0.785	0.552	0.814	0.693	0.651	0.693
DifFIQA(R) [5]	0.672	0.820	0.362	**0.699***	0.588	**0.571***	0.619
eDifFIQA(S)	0.688	0.767	0.376	0.724	0.585	0.609	0.625
eDifFIQA(M)	0.704	**0.708***	**0.353**	0.712	0.558	0.615	0.608
eDifFIQA(L)	0.676	0.739	0.352 *	0.705	0.517 *	0.578	0.595
TransFace - pAUC@FMR=10^{-3}(↓)							
FIQA Model	**Adience**	**CALFW**	**CFP-FP**	**CPLFW**	**LFW**	**XQLFW**	$\overline{pAUC}$
FaceQnet [10]	1.016	0.986	0.943	0.968	1.038	0.992	0.991
SDD-FIQA [13]	0.706	0.810	0.465	0.660	0.645	0.761	0.674
PFE [18]	0.729	0.823	0.488	0.733	0.631	0.734	0.689
PCNet [11]	0.920	0.967	0.884	0.957	0.969	0.952	0.942
MagFace [19]	0.736	0.794	0.583	0.644	0.726	0.872	0.726
LightQNet [12]	1.008	0.970	1.030	2.472	1.020	0.992	1.249
SER-FIQ [14]	**0.662**	0.880	0.331	0.613	0.712	0.522†	0.620
FaceQAN [15]	0.739	0.905	0.437	0.595	0.602	0.534	0.635
CR-FIQA [21]	0.697	0.782	0.328	**0.570**	0.572	0.503	0.575
FaceQgen [32]	1.009	0.979	1.013	1.239	0.960	0.962	1.027
DifFIQA [5]	0.708	0.880	0.407	0.684	0.737	0.588	0.667
DifFIQA(R) [5]	0.686 *	0.852	0.329	0.584	0.613	**0.458***	0.587
eDifFIQA(S)	0.719	0.829	0.333	0.583	0.598	0.530	0.599
eDifFIQA(M)	0.722	**0.752***	0.318	0.581	0.557	0.480	0.568
eDifFIQA(L)	0.692	0.775	**0.308***	0.578 *	**0.533***	0.468	0.559

†SER-FIQ was used to create XQLFW.

performance drops for other values of the two hyperparameters. Interestingly, the second-best result is achieved when changing the hyperparameter θ to focus more on the representation consistency term of the loss function, clearly showing the importance of identity information (and consistency) in the quality prediction process.

TABLE 10
Ablation study. We explore the sensitivity of the proposed eDifFIQA technique to the two hyperparameters ϵ and θ and compare the results to the baseline methods of DifFIQA and DifFIQA(R) using the AdaFace FR model. The best performing method for each dataset is presented in **bold**, the overall best result is colored green, the second best blue and the third best red.

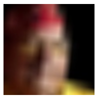
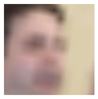
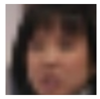
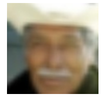
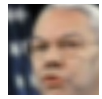
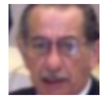
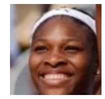
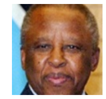
			Adience	**CALFW**	**CFP-FP**	**CPLFW**	**IJB-C**	**LFW**	**XQLFW**	$\overline{pAUC}$
eDifFIQA	$\theta = 0.5$	$\epsilon = 0.1$	0.876	0.919	0.452	0.693	0.778	**0.690**	0.699	0.730
		$\epsilon = 0.2$	0.873	0.884	0.448	0.690	0.771	0.750	0.698	0.731
		$\epsilon = 0.5$	0.871	0.872	0.438	0.687	0.766	0.735	**0.690**	0.723
		$\epsilon = 1.0$	0.905	**0.852**	**0.428**	**0.676**	0.792	0.712	0.766	0.733
	$\epsilon = 0.5$	$\theta = 0.1$	0.871	0.875	0.438	0.690	0.772	0.811	0.714	0.739
		$\theta = 0.5$	0.871	0.872	0.438	0.687	0.766	0.735	**0.690**	0.723
		$\theta = 0.9$	**0.864**	0.885	0.446	0.687	0.774	0.711	0.711	0.725
DifFIQA			0.881	0.922	0.460	0.695	0.785	0.754	0.701	0.743
DifFIQA(R)			0.879	0.901	0.460	0.690	**0.759**	0.737	0.701	0.732

5.6 Qualitative Evaluation

Last but not least, we present a qualitative evaluation of the proposed eDifFIQA approach in a comparitive analysis with DifFIQA and DifFIQA(R).

Analysis of Quality Score Distributions. We first look at how the quality-optimized knowledge distillation changes the distribution of quality scores on the seven different experimental datasets. The results of the analysis in Fig. 7 show that the distributions of the extended models are significantly different from the baseline DifFIQA approach. The baseline approach has a very limited range of predicted quality scores mostly limited to $[0.6, 1.0]$, This does not cause issues for most quality-based tasks, since we are mostly interested in the ranking of the image qualities.

TABLE 11
Illustration of the quality scores produced by the proposed FIQA techniques. The scores are compared between the baseline DifFIQA and proposed eDifFIQA techniques.

										
DifFIQA [5]	0.071	0.561	0.657	0.726	0.775	0.815	0.847	0.867	0.884	0.936
DifFIQA(R) [5]	0.007	0.448	0.452	0.659	0.684	0.760	0.837	0.845	0.845	0.942
eDifFIQA(S)	0.032	0.095	0.068	0.365	0.413	0.453	0.666	0.728	0.828	0.897
eDifFIQA(M)	0.000	0.176	0.056	0.469	0.539	0.575	0.734	0.775	0.826	0.870
eDifFIQA(L)	0.036	0.264	0.236	0.440	0.500	0.585	0.768	0.790	0.824	0.913

However, in the case where quality scores are used as weights to calculate average representations, like in the experiment shown in Section 5.3, we can observe a notable drop in performance when compared to other methods. The reason is that images of relatively low quality still obtain a relatively high quality score and in the end contribute significantly to the average representation. From this standpoint the distributions of the extended approach, can be seen as much more valuable as the range of quality scores is much wider. The benefits of this can also be seen from the results in Section 5.3, where the extended version outperforms the baseline.

Analysis of Image Rankings. In Table 11, we analyze the differences in quality scores per chosen images. The presented images are all part of the XQLFW dataset, and have been chosen by maximizing the differences in quality scores for the baseline DifFIQA approach. The chosen images are ranked by their respective quality scores, alongside the images we also report the quality scores calculated using both the baselines and all proposed eDifFIQA variants. The reported results tell a similar story as the results of the quality score distribution analysis. The quality scores of all eDifFIQA methods cover a wider range of values and are better distributed over the whole range of $[0, 1]$. The ranking is for the most part consistent between all presented methods, with some individual changes in the ranking.

6 Conclusion

We have presented DifFIQA, a novel unsupervised FIQA technique based on denoising diffusion probabilistic models and the extended eDifFIQA supervised approach. The base DifFIQA uses the forward and backward processes of diffusion models to estimate the quality of input samples. A common issue of diffusion models and consequentially of DifFIQA is their relatively high computational complexity, stemming from the iterative nature of the forward and backward diffusion processes. The extended approach improves on the high computational complexity of DifFIQA, as well as its performance, by employing a knowledge distillation process. In this process quality labels extracted using DifFIQA are first optimized using additional sources of quality information and then used in the knowledge distillation process. During the distillation process a model consisting of a feature extraction backbone and a quality-regression MLP head is trained to predict the optimized quality scores. By training models with varying sizes of the feature extraction backbones we are able to control the performance/complexity trade-off in the final trained eDifFIQA model. Through comprehensive experiments on multiple datasets, we showed that the extended eDifFIQA models outperform the baseline DifFIQA techniques and achieve state-of-the-art performance in almost all tested scenarios. The optimization is shown to not only improve the predictive capabilites of the end model but also improve the characteristics of the quality distributions produced by the model. We presented three variants of eDifFIQA based on different sizes of ResNet backbones, and shown that using ResNet18 significantly lowers the runtime, without a significant decrease in performance. As part of our future work, we plan to further extend the model toward non-scalar quality predictions and explicit identification of the main image characteristics governing the predicted quality score.

Acknowledgements

The research presented in this work was supported in parts by the ARIS junior researcher program, the ARIS Research Programmes P2-0250 "Metrology and Biometric Systems" and P2-0214 "Computer Vision", and the ARIS Research Project J2-50069 "MIXBAI".

References

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

[2] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A Comprehensive Study on Face Recognition Biases Beyond Demographics," *Transactions on Technology and Society (TTS)*, vol. 3, no. 1, pp. 16–30, 2021.

[3] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *IET Biometrics*, vol. 7, no. 1, pp. 81–89, 2018.

[4] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face Image Quality Assessment: A Literature Survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–49, 2022.

[5] Ž. Babnik, P. Peer, and V. Štruc, "DifFIQA: Face Image Quality Assessment Using Denoising Diffusion Probabilistic Models," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2023.

[6] Z. Babnik and V. Štruc, "Assessing bias in face image quality assessment," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1037–1041.

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

[8] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.

[9] S. Ling, P. Le Callet, and Z. Yu, "The Role of Structure and Textural Information in Image Utility and Quality Assessment Tasks," *Journal of Perceptual Imaging (JPI)*, vol. 1, pp. 1–13, 2018.

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

[11] W. Xie, J. Byrne, and A. Zisserman, "Inducing Predictive Uncertainty Estimation for Face Verification," in *British Machine Vision Conference (BMVC)*, 2020.

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

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

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

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

[16] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of Face Image Sample Quality," in *Proceedings of the IAPR International Conference on Biometrics (ICB)*. Springer, 2007, pp. 242–251.

[17] K. Nasrollahi and T. B. Moeslund, "Extracting a Good Quality Frontal Face Image From a Low-Resolution Video Sequence," *Transactions on Circuits and Systems for Video Technology (TCSVT)*, vol. 21, no. 10, pp. 1353–1362, 2011.

[18] Y. Shi and A. K. Jain, "Probabilistic Face Embeddings," in *Proceedings of the CVF/IEEE International Conference on Computer Vision (CVPR)*, 2019, pp. 6902–6911.

[19] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "MagFace: A Universal Representation for Face Recognition and Quality Assessment," in *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 225–14 234.

[20] P. Terhörst, M. Ihlefeld, M. Huber, N. Damer, F. Kirchbuchner, K. Raja, and A. Kuijper, "QMagFace: Simple and Accurate Quality-Aware Face Recognition," in *Proceedings of the CVF/IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 3484–3494.

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

[22] M. Kim, A. K. Jain, and X. Liu, "AdaFace: Quality Adaptive Margin for Face Recognition," in *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 750–18 759.

[23] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters (SPL)*, vol. 20, no. 3, pp. 209–212, 2012.

[25] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-Reference Image Quality Assessment in the Spatial Domain," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 12, pp. 4695–4708, 2012.

[26] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from Rankings for No-Reference Image Quality Assessment," in *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1040–1049.

[27] H.-I. Kim, S. H. Lee, and Y. M. Ro, "Investigating Cascaded Face Quality Assessment for Practical Face Recognition System," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, 2014, pp. 399–400.

[28] O. Henniger, B. Fu, and C. 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)*, 2020, pp. 1–5.

[29] L. Best-Rowden and A. K. Jain, "Learning Face Image Quality from Human Assessments," *Transactions on Information Forensics and Security (TIFS)*, vol. 13, no. 12, pp. 3064–3077, 2018.

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

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

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

[33] Ž. Babnik, N. Damer, and V. Štruc, "Optimization-based improvement of face image quality assessment techniques," in *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, 2023, pp. 1–6.

[34] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion Models in Vision: A survey," *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[35] M. Ivanovska and V. Štruc, "Face Morphing Attack Detection with Denoising Diffusion Probabilistic Models," in *Proceedings of the International Workshop on Biometrics and Forensics (IWBF)*, 2023, pp. 1–6.

[36] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

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

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

[41] E. Eidinger, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," *Transactions on Information Forensics and Security (TIFS)*, vol. 9, no. 12, pp. 2170–2179, 2014.

[42] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney *et al.*, "IARPA Janus Benchmark-C: Face Dataset and Protocol," in *Proceedings of the International Conference on Biometrics (ICB)*, 2018, pp. 158–165.

[43] T. Zheng, W. Deng, and J. Hu, "Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments," *CoRR*, vol. abs/1708.08197, 2017. [Online]. Available: http://arxiv.org/abs/1708.08197

[44] 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.

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

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

[47] L. Wolf, T. Hassner, and I. Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," in *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 529–534.

[48] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. 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)*, 2018, pp. 5265–5274.

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

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

[51] J. Dan, Y. Liu, H. Xie, J. Deng, H. Xie, X. Xie, and B. Sun, "TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective," in *Proceedings of the CVF/IEEE International Conference on Computer Vision (ICCV)*, 2023, pp. 20 642–20 653.

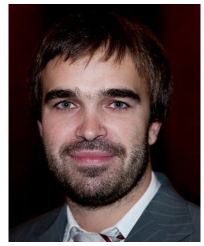
[52] T. Schlett, C. Rathgeb, J. Tapia, and C. Busch, "Considerations on the Evaluation of Biometric Quality Assessment Algorithms," *arXiv preprint arXiv:2303.13294*, 2023.

[53] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A Dataset for Recognising Faces Across Pose and Age," in *Proceedings*

of the IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2018, pp. 67–74.
[54] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proceedings of the CVF/IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212.
[55] J. Xiang and G. Zhu, "Joint Face Detection and Facial Expression Recognition with MTCNN," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2017, pp. 424–427.
[56] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Proceedings of the CVF/IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

Žiga Babnik received his bachelor's degree at the University of Ljubljana, Faculty of Computer Science in 2019 – "Song similarity analysis based on lyrics, ratings and meta data". In 2021, he received his Masters's degree at the University of Ljubljana, Faculty of Computer Science – "Face Deidentification using Face Swapping". Msc. Žiga Babnik is currently employed at the Laboratory for Machine Intelligence (LMI), as a young researcher. The main topics of his research revolve around deep learning, computer vision and biometrics, he is currently active in the research fields of face recognition and face image quality assessment.

Peter Peer is a Full Professor at the Faculty of Computer and Information Science, University of Ljubljana (UL), Slovenia, where he heads the Computer Vision Laboratory, is a member of the Faculty senate, a member of the Faculty management board, and coordinates the double degree study program with the Kyungpook National University (KNU), South Korea. He received his PhD in computer science from UL in 2003. Within his post-doctorate he was an invited researcher at CEIT, San Sebastian, Spain. He teaches courses on Operating Systems, Game Technology and Virtual Reality, and Image-based Biometrics. His research interests focus on biometrics and computer vision. He participated in several national and EU funded R&D projects and published over 120 research papers in leading international peer reviewed journals and conferences. He is a co-organizer of a number of competitions at top-tier biometrics conferences. He serves as an executive editor at ICT Express, as an associated editor of IET Biometrics, IEEE Access, and as a topical advisory panel member of MDPI Applied Sciences. He is a member of the EAB, IAPR and IEEE. Each year he reviews for top-tier conferences. In the past he also served as a chairman of the Slovenian IEEE Computer chapter for four years, guest edited a number of special issues in different SCI journals, was a guest professor at the North-Caucasus Federal University and KNU, and served at the Faculty as a vice-dean for economic affairs for four years.



Vitomir Štruc is a Full Professor at the University of Ljubljana, Slovenia. His research interests include problems related to biometrics, computer vision, image processing, and machine learning. He (co-)authored more than 150 research papers for leading international peer reviewed journals and conferences in these and related areas. Vitomir is a Senior Area Editor for the IEEE Transactions on Information Forensics and Security, a Subject Editor for Elsevier's Signal Processing and an Associate Editor for Pattern Recognition, and IET Biometrics. He serves regularly on the organizing committees of visible international conferences, including IJCB, FG, WACV and CVPR. He currently acts as a General Chair for IJCB 2023, a Program Chair for IEEE Face and Gesture 2024, a Tutorial Chair for CVPR 2024 and Program Co-Chair for WACV 2025. Dr. Struc is a Senior member of the IEEE, a member of IAPR, EURASIP, Slovenia's ambassador for the European Association for Biometrics (EAB) and the former president and current executive committee member of the Slovenian Pattern Recognition Society, the Slovenian member of IAPR. Vitomir is also the current VP Technical Activities for the IEEE Biometrics Council, the secretary of the IAPR Technical Committee on Biometrics (TC4) and a member of the Supervisory Board of the EAB.