Deep Tiny Network for Recognition-Oriented Face Image Quality Assessment[⋆]

Baoyun Peng[1][0000−0001−5240−1779], Min Liu[1,2][0000−0003−3086−315X](✉), Zhaoning Zhang[3][0000−0001−7518−1385], Kai Xu[3][0000−0002−9054−0216], and Dongsheng Li[3][0000−0001−9743−2034]

[1] Academy of Military Science, Beijing 100097, People's Republic of China
pengbaoyun13@alumni.nudt.edu.cn,gfsliumin@gmail.com
[2] Intelligent Game and Decision Laboratory, Beijing 100071, People's Republic of China
gfsliumin@gmail.com
[3] National University of Defense Technology, Changsha 410073, People's Republic of China
{zhangzhaoning,kevin.kai.xu,dsli}@nudt.edu.cn

Abstract. Face recognition has made significant progress in recent years due to deep convolutional neural networks (CNN). In many face recognition (FR) scenarios, face images are acquired from a sequence with huge intra-variations. These intra-variations, which are mainly affected by low-quality face images, cause instability of recognition performance. Previous works have focused on ad-hoc methods to select frames from a video or use face image quality assessment (FIQA) methods, which consider only a particular or combination of several distortions. In this work, we present an efficient non-reference image quality assessment for FR that directly links image quality assessment (IQA) and FR. More specifically, we propose a new measurement to evaluate image quality without any reference. Based on the proposed quality measurement, we propose a deep Tiny Face Quality network (tinyFQnet) to learn a quality prediction function from data. We evaluate the proposed method for different powerful FR models on two classical video-based (or template-based) benchmarks: IJB-B and YTF. Extensive experiments show that, although the tinyFQnet is much smaller than the others, the proposed method outperforms state-of-the-art quality assessment methods in terms of effectiveness and efficiency.

**Keywords:** Face image quality assessment · Lightweight network · Face recognition.

1 Introduction

The performance of face recognition (FR) has been considerably improved in recent years, mainly owing to the combination of deep neural networks and large-scale labeled face images. On the representative academic benchmarks IJB-C [32] and IQIYI-VID [31], several FR methods [2, 9, 22] have even surpassed humans in terms of face verification. However, in many real scenarios where face images are captured as a

---

sequence with high uncertainty, the FR performance degrades sharply due to the sequence's low-quality face images. Figure 1 shows such a typical scenario in which the input of the FR model is a sequence of images. In such a sequence, the difficulty of recognizing the person in various face images is different. Usually, images with a high-resolution, neural head pose, and little distortions (occlusion, blur, or noises), are easier to recognize. For the sake of throughput capacity, the FR system may only sample one or a few images from a sequence to recognize. Without a good selection method, the FR system may fail to recognize when selecting a low-quality image that is hard to recognize. Face image quality assessment (FIQA) can correctly assess face images' quality to improve an FR system's performance.
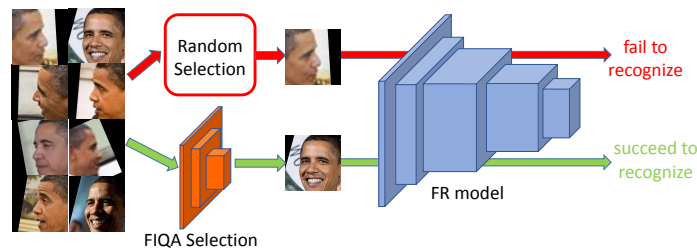

Fig. 1. A typical selection process. Random selection may select a low-quality image that is hard for the FR model to recognize. A good FIQA selection method can improve the performance of recognition.

Evaluating the influence of face image quality on an FR system's performance is non-trivial since the performance is affected by many underlying variations, and there is no unified definition or standard metric on face image quality. Several efforts have been made to develop common standards [15, 38]. In these standards, the factors that influence the face image quality can be categorized into perceptual variations and biometric variations. Many prior works have contributed to automatic FIQA methods [1, 4, 53], and these works are similar to general image quality assessment that relies on either the difference among several known properties of the human visual system between target and reference images [13] or the degradation in structural information of face images [45]. However, considering one or several particular factors may result in unsuitable face image selection since the face image quality is influenced by many potential factors.

Several new learning-based FIQA methods have been proposed in recent years that attempt to address limitations of earlier techniques through automatically learning how to assess face image quality from amounts of data [5, 19, 33, 42]. For instance, [44] uses the matching score between a face image and a reference image as a quality score based on hand-crafted features, while [19] and [33] are based on features extracted from a well-performed FR model. Both of them need to select an image as a reference for each class to evaluate quality scores before training their models. For non-reference FIQA, [23] proposes pixel-level face image quality assessment that measures

face recognition performance changes when inserting or deleting pixels based on their predicted quality. [7] uses knowledge distillation on difficult samples to train an efficient lightweight FIQA network focused on the quality classification boundary. Besides, there are also several non-reference IQA methods [3, 5] that use GAN or CNN to assess face image quality based on noise exploration or sample relative classifiability.
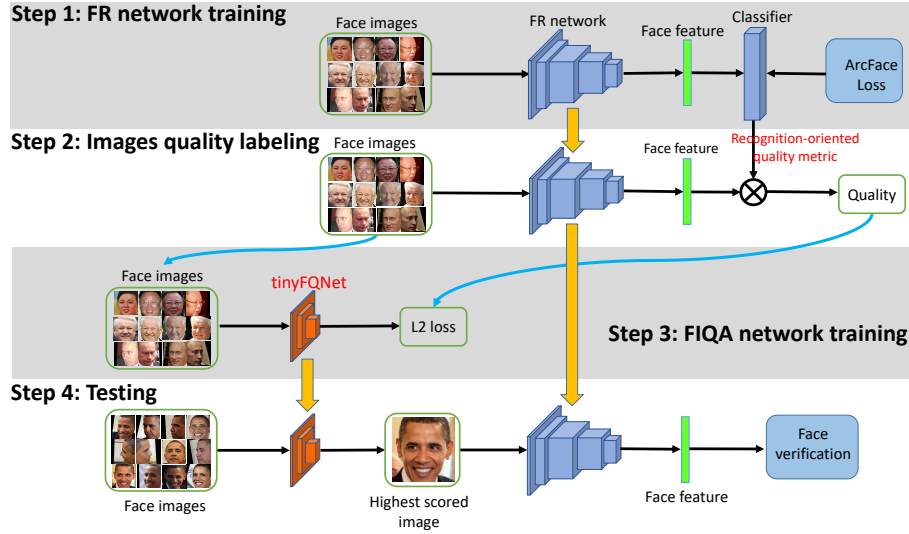
Fig. 2. The pipeline for video-based FR. The base set contains multiple identities, and each identity is with several images. In common, the frames are captured under significant variations, such as large head pose, illumination, motion blur, and occlusion. Our method consists of four steps. In Step 1, we train the FR network. Then, the trained FR network is used to label face images with quality. We use L2 regression loss in Step 3 to train a FIQA network. For testing in Step 4, we first use our trained FIQA network to select high-quality face images and feed them to our trained FR network to extract features, then the extracted feature will be used for face verification.

Unlike the above FIQA methods, we propose an efficient deep image quality assessment method for face recognition in this paper. More specifically, we propose a novel non-reference quality measurement for face image quality scoring that directly links Face IQA with FR without a reference. The details of the proposed quality metric are presented in section 3.1. Using the proposed quality measurement, we can generate amounts of training data with quality labels in a fully automatic way. Besides, previous methods are usually based on complicated networks (e.g., ResNet-50 in [54] and VGG-16 in [19]) while training them on a relatively small dataset. They may ignore that computation and memory costs are important when applying face IQA to a real FR system. However, the computation or memory cost of FIQA is truly a key factor when designing a resource-limited FR system, such as face unlock on mobile phones and other embedded devices. Based on this consideration, we propose a tiny but effective

FIQA network that only has **21.8k** parameters, and the average time of processing an image is only 4ms on Samsung S10.

To verify the effectiveness of the proposed method, the performance evaluation of the FR model when combined with face IQA as a plug-in is carried out. We provide extensive analysis of the impact on the performance of the FR model in terms of different network sizes and show that even a tiny network can achieve comparable performance with much more complicated networks on face image quality assessment. To evaluate the influence of different factors on the FR model's performance, we provide extensive experiments on three perceptual factors: head pose, blurriness, and JPEG compression rate. We also provide a comprehensive comparison among perceptual-based and learning-based FIQA methods and compare the proposed method with several other FIQA methods.

The proposed method can also be used to generate video representation by quality-weighted averages like [39, 49], but this is not our scope in this work. We evaluate the proposed method on three common datasets, including the IARPA Janus Benchmark-B (IJB-B) [46], IARPA Janus Benchmark-C (IJB-C) [32], and Youtube Faces Database (YTF) [47]. The results show that our proposed method can improve the baseline FR performance significantly and is also superior to other FIQA methods. The contributions of this paper are summarized as follows:

- we propose a new quality measurement that directly targets face recognition for face image quality assessment (FIQA), and no reference is required. Using the proposed quality measurement, we can automatically generate amounts of data with quality labels to train the IQA model;
- Based on our quality measurement, we propose an efficient and effective deep network as a face IQA model and demonstrate the effectiveness of the proposed method on the unconstrained IJB-B dataset and YTF dataset;
- we provide extensive analysis about the influence of several perceptual factors on FR's performance, and a comprehensive comparison between perceptual-based and learning-based FIQA methods;
- we provide extensive analysis on the influence of FIQA network size on FR's performance, and show that even a tiny network with only 21.8K parameters can achieve comparable performance with a much more complicated network with 23M parameters;
- we provide extensive analysis on the influence of data size and quality score distribution of training dataset on FR's performance, and provide an effective sampling method to make the training dataset more balanced, which is proved that it can further improve FR's performance;

The rest of this paper is organized as follows. In Section 2, we review related works on FR and FIQA. In Section 3, we present the details of the proposed FIQA method, including a novel recognition-oriented quality metric in section 3.1, a tiny network for FIQA in section 3.2, and an easy and efficient method that can quickly generate amounts of labeled data in section 3.3. In Section 4, evaluations on IJB-B and YTF datasets are presented. Finally, the concluding remarks are presented in Section 5.

2   Related Work

This work mainly focuses on face image quality assessment (FIQA) for face recognition. Many of previous works are extended from general image quality assessment (IQA). Therefore, we first discuss the related works of general IQA, then face recognition and FIQA.

2.1   Image Quality Assessment

Image quality assessment methods are mainly categorized into full-reference methods and non-reference (or blind) methods. The former needs full access to the reference images [11, 24, 41]. Several full-reference IQA methods are based on human visual systems [11, 26], which predict quality scores from visible image differences. Changes of structure in a distorted image are also taken to measure the quality in [37, 45]. Non-reference IQA methods require no or limited information about the reference image. These methods try to detect a particular or several distortions, such as blurriness [14, 17], JPEG compression [40, 51], noise [8], and combination of several distortions [12, 34]. Recently, several methods [27, 28, 52] have adopted a deep convolutional neural network (CNN) to predict image quality on training datasets with quality labels. Note that the ultimate goal of these quality assessment measurements is for human perceptual cognition rather than face recognition.

2.2   Face Image Quality Assessment

Existing methods for face image quality assessment are mainly based on the similarity to a reference (ideal) image, and these methods measure the face image quality by comparing several known properties between target images and reference images. Specifically, perceptual image quality (such as contrast, resolution, sharpness, and noise) and biometric quality (such as pose, illumination, and occlusion) have been used to evaluate face image quality [1, 50]. Prior works [35] proposed a weighted quality fusion approach that merged the weights of factors (rotation, sharpness, brightness, and resolution) into a quality score. Wong et al. [48] predicted a face image quality score by determining its probabilistic similarity to an ideal face image via local patch-based analysis for FR in a video.

Another way for FIQA is to leverage learning-based approaches [4, 6, 44]. Unlike conventional approaches that measure the quality by analyzing pre-defined biometric and perceptual image characteristics, learning-based methods learn a prediction function from amounts of face images with quality scores. Hence, the learning process is highly dependent on the training dataset. To assess a face image's quality, [21, 25] used a discrete value to indicate matching results. Best-Rowden et al. [4] obtained a training dataset through pairwise comparisons performed by workers from Amazon Mechanical Turk. They used a support vector machine to predict the quality score using features extracted from a face image. Chen et al. [6] learned a ranking function for quality scores by dividing datasets according to quality. These methods aim to establish a function from image features to quality scores. Besides, these learning-based methods use either hand-crafted features or features extracted from pre-trained recognition models

as inputs to learn the prediction function and train the FIQA model on small datasets collected in laboratories (e.g., FRGC [36]). In [42], deep CNN is used to determine the category and degree of degradation in a face image by considering five perceptual image characteristics (resolution, blurriness, additive white Gaussian noise, salt-and-pepper noise, and Poisson noise). In [54] and [19], the face recognition model is also used to generate quality labels, and the deep CNN model is learned from labeled data to predict quality scores. FaceQAN [3] links face image quality with adversarial examples learned by gradient descent information from FR model. Recently, [5] proposes CR-FIQA that estimates the face image quality by predicting its relative classifiability using the sample feature representation with respect to its class center and the nearest negative class center.

3   Method

In this section, we describe the details of our proposed FIQA method, including a recognition-oriented non-reference quality metric, a tiny but efficient deep network (tinyFQnet), a simple technique to quickly generate amounts of quality-labeled data that can be used to train the tinyFQnet, and a data sampling method to make the distribution of scores in the training dataset more balanced.

3.1   Recognition-oriented Non-Reference Quality Measurement

Empirically, focusing on one particular or a combination of several factors (blurriness, head pose, noise, etc.) may not be effective since FR is affected by amounts of variations, and some of these variations are not quantifiable or hand-crafted. Besides, it is hard to determine the relative importance of each factor. We argue that the QA model should be linked with the face recognition process directly.
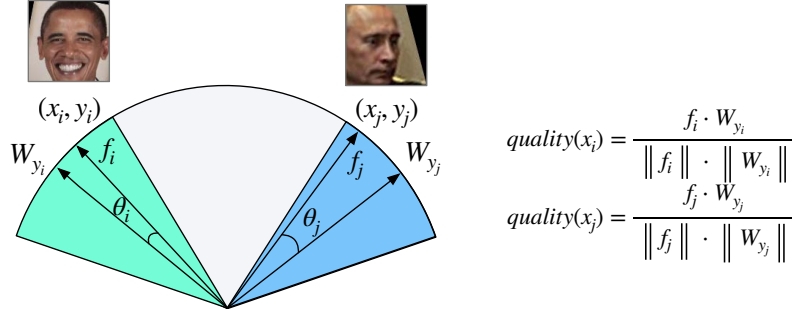
Inspired by [19], of which the quality score of a given image is defined as the normalized Euclidean distance between it and a reference image of the same subject, we propose a recognition-oriented non-reference quality measurement that directly links quality assessment with face recognition. More specifically, given an image x and its label y, the quality score of x is defined as the cosine similarity between x and the distribution center of the same class with x in angular space, as follows:

$$quality(x) = cosine(\boldsymbol{f_x}, \boldsymbol{u_y}) = \frac{\boldsymbol{f_x} \cdot \boldsymbol{u_y}}{\|\boldsymbol{f_x}\| \, \|\boldsymbol{u_y}\|} \tag{1}$$

where $\boldsymbol{f_x}$ is the feature vector of image x extracted by the FR model, $\boldsymbol{u_y}$ is the center feature vector of y class. Compared with using a hand-selected image [54] or the highest-quality image selected by a commercial system [19] as the reference to compute image quality, the proposed method in Equation 1 is more reasonable in the following aspect: $cosine(\boldsymbol{f_x}, \boldsymbol{u_y})$ directly reflects how difficult for the FR model to recognize the image x correctly. The lower the $cosine(\boldsymbol{f_x}, \boldsymbol{u_y})$ is, the more difficult the FR model to recognize x correctly.

Although the center $\boldsymbol{u_y}$ is hard to obtain due to the ground-truth distribution of a given class being unknown for a particular FR model, we can regard the last fully

connected layer output of an FR model as the class center in the angular space as in [29, 30].

$$quality(x_i) = \frac{f_i \cdot W_{y_i}}{\| f_i \| \cdot \| W_{y_i} \|}$$

$$quality(x_j) = \frac{f_j \cdot W_{y_j}}{\| f_j \| \cdot \| W_{y_j} \|}$$

Fig. 3. Geometry interpretation of recognition-oriented quality metric in a 2D feature embedding space. $\boldsymbol{W}_{y_i}$ and $\boldsymbol{W}_{y_j}$ are the centers of y_i and y_j class, $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ is the feature vector of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, θ_i and θ_j is the angle between $\boldsymbol{f}_i$ & $\boldsymbol{W}_{y_i}$ and $\boldsymbol{f}_j$ & $\boldsymbol{W}_{y_j}$, respectively.

Figure 3 shows the geometry interpretation of the 2D feature and two samples $(\boldsymbol{x}_i, y_i)$ and $(\boldsymbol{x}_j, y_j)$ with different difficulties to recognize. θ_j is bigger than θ_i, which means that $\boldsymbol{x}_j$ is harder to recognize compared to $\boldsymbol{x}_i$. Consequently, the quality score of $\boldsymbol{x}_j$ is smaller than $\boldsymbol{x}_i$.

3.2 Tiny Face Quality Network

The primary purpose of learning-based FIQA is to provide a prediction function for the face image quality score. Unlike previous learning-based methods that access quality scores via deep or hand-crafted features, we adopt an end-to-end method to directly predict the quality score of a raw image through deep CNN. The reason is that a deep CNN can learn more relative features from raw images compared with features extracted from pre-trained models.

We noticed a few works that adopt an end-to-end deep CNN [19, 54] to predict face image quality. The core difference is that the proposed method in this work doesn't need to pick an image as a reference using other methods before generating quality scores, while [19] needs to use ICAO compliance software to pick an image with the highest score and [54] need to select an image as a reference through human perceptual vision. Given this point, both [19] and [54] are more like reference-based methods, while our method is more similar to non-reference methods.Besides, these works didn't take the memory and computation costs into account and used very complex networks (ResNet-50 in [54] and VGG-16 in [19]) as face quality networks. However, both computation and memory costs are key factors when one needs to apply FIQA to real FR scenarios.

Table 1. Our tinyFQnet architecture. Each conv layer is followed by a batch norm layer and a ReLU layer. Block1 and Block2 are shown in Fig 4.

layer name	output size	tinyFQnet
input	$64 \times 64 \times 3$	
conv	$32 \times 32 \times 11$	$3 \times 3 \times 11$,stride=2
Block1	32x32x2	$1 \times 1, 8$ $[3 \times 3, 8] \times 1$, stride=1 $1 \times 1, 2$
Block2	$16 \times 16 \times 5$	$1 \times 1, 8$ $[3 \times 3, 8] \times 1$, stride=2 $1 \times 1, 5$
Block1	$16 \times 16 \times 5$	$1 \times 1, 20$ $[3 \times 3, 20] \times 1$, stride=1 $1 \times 1, 5$
Block2	$8 \times 8 \times 11$	$1 \times 1, 20$ $[3 \times 3, 20] \times 1$, stride=2 $1 \times 1, 11$
Block1	$8 \times 8 \times 11$	$1 \times 1, 44$ $[3 \times 3, 44] \times 2$, stride=1 $1 \times 1, 11$
Block1	$8 \times 8 \times 22$	$1 \times 1 \times 44$ $[3 \times 3 \times 44]$x1, stride=1 1x1x22
conv	$8 \times 8 \times 256$	$1 \times 1 \times 256$, stride=1
avgpool	256	
fc	1	1×256

In this work, we design a Tiny Face Quality Assessment Network (tinyFQnet) that is tiny but efficient for FIQA. The design of the tinyFQnet is highly modularized by following MobileNetV2 [20]. It stacks two kinds of blocks shown in Figure 4, both of which share the same topology but are only different in whether there is a residual connection. Each block has three convolutional layers, and each convolutional layer is followed by a batch normalization layer and a ReLU layer. Both blocks share the same hyper-parameters (such as filter sizes, padding, bias, etc.) except for the input channels and output channels.

Table 1 shows the details of the architecture of tinyFQnet, noted as Q. There are 7 blocks in Q, including 5 non-residual blocks and 2 residual blocks in Q. The output of Q will be normalized to (0,1) by a sigmoid layer. Q only has 2.356 Mflops and 21.806k parameters, while ResNet-50 has 3.898 Gflops and 22.421M parameters, and VGG-16 has 14.528 Gflops and 128.041M parameters. Compared with ResNet-50 and VGG-16 used in previous methods, our tinyFQnet has a significant advantage in memory and

computation costs. Thanks to the low computation cost, our tinyFQnet can significantly speed up FR systems.

(a)   (b)
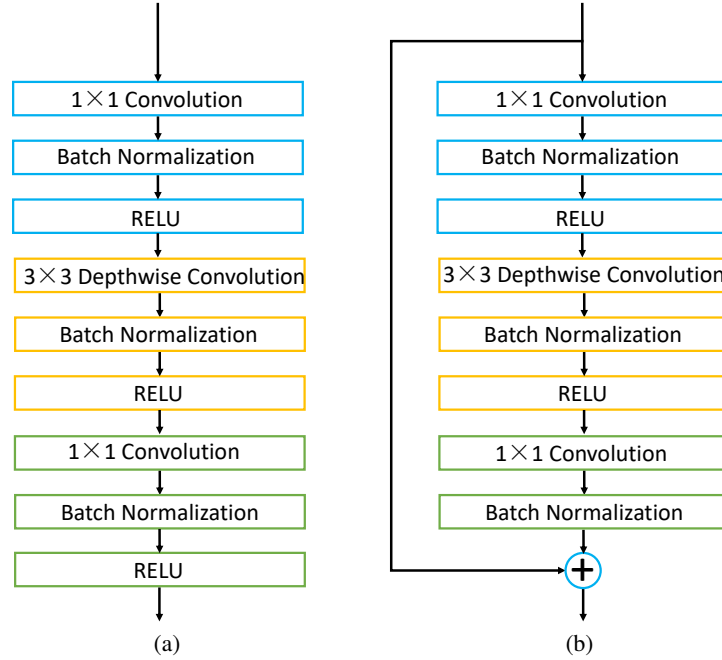
Fig. 4. The details of two basic blocks in tinyFQNet. (a): non-residual block; (b): residual block.

To cooperate with the FR network, we use RetinaFace [10] to detect and align images. Then, the aligned images are resized to 64×64 which is the same as tinyFQnet's input size. L2 regression loss in Equation 2 is adopted to train the tinyFQnet as follows:

$$L = \frac{1}{N} \sum_{i=1}^{N} \|\phi(x_i, \theta) - q_i\|_2 , \qquad (2)$$

where $\phi(x, \theta)$ denotes the prediction model with parameter θ. Stochastic gradient descent is used to train the network. The learning rate is initialized to 0.01 and degrades by 0.1 at every 5 epochs. The training batch size is set as 1024, and the weight decay is set as 0.0001. The training takes 17 epochs in total.

3.3 Generating training dataset with quality label

Usually, learning-based methods require amounts of labeled data to learn the desired function. Previous learning-based FIQA methods tried to train a prediction model from small lab-collected datasets (e.g., SCFace, FRGC, GBU, and Multi-PIE) due to the lack of a large number of face images that are labeled with quality scores. The lack

of training data limits the capability of deep learning models. Similar to [54] and [19], we adopt a feasible but fast method to generate a large number of labeled data. More specifically, since the proposed quality metric only involves a well-trained FR model and a training dataset, we can automatically generate amounts of data with quality labels and use labeled data to train our tinyFQnet.

The whole process of generating the labeled dataset for training the tinyFQnet is shown as follows:

Algorithm 1 Quality label generating process

Train model Q on $\mathcal{D}$;
Extract weights $\mathcal{W}$ of the classifier layer in Q;
Extract features $\mathcal{F}$ of training data $\mathcal{D}$ using Q;
Compute quality scores of $\mathcal{D}$ using Equation 1;

The first step is to train the FR model Q. Usually, we can use an existing well-trained FR model and its training dataset. Then, we extract the weights $\mathcal{W}$ of the classifier and all training data features from the FR model. Finally, the quality of the image can be computed with Equation 1.

3.4 Data sampling and augmentation strategy for balancing the distribution of scores

Although we can generate amounts of data with the quality scores through the method above, we found that the quality score distribution is highly unbalanced. The left sub-figure in Figure 5 shows the histogram distribution of quality scores generated by Section 3.3. It can be observed that most images are concentrated in a narrow range of scores. The fact, that the proposed quality metric is directly related to the cosine similarity between the image and its class centers in embedding space, means that the lower the training loss of the FR model is, the closer the features of the intra-class are. Consequently, a well-performed FR model will lead to an unbalanced distribution of quality scores, which will mislead the optimized results of the tinyFQnet.

We adopt a data sampling strategy that samples the data by identity and scores. To be more specific, we first filter those identities of which the number of related images is less than a threshold (in our experiments, we set the threshold to be 100 for Ms-Celeb-1M dataset). By doing so, we get nearly 7,900 identities and 860 thousand images. Then we divide images into 100 bins in a range from 0 to 1.0 according to quality scores with equal intervals. We over-sample those images of which quality scores range in the lowest 10% and the highest 5%, and downsample the remaining 80% of images at each bin respectively (the sampling rate is dependent on the number of images in a bin). After applying identities filtering and scores sampling strategy, the distribution of quality scores in the training dataset is more balanced than before, as shown in the right subfigure of Figure 5.
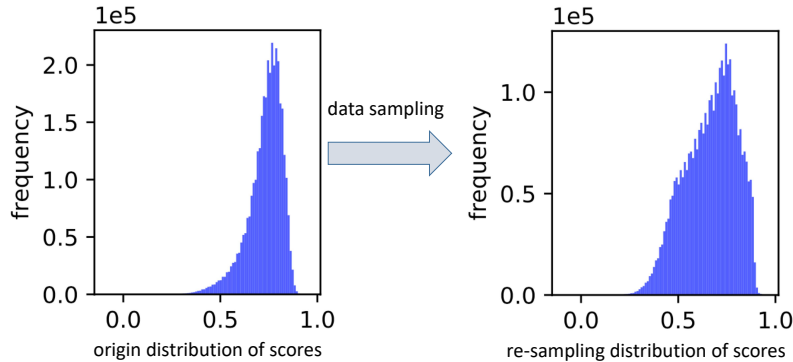
Fig. 5. Distribution of quality scores under different data sizes, including large size(left column), middle size(middle column), and small size(right column). The dataset is Ms-Celeb-1M [16], and the FR model is R50. The top row shows the results of a random sampling strategy, and the bottom row shows the results of a smooth sampling strategy.

4   Experimental Results

4.1   Experimental Setup

To evaluate the effectiveness of our FIQA on real FR systems, we consider the tiny recognition-oriented quality network (tinyRQNet) as a component independent of FR. We only use tinyRQNet to select the appropriate frames from a video. More specifically, we use the tinyRQNet to predict all the images in a video or a template (a template consists of several images and videos belonging to the same subject in the IJB-B dataset) and select the image with the highest quality score. Then, we evaluate those selected images using the FR model.

We compare the proposed method with three learning-based methods, including SVM [4], RQS [6] and FaceQnet [19]. RQS learns to predict and rank the quality scores from hand-crafted features. SVM [4] uses a support vector machine to predict the quality scores with features extracted by a deep FR model. Similar to our method, FaceQnet adopts deep CNN to predict quality scores from raw images.Beyond these comparisons, we also present the results of the most common perceptual factors, including blurriness, JPEG compression, head pose, and their combination. All the quality scores predicted by each method are rescaled to a range from 0 to 100, of which a higher score means higher quality.

We use two classical networks, including a complicated network ResNet-50 [18] and a lightweight network EfficientNet-b0 [43], to train the FR model on MS-Celeb-1M [16] separately. ResNet-50 is widely used in many tasks as a baseline model due to its powerful capacity, while EfficientNet-b0 achieves considerably better accuracy compared to classical MobileNet-V2 [20] while using fewer parameters and flops.All the images are aligned using 5 landmarks provided by RetinaFace [10] and cropped to 112×112 resolution. For all learning-based methods, we keep the same settings as the original papers. All the training is carried out on 8 TiTANX GPUs. Note that we

focus on face image quality assessment in this work, rather than learning video-level representations for face recognition. One can apply the proposed method to video-level face recognition, but this is beyond the scope of our work.
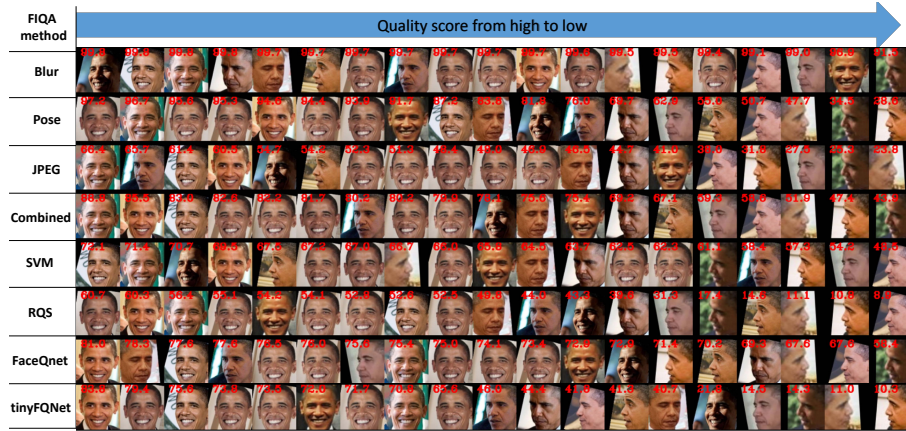
Fig. 6. Some face images randomly selected from one subject in the IJB-B dataset. We rank them with quality scores predicted by several FIQA methods.

4.2   Datasets and Protocols

Three popular video-based (or template-based) datasets are used for evaluation: IJB-B [46], IJB-C [32], and YTF [47].The IJB-B dataset consists of 21,798 images and 7,011 videos from 1,845 subjects captured under unconstrained conditions. Instead of image-to-image or video-to-video recognition and verification, the IJB-B challenge protocol aims to evaluate the FR model on templates. Due to large variations existing in different templates, this protocol is more difficult for the FR model than other benchmarks. We follow the template-based 1:1 verification task to evaluate the proposed method against other methods on the IJB-B dataset. The IJB-C dataset adds an extra 1,686 subjects based on IJB-B and contains 31,334 images and 11,779 videos. To improve the representation of the global population, the IJB-C emphasizes occlusion and diversity of subjects. By increasing the size and variability of the dataset, the IJB-C is more challenging than the IJB-B and other datasets in unconstrained face recognition.

The YTF dataset is a video-based dataset with 3,425 videos belonging to 1,595 different subjects. Each subject has 2.15 videos on average. All videos are collected in unconstrained conditions with large variations in poses, expressions, illuminations, etc. Like the IJB-A verification task, the YTF protocol splits the dataset into ten-fold cross-validation sets. Each set contains 500 randomly selected video pairs, of which 250 pairs are positive and the other 250 pairs are negative.

4.3   Visualization of different FIQA methods

Figure 6 shows the images and corresponding quality scores predicted by different FIQA methods in the IJB-B dataset. We rank the face images by the quality scores from high to low. The first four methods are perceptual-based, and the last three methods are learning-based. The Blur only considers the blurriness of face images, and it chooses those with high resolution while ignoring other factors, such as head pose. The Pose only considers the head pose of a face. Although it can choose images under neural head poses, these images may be too blurry to recognize. The JPEG only considers the jpeg compression rate of face images, and the less compressed images are with higher quality scores. The combination consists of Blur, Pose, and JPEG. The learning-based methods learn quality prediction functions from labeled datasets. It shows the image with a higher quality score contains relatively more information about the identity and is easier to recognize in most situations.

4.4   Memory and Computation Costs

The memory cost and computation cost are the key factors when applying FIQA to a real FR system. Fewer parameters and less computation cost can improve the throughput of a real FR system.

Table 2. The memory and computation costs of different methods. Since several methods are based on aligned images while the others are not, we also present whether it is necessary to align images.

method	need aligning?	parameters (MB)	computation (Gflops)
Blur	no	few	low
Pose	yes	0	low
JPEG	no	0	low
SVM	yes	few	low
RQS	yes	few	low
FaceQNet	yes	23	3.9
tinyFQNet	yes	0.022	0.0024

Table 2 shows the memory and computation costs of different methods. Generally speaking, perceptual-based methods need fewer parameters, while learning-based methods need more memory and computation costs. Although SVM has fewer parameters and low computation cost since it is based on the feature extracted from the FR model, it needs to process all images using the FR model regardless of that the image may be dropped, which would lower throughput. The RQS is based on handle-crafted features, including HoG, Gabor, Gist, LBP, and the feature of a face alignment network. Although less memory and low computation cost, the RQS is not easy to deploy on GPU devices. Compared with FaceQnet which needs nearly 23M parameters and 3.9

Gflops, tinyFQnet only has 21.8k parameters and about 0.0024 Gflops. Due to the efficient design of tinyFQNet, the proposed method is easy to serve as a plug-in with the FR model while the extra computation and memory cost are needed.

4.5   Quantitative evaluation on IJB-B and IJB-C Datasets

Table 3 summarizes the overall results of the 1:1 verification task on the IJB-B dataset. We use true positive rates (TPR) under 10^i (i= -1, -2, -3, -4, -5) and false positive rates (FPR) as the evaluation metric. We choose the random selection as the baseline. We evaluate on two different recognition models, including ResNet-50 [18] and EfficientNet-b0 [43]. We compare the tinyFQnet with three learning-based methods, including SVM [4], RQS [6], FaceQnet [19]. All these methods showed higher performance compared with perceptual-based methods. Since both the [6] and [19] have released their code, we use the released code to compute the quality scores. For SVM [4], we implement their method and replace the training dataset with MS1M used in this paper. We also report the results of perceptual-based methods, including Blur, Pose, JPEG, and their combination.

Table 3. The overall results on the 1:1 verification task of the IJB-B dataset. Two different recognition models ResNet-50 [18] and EfficientNet-b0 [43] are chosen as recognition models. The true accept rates(TAR) vs. false positive rates(FAR) are used as the evaluation metric.

FR model	method	tpr				
		fpr=e-1	fpr=e-2	fpr=e-3	fpr=e-4	fpr=e-5
ResNet-50	random	92.09	87.64	83.08	77.22	64.02
	Blur	93.62	89.96	86.71	82.05	65.25
	Pose	93.45	90.34	86.91	82.39	75.57
	JPEG	95.27	93.14	90.87	88.46	81.93
	Combination	95.55	93.75	91.37	88.56	83.23
	SVM	94.75	91.44	88.51	84.13	75.29
	RQS	96.06	93.78	91.82	89.42	83.13
	FaceQnet	95.15	92.76	90.36	87.29	79.63
	tinyFQNet	**95.99**	**94.32**	**92.73**	**90.78**	**85.21**
EfficientNet-b0	random	92.65	87.81	82.72	75.34	53.74
	Blur	93.98	90.04	86.39	79.71	48.04
	Pose	93.99	90.31	86.64	81.65	70.29
	JPEG	95.7	93.15	90.99	87.85	79.06
	Combination	96.11	93.73	91.55	88.42	80.84
	SVM	95.11	91.56	87.61	82.71	61.54
	RQS	96.14	94.12	91.85	88.89	81.8
	FaceQnet	95.47	92.74	90.22	86.61	76.36
	tinyFQNet	**96.31**	**94.41**	**92.65**	**90.16**	**84.06**

Table 3 shows that the learning-based methods outperform the perceptual-based methods, and both of them are superior to random selection. Among all these methods,

our tinyFQnet achieves 85.21 tpr@fpr=e-5 when using the ResNet-50 network as the FR model (84.06 tpr@fpr=e-5 for EfficientNet-b0), which is the best performance. The SVM performs the worst among learning-based methods. A possible explanation for the results of SVM is that the goal of the FR model is to learn feature representation invariant to perceptual or other potential factors, and the feature of an image extracted from the FR model contains little information on image quality. SVM, which tries to learn a quality prediction function from the feature of a face image, may not learn an effective prediction function about the image quality.

Table 4. The overall results on the 1:1 verification task of the IJB-C dataset. The true accept rates(TAR) vs. false positive rates(FAR) are used as the evaluation metric.

FR model	method	tpr				
		fpr=e-1	fpr=e-2	fpr=e-3	fpr=e-4	fpr=e-5
ResNet-50	random	92.24	87.88	83.11	76.75	68.42
	Blur	94.33	90.69	87.32	83.28	74.65
	Pose	93.8	90.95	87.45	83.06	75.22
	JPEG	96	94.04	91.88	89.27	84.64
	Combination	95.97	93.79	91.41	88.81	83.77
	SVM	94.94	91.98	88.92	84.67	77.21
	RQS	96.17	94.21	92.13	89.72	85.36
	FaceQnet	95.37	93.12	90.8	87.34	82.3
	tinyFQNet	**96.29**	**94.77**	**93.05**	**91.12**	**87.68**
EfficientNet-b0	random	92.31	87.66	82.97	75.51	51.6
	Blur	94.25	90.66	87.41	81.3	57.88
	Pose	94.45	91.09	87.16	82.13	69.76
	JPEG	96.17	93.73	91.84	88.67	81.02
	Combination	96.21	93.61	91.31	88.23	80.23
	SVM	95.29	91.92	88.1	82.97	68.78
	RQS	96.46	94.35	92.21	89.37	84.54
	FaceQnet	95.49	93.13	90.61	86.8	80.12
	tinyFQNet	**96.53**	**94.83**	**93.05**	**90.73**	**86.06**

We also conduct experiments on a more recent and challenging benchmark IJB-C dataset [32]. Table 4 summarizes the overall results of the 1:1 verification task on the IJB-C dataset. Similar to the results on IJB-B, the learning-based methods outperform the perceptual-based methods, and the random selection strategy performs the worst.

4.6 Quantitative evaluation on YTF Dataset

Table 5 shows the results on the YTF dataset. Similar to the results shown in Table 3, our tinyFQnet achieves the best performance among all methods. We find that learning-based methods slightly outperform perceptual-based methods. Unlike the results on the IJB-B dataset, the random selection achieves a comparable performance with the others on the YTF dataset. This is probably due to the fewer variations in a YTF video

Table 5. Performance evaluation for two recognition models with different FIQA methods on the YTF dataset. We adopt 10-folder cross-validation to calculate the verification accuracy.

method		accuracy with ResNet-50	accuracy with Efficientnet-b0
perceptual-based	random	96.34 ±0.11	96.53 ±0.09
	blur	96.39 ±0.07	96.53 ±0.04
	pose	96.58 ±0.05	96.55 ±0.04
	JPEG	96.72 ±0.03	96.72 ±0.02
	combined	96.11 ±0.06	96.54 ±0.05
learning-based	SVM	95.57 ±0.08	96.01 ±0.04
	RQS	96.89 ±0.06	**97.02±0.05**
	FaceQnet	96.35 ±0.03	96.64 ±0.04
	tinyFQNet	**97.02±0.03**	97.00 ±0.04

compared to an IJB-B template. Among all the methods, SVM performs the worst, as explained in Section 4.5. An interesting phenomenon in Table 5 is that the combination of Blur, Pose, and JPEG performs even worse than each separate method.

4.7   Ablation Studies

The impact of FIQA network's size   Since the memory cost and computation cost are the key factors when applying FIQA to a real FR system, we explore the impact of the size of a network on performance in this section. More specifically, we use two different networks: MobileNetV2 and ResNet-50. There are four different architectures for MobileNetV2, including mbn_t4_w0.35_64, mbn_t6_w1_64, mbn_t4_w0.35_112, and mbn_t6_w1_112, where t represents the channels expansion and w represents the width multiplier for basic block in MobileNetV2, the postfix number represents the input size of network.

Table 6. Results of different networks on the IJB-B dataset. Five architectures, including mbn_t4_w0.35_64, mbn_t6_w1_64, mbn_t4_w0.35_112, mbn_t6_w1_112, and ResNet-50 are explored.

network	input size	parameters	(Mflops)	tpr@fpr=e-5
mobilenet_t4_w0.35	64x64	0.013M	1.43	85.21
mobilenet_t6_w1	64x64	0.02M	2.5	85.81
mobilenet_t4_w0.35	112x112	0.5M	44.4	85.19
mobilenet_t6_w1	112x112	2.1M	206.9	86.38
ResNet-50	112x112	22.4M	3907	86.93

Table 6 shows the results of different networks on IJB-B. As illustrated in the table, for a particular FR model, the size of FIQA network makes little difference on FR's performance. On the IJB-B dataset, the largest network ResNet-50 with 23M parameters achieves 86.93 tpr@fpr=e-5, only 1.7 higher than the smallest network

mbn_t4_w0.35_64 with 21.8k parameters. On the YTF dataset, the size of FIQA network almost makes no difference on the FR's performance.

The impact of sampling strategy under different data sizes To evaluate the impact of the data sampling strategy, we evaluate the tinyFQNet on FR performance when using different data sampling strategies to generate a training dataset. Two different sampling strategies are chosen for comparison, including the random strategy and the smooth strategy, as mentioned in 3.4. Figure 7 shows the histogram distribution of quality scores when applying two different strategies, the top row for the random strategy and the bottom row for the smooth strategy.
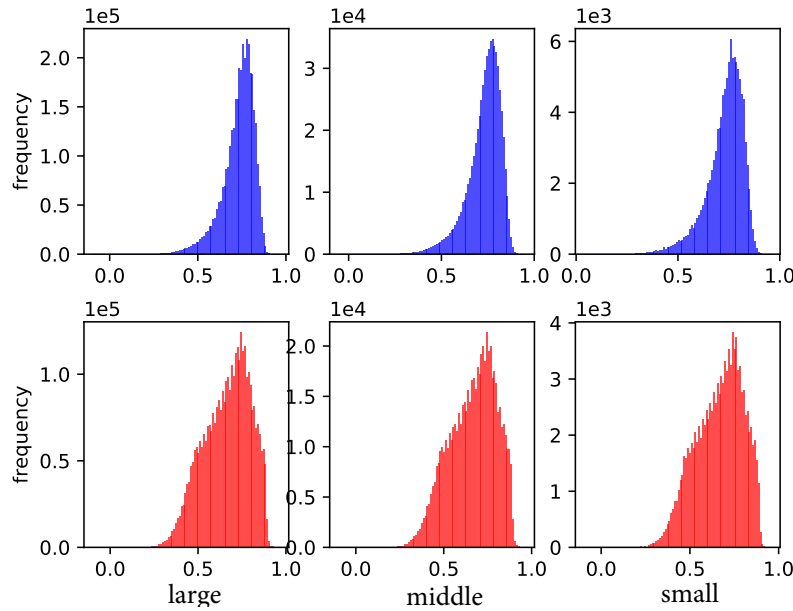
Fig. 7. The distribution of quality scores under different data sizes, including large, middle, and small. The top row shows the results of the random sampling strategy, and the bottom row shows the results of a smooth sampling strategy.

Table 7 shows the results of tinyFQNet on the IJB-B data set when using different sampling strategies to generate a training dataset for training tinyFQNet. For a fair comparison, we keep the same number of training images for both two sampling strategies. It can be observed that the number of identities of the smooth strategy is much lower than the random strategy due to applying identity filtering. In all three different data sizes, the smooth strategy is better than the random strategy.

Table 7. Results on IJB-B dataset when using two different sampling strategies for generating training datasets to train tinyFQNet. To make the results more convincing, we conduct this experiment on three datasets with different sizes.

sampling	data size	number(images, ids)	IJBB tpr@fpr=e-5
random	large	3.5M, 92.9K	83.95
random	middle	0.61M, 87.4K	84.73
ranom	small	0.11M, 57.0K	84.78
smooth	large	3.5M, 34K	85.85
smooth	middle	0.61M, 5.3K	85.64
smooth	small	0.11M, 0.9K	85.45

The influence of the labeling model Since the FR model generates quality scores, a natural thought is how the labeling model (we denote the model used to generate quality scores as the labeling model.) would influence the overall performance of an FR model.

Table 8. Results of different quality labelling models on the IJB-B dataset.

FR model	labeling model	IJBB tpr@fpr=e-5	YTF 10-fold accuracy
ResNet-50	ResNet-50	85.21	97.0
ResNet-50	EfficientNet-b0	85.15	97.0
EfficientNet-b0	EfficientNet-b0	83.71	96.9
EfficientNet-b0	ResNet-50	84.06	97.0

Table 8 shows the results of two different labeling models (used to generate quality labels for face images to train tinyFQnet) on the performance of two different FR models. It shows that both FR models perform better when using ResNet-50 as a labeling model to generate quality scores, but there is no significant difference between them. It can be concluded that with a better labeling model, the FIQA model can improve the performance of an FR system.

5   Conclusion

In this paper, we present a novel deep FIQA method, in which a novel and effective recognition-oriented metric is proposed that directly links face image quality assessment with the performance of the face recognition model. We design a tiny but efficient face quality network (tinyFQnet) and use it to train a FIQA model. We provide analysis by many experiments about the influences of network size on FR's performance and show that even a tiny network with only 21.8K parameters can achieve comparable performance to a much more complicated network with 23M parameters. We show that learning-based methods are superior to perceptual-based methods, and the proposed method outperforms other learning-based methods on two classical benchmarks.

However, limitations exist. The quality metric relies only on cosine similarity comparisons and could explore alternatives and needs to test on more diverse datasets to confirm generalizability. Further efforts on more diverse datasets, alternate metrics, and model optimizations can build on this initial exploration of linking quality directly to recognition accuracy.

References

1. Abaza, A., Harrison, M.A.F., Bourlai, T., Ross, A.: Design and evaluation of photometric image quality measures for effective face recognition. IET Biometrics **3**(4), 314–324 (2014)
2. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., et al.: Partial fc: Training 10 million identities on a single machine. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1445–1449 (2021)
3. Babnik, Ž., Peer, P., Štruc, V.: Faceqan: Face image quality assessment through adversarial noise exploration. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 748–754. IEEE (2022)
4. Best-Rowden, L., Jain, A.K.: Learning face image quality from human assessments. IEEE Transactions on Information Forensics and Security **13**(12), 3064–3077 (2018)
5. Boutros, F., Fang, M., Klemt, M., Fu, B., Damer, N.: Cr-fiqa: face image quality assessment by learning sample relative classifiability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5836–5845 (2023)
6. Chen, J., Deng, Y., Bai, G., Su, G.: Face image quality assessment based on learning to rank. IEEE signal processing letters **22**(1), 90–94 (2015)
7. Chen, K., Yi, T., Lv, Q.: Lightqnet: Lightweight deep face quality assessment for risk-controlled face recognition. IEEE Signal Processing Letters **28**, 1878–1882 (2021)
8. Corner, B.R., Narayanan, R.M., Reichenbach, S.E.: Noise estimation in remote sensing imagery using data masking. International Journal of Remote Sensing **24**(4), 689–702 (2003)
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
10. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
11. Eckert, M.P., Bradley, A.P.: Perceptual quality metrics applied to still image compression. Signal Processing **70**(3), 177–200 (1998)
12. Erez, CohenYitzhak, Yitzhaky: No-reference assessment of blur and noise impacts on image quality. Signal Image & Video Processing (2010)
13. Eskicioglu, A.M., Fisher, P.S.: Image quality measures and their performance. IEEE Transactions on communications **43**(12), 2959–2965 (1995)
14. Ferzli, R., Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb). IEEE Transactions on Image Processing **18**(4), 717–728 (2009)
15. Gao, X., Li, S.Z., Liu, R., Zhang, P.: Standardization of face image sample quality. In: International Conference on Biometrics. pp. 242–251. Springer (2007)
16. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European conference on computer vision. pp. 87–102. Springer (2016)
17. Hassen, R., Wang, Z., Salama, M.: No-reference image sharpness assessment based on local phase coherence measurement. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 2434–2437 (2010)

18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

19. Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., Beslay, L.: Faceqnet: quality assessment for face recognition based on deep learning. In: 2019 International Conference on Biometrics (ICB). pp. 1–8. IEEE (2019)

20. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017)

21. Hsu, R.L.V., Shah, J., Martin, B.: Quality assessment of facial images. In: Biometric Consortium Conference, 2006 Biometrics Symposium: Special Session on Research at the. pp. 1–6 (2006)

22. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)

23. Huber, M., Terhöst, P., Kirchbuchner, F., Damer, N., Kuijper, A.: On evaluating pixel-level face image quality assessment. In: 2022 30th European Signal Processing Conference (EU-SIPCO). pp. 1052–1056. IEEE (2022)

24. Ke, G., Wang, S., Zhai, G., Lin, W., Yang, X., Zhang, W.: Analysis of distortion distribution for pooling in image quality prediction. IEEE Transactions on Broadcasting **62**(2), 446–456 (2016)

25. Kim, H.I., Lee, S.H., Ro, Y.M.: Face image assessment learned with objective and relative face image qualities for improved face recognition. In: Image Processing (ICIP), 2015 IEEE International Conference on. pp. 4027–4031. IEEE (2015)

26. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging **19**(1) (2010)

27. Li, Y., Po, L.M., Feng, L., Fang, Y.: No-reference image quality assessment with deep convolutional neural networks. In: 2016 IEEE International Conference on Digital Signal Processing (DSP) (2016)

28. Liu, J., Zhou, W., Li, X., Xu, J., Chen, Z.: Liqa: Lifelong blind image quality assessment. IEEE Transactions on Multimedia (2022)

29. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6738–6746 (2017)

30. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: Proceedings of The 33rd International Conference on Machine Learning (2016)

31. Liu, Y., Peng, B., Shi, P., Yan, H., Zhou, Y., Han, B., Zheng, Y., Lin, C., Jiang, J., Fan, Y., et al.: iqiyi-vid: A large dataset for multi-modal person identification. arXiv preprint arXiv:1811.07548 (2018)

32. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 International Conference on Biometrics (ICB). pp. 158–165. IEEE (2018)

33. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14225–14234 (June 2021)

34. Moorthy, A.K., Bovik, A.C.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing **20**(12), p.3350–3364 (2011)

35. Nasrollahi, K., Moeslund, T.B.: Face quality assessment system in video sequences. In: European Workshop on Biometrics and Identity Management. pp. 10–18. Springer (2008)

36. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on pattern analysis and machine intelligence **22**(10), 1090–1104 (2000)
37. Sampat, M., Wang, Z., Gupta, S., Bovik, A., Markey, M.: Complex wavelet structural similarity: A new image similarity index. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society **18**(11), p.2385–2401 (2009)
38. Sang, J., Lei, Z., Li, S.Z.: Face image quality evaluation for iso/iec standards 19794-5 and 29794-5. In: International Conference on Biometrics. pp. 229–238. Springer (2009)
39. Sankaran, N., Tulyakov, S., Setlur, S., Govindaraju, V.: Metadata-based feature aggregation network for face recognition. In: ICB. pp. 118–123 (Feb 2018)
40. Shan, S.: No-reference visually significant blocking artifact metric for natural scene images. Signal Processing **89**(8), 1647–1652 (2009)
41. Sheikh, H.R., Member, IEEE, Bovik, A.C., Fellow: An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans Image Process **14**(12), 2117–2128 (2006)
42. Sun, K., Gao, F., Zhu, S., et al.: Face biometric quality assessment via light cnn. Pattern Recognition Letters (2017)
43. Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
44. Vignesh, S., Priya, K.M., Channappayya, S.S.: Face image quality assessment for face selection in surveillance video using convolutional neural networks. In: Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on. pp. 577–581. IEEE (2015)
45. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
46. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset pp. 592–600 (2017)
47. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. In: Computer Vision and Pattern Recognition. pp. 529–534 (2011)
48. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. pp. 74–81. IEEE (2011)
49. Yang, J., Ren, P., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: CVPR. pp. 1–8. IEEE (2017)
50. Zhang, G., Wang, Y.: Asymmetry-based quality assessment of face images. Advances in Visual Computing pp. 499–508 (2009)
51. Zhang, J., Ong, S.H., Le, T.M.: Kurtosis-based no-reference quality assessment of jpeg2000 images. Signal Processing Image Communication **26**(1), 13–23 (2011)
52. Zhang, W., Ma, K., Yan, J., Deng, D., Wang, Z.: Blind image quality assessment using a deep bilinear convolutional neural network. IEEE Transactions on Circuits and Systems for Video Technology **30**(1), 36–47 (2020)
53. Zhu, J., Fang, Y., Ji, P., Abdl, M.E., Wang, D.: Rrar: A novel reduced-reference iqa algorithm for facial images. In: IEEE International Conference on Image Processing (2011)
54. Zhuang, N., Zhang, Q., Pan, C., Ni, B., Xu, Y., Yang, X., Zhang, W.: Recognition oriented facial image quality assessment via deep convolutional neural network. Neurocomputing **358**, 109–118 (2019)