Hindawi
Scientific Programming
Volume 2021, Article ID 4606828, 21 pages
https://doi.org/10.1155/2021/4606828

Research Article

Sharpness and Brightness Quality Assessment of Face Images for Recognition

Ke Li,[1,2] **Hu Chen**,[1] **Faxiu Huang**,[1,2] **Shenggui Ling**,[1] **and Zhisheng You**[1,2]

[1] *National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China*
[2] *Wisesoft Co., Ltd., Chengdu 610045, China*

Correspondence should be addressed to Faxiu Huang; huangfaxiu@stu.scu.edu.cn

Received 19 August 2021; Revised 30 September 2021; Accepted 6 October 2021; Published 25 October 2021

Academic Editor: Jianping Gou

Face image quality has an important effect on recognition performance. Recognition-oriented face image quality assessment is particularly necessary for the screening or application of face images with various qualities. In this work, sharpness and brightness were mainly assessed by a classification model. We selected very high-quality images of each subject and established nine kinds of quality labels that are related to recognition performance by utilizing a combination of face recognition algorithms, the human vision system, and a traditional brightness calculation method. Experiments were conducted on a custom dataset and the CMU multi-PIE face database for training and testing and on Labeled Faces in the Wild for cross-validation. The experimental results show that the proposed method can effectively reduce the false nonmatch rate by removing the low-quality face images identified by the classification model and vice versa. This method is even effective for face recognition algorithms that are not involved in label creation and whose training data are nonhomologous to the training set of our quality assessment model. The results show that the proposed method can distinguish images of different qualities with reasonable accuracy and is consistent with subjective human evaluation. The quality labels established in this paper are closely related to the recognition performance and exhibit good generalization to other recognition algorithms. Our method can be used to reject low-quality images to improve the recognition rate and screen high-quality images for subsequent processing.

1. Introduction

Extensive research on face image quality (FIQ) has shown that samples given as inputs to an automated recognition system influence recognition performance. Face recognition has been increasingly applied in uncontrollable environments (e.g., automated security checkpoints) where the acquired images may include blur, uneven illumination, and nonfrontal poses. Such nonideal factors can significantly decrease the recognition accuracy. The most direct manifestation of this decreased accuracy is that the face recognition performance of the same recognition algorithm on datasets with different qualities has obvious differences. For example, Aghdam et al. [1] used several models to prove that the recognition performance of the same recognition model can differ by 70% or more on data of various qualities captured in the same scene. Some researchers have proposed effective methods to solve the problems caused by nonideal factors in recognition. For example, Cao et al. [2] proposed a posture robustness recognition algorithm, and Fekri-Ershad [3] classified face gender to help improve the recognition rate. These methods have achieved some results. However, filtering low-quality images by face image quality assessment (FIQA) is also an important way to improve the performance of recognition systems.

The quality of face images as biometric samples is closely related to the recognition result. Three characteristics of FIQ have been described in standard ISO/IEC 29794 [4]: (1) the character, which indicates the attributes associated with an inherent characteristic; (2) the fidelity, which reflects the degree of similarity with the source biometric characteristic; and (3) the utility, which indicates the fitness for recognition and is influenced by the character and fidelity. FIQ is defined as a measure of the utility of a face image to face recognition systems [5–7]. This definition is consistent with the utility described above. An FIQ measure can essentially be considered a predictor of face recognition accuracy. In other words, a face image determined to be of high quality should enable recognition systems to succeed (or vice versa). The ultimate goal of FIQA is to exploit the relationship between image quality and the output of recognition algorithms.

The FIQA has great practical value because it can screen face images of various qualities, whether it is applied to real-time recognition systems online or offline face image applications. Restricting face images, which are determined to be of poor quality for recognition, can improve the recognition performance and simultaneously reduce the waste of face recognition system resources. Some adjustment instructions can be provided to persons being identified or staff according to the quality of the image, which has guiding significance for effective dynamic adjustment of the face image acquisition environment. FIQA for images that have failed to be identified can provide feedback to recognition algorithm researchers who are purposefully improving the recognition performance. The development of multi-recognition algorithm systems can be promoted by selecting appropriate recognition algorithm configurations based on the image quality so that the recognition system can utilize images of different qualities. Image enhancement can be promoted by selectively enhancing the image or choosing different enhancement configurations for images of different qualities. In addition, FIQA can be applied to quality-based fusion, database maintenance [7], and dynamic recognition approaches [8, 9].

One of the challenges of face image quality evaluation is that the FIQA output should be closely related to recognition. Recently, some studies have evaluated specific factors, such as clarity, and combined the evaluation results of each factor to obtain the OQ [10, 11]; however, these methods are not closely related to recognition performance. Researchers have proposed deep learning methods to predict quality using the similarity score of two images of a given individual as labels [5]. Although these methods have been used to achieve some breakthroughs, there is still a lack of identity-oriented methods that do not rely heavily on recognition algorithms.

In this work, experiments are conducted on a database (denoted the SC database) of images that were collected in identification channels. Because the people are ready for recognition in this scene, the captured pictures are mostly frontal portraits without occlusions but include light distortion by uneven light and blur due to the transitions between the identified persons. We mainly conduct a composite assessment of face image brightness and sharpness by supervised deep learning methods on these images. We also use the same method to perform experiments on the CMU multi-PIE [12] face database (M-PIE) data and cross-validation on Labeled Faces in the Wild (LFW) [13]. The main contribution of this article is as follows: (i) the effects of brightness and sharpness on recognition are simply verified on the M-PIE, and suitable very high-quality images (VHQI) per subject for identification are selected with International Standard [4, 14] and human consensus [7] before image labeling. (ii) We establish brightness and sharpness labels associated with identification. As a result, the images are divided into nine categories that represent varying degrees of brightness and sharpness. (iii) A classification model is trained to predict quality based on the self-built SC database and established quality labels, and the quality of the classified data is verified. In particular, the network structure is derived from the literature [15] and improved. The method for establishing the labels is different from the method that uses only the similarity score, which depends seriously on a recognition algorithm and uses only subjective assessment, deviating from the recognition in this paper. The trained model can predict which class the image belongs to, where the classes represent different levels of brightness and sharpness.

This paper is organized as follows: Section 2 surveys the quality assessment methods for face images. Section 3 describes the materials and methods, including face databases and preprocessing, the method of selecting the VHQIs and establishing quality labels, and the network structure. Experimental settings and results are provided in Section 4. Section 5 presents a concluding summary of this work and directions for future work.

2. Related Work

FIQA is a branch of image quality assessment (IQA) but is also an extension of image quality. IQA can be subdivided into (i) full-reference (FR) [16, 17], (ii) reduced-reference (RR) [18, 19], and (iii) no-reference (NR) [20–23] categories according to the amount of information provided by the reference image. FIQAs also include FR-based approaches; for example, there is relevant literature [24–26] that reports the use of computing luminance distortion, structural similarity (SSIM), and probabilistic similarity to reference face images. However, FR and RR methods are not easy to apply because of the difficulty in obtaining undistorted reference images. Studies of NR-IQA are necessary. The FIQAs described below are all based on NR. FIQAs can be categorized into non-deep learning (non-DL FIQA) and deep learning (DL FIQA).

Non-DL FIQAs mostly assess specific factors, such as sharpness, occlusion, pose, symmetry, expression, illumination, and resolution by defined methods. One of the early studies proposed by Gao et al. [27] demonstrated the assessment of symmetry for light and pose, eye distance, illumination, contrast, and blur. Another method for evaluating symmetry proposed by Zhang and Wang [10] is based on local scale invariant feature transform (SIFT) features. Sang et al. [28] also evaluated symmetry through illumination and pose based on a Gabor filter and measured blur by a discrete cosine transform (DCT) and inverse DCT. In the literature [29], researchers have employed DCT to evaluate sharpness. Nasrollahi et al. [30] utilized the least out-of-plane rotated (LOPR) faces method to evaluate posture. Furthermore, overall quality (OQ) is always obtained by combining the evaluation results of each factor. Nasrollahi and Moeslund [30] also measured the illumination, blur, and resolution and performed weight fusion to obtain the OQ. A similar method exists in the literature [31]. Chen et al. [32] divided images into three categories: nonface images, unconstrained face images, and identification (ID) card face images, and assumed that these ranks gradually increase. Rank-based OQ normalized to [0, 100] is acquired by five feature fusions and is applied to learn rank weights.

DL FIQAs have emerged in recent years and are almost supervised. Zhang et al. [33] created the Face Image Illumination Quality Database based on human assessments and trained a model based on ResNet-50 [34]. The experimental results show that the predicted illumination quality is closely related to the labels defined by humans but lacks a relationship between the predicted quality and the recognition performance. Rowden and Jain [5] established quality labels for the LFW training database through the two methods of human assessments and matcher dependence. Given established target face quality values, a support vector model was trained on face features extracted by a convolutional neural network (CNN) to predict the quality of the face images. Yu et al. [11] synthesized 5 degradations (nearest-neighbor downscaling, Gaussian blur, additive white Gaussian noise, salt-and-pepper noise, and Poisson noise) with 3 configurations on the CASIA WebFace [11] and trained a classification model on 16 classes of images, including the original unmodified image and 15 synthetic degradation images. OQ scores were obtained by pooling 16 products of the image degradation classification confidence and the face image recognition accuracy under the corresponding degradation. In the literature [35], a two-stream CNN named "deep face quality assessment (DFQA)" was proposed. Yang et al. [35] divided the quality scores into 5 segments, which were categorized by angle, clarity, illumination, visibility, expression, etc., and established manual labels for 3000 images from ImageNet to train the pretrained SqueezeNet model. The DFQA was trained to predict OQ scores on the MS-Celeb-1M [36] dataset with quality labels produced by the pretrained model. Hernandez-Ortega et al. [37] proposed FaceQnet based on ResNet-50 for quality learning on a 300-subject subset of VGGFace2 [38]. The quality labels in this experiment are comparison scores derived from multiple feature extractors between the probe images and high-quality images selected by the BioLab-International Civil Aviation Organization (ICAO) framework. Zhang et al. [39] and Zhuang et al. [40] utilized a multitask structure with several factors and OQ labels that were established by humans and a related algorithm for 3000 images from the Intelligence Advanced Research Projects Activity (IARPA) Janus Benchmark-A [41] (IJB-A) dataset. The features extracted from the front shared layers are set as the subtask layer inputs for predicting various quality factors, such as pose. The outputs of the subtask are fused to produce an OQ score via fully connected layers. Unsupervised methods, including SER-FIQ [42] and MagFace [43], have emerged in the last two years. SER-FIQ uses stochastic embedding robustness to estimate face image quality. MagFace obtains the quality scores by learning a universal representation of face recognition and quality assessment.

In this work, we combine the similarity score of face recognition algorithms, the definition grade classification method of the human visual system, and the traditional brightness classification method to establish brightness and sharpness labels associated with identification. Furthermore, given that we have established FIQ labels for a self-built database, we train a classification model based on MobileNetV3 [15], which can predict a face OQ that simultaneously represents the brightness and sharpness rank. To our knowledge, this is the first attempt to combine recognition performance with human assessments for FIQ labels.

3. Materials and Methods

3.1. Face Databases and Preprocessing. This work utilized three face databases: the M-PIE, a self-built SC database, and the LFW. The M-PIE was collected under an environment with strict lighting, posture, and expression control in four sessions over a five-month period; these data consist of 337 subjects and more than 750,000 high-resolution face images. The SC database consists of approximately 5000 face images of 945 subjects selected from identification channels of Wisesoft Co., Ltd. The subjects were employees of the company and agreed to the use of their images in the study. The specific screening methods will be described later. The images in LFW were derived from natural scenes in life, and a total of 13,233 images of 5,749 subjects were included, of which more than 70% of subjects had only one image. The M-PIE contains face images that were acquired under the condition that only one factor changes, while the other factors remain optimal. For example, when capturing images under different lighting conditions, the face remained in a frontal posture with a neutral expression. We extended the M-PIE to 9 classes of data similar to the SC database. The experiments were trained using the SC database and M-PIE and then evaluated on the LFW and subsets of the SC database and M-PIE other than the training set. The prediction results on the LFW dataset were used to see how the evaluation results correlate with the human visual system and recognition performance.

In this work, all images were detected, and five key points (pupils of two eyes, nasal tip, and two corners of the mouth) were marked by a model based on a multitask convolutional neural network (MTCNN) [44] that included only the convolution layer in the first stage; thus, the input of the model was not limited to a defined size. MTCNN mainly adopted three cascade networks: the proposal network (P-NET) for rapidly generating candidate windows, the refine network (R-NET) for high-precision candidate window filtering and selection, and the output network (O-NET) for generating final boundary boxes and face key points. O-NET was a regression task that minimized the Euclidean loss of the facial landmark coordinates ($\hat{y}_i^{\text{landmark}}$) obtained from the network and the ground-truth coordinate (y_i^{landmark}) for the $i-$th sample. Euclidean loss is as follows:

$$L_i^{\text{landmark}} = \left\| \hat{y}_i^{\text{landmark}} - y_i^{\text{landmark}} \right\|_2^2. \tag{1}$$

In the literature [45], Best-Rowden divided FIQ into three scenarios: (i) whether an image contains a face, (ii) evaluation of the accuracy of face alignment, and (iii) the quality of an aligned face image. We will discuss the third scenario. Face images need to be preprocessed to align the faces as much as possible in the process of face recognition. The input image evaluation of face recognition systems has more practical significance; therefore, it is necessary to carry out the same image pretreatment as that used for recognition in FIQA. Based on the key points detected, the process of pretreatment was as follows: first, the midpoint denoted P1 between the two pupils and the midpoint called P2 between the two corners of the mouth were found. Then, we connected P1 and P2 to obtain line segment L, calculated the angle between L and the vertical line as the rotation angle denoted by θ, and rotated the face clockwise or counterclockwise by θ so that all faces had the same posture on the plane. Finally, the image of the face was magnified or reduced to the specified size. Specifically, we scaled each image to 150×150 pixels.

3.2. Brightness and Sharpness Factor Verification. This paper focuses on the brightness and sharpness of the image, and we use specific data to illustrate the degree of influence for the recognition of these two factors before introducing the method for establishing quality labels. The M-PIE contains images taken under 19 light conditions, where other quality factors are optimal (see Figure 1). This database is suitable for verifying the influence of individual factors on identification and is therefore chosen to verify the effect of sharpness and brightness on recognition. We tested the recognition of images in different lighting environments with a classical face recognition algorithm (FRA-A) based on a Light CNN-9 [46] with max-feature-map (MFM) units. We know that the human visual system is very accurate at recognizing people, as it is even better than current state-of-the-art recognition systems [47, 48]. Similarly, some studies [5, 31, 49] have verified the usability of the human visual system in FIQA. In the following work, we used the human recognition system to assist in the selection of images and the establishment of labels. The M-PIE does not contain off-light images. Therefore, we brightened some of the images by a power exponential operation via image transformation to verify the quality of the off-light images. Gamma (G) parameters of 0.14 and 0.28 were selected to augment images called Bri0* and Bri1*, respectively (see Figure 2). Usually, to minimize the error caused by labeling single images based on the similarity scores (SS) determined for a pair of images, it is necessary to select suitable VHQI per subject for identification or verification. It is also necessary to choose the images with the most appropriate brightnesses as the VHQI and then test the verification accuracy (VA) of images with different brightnesses.

According to the given brightness indicator (last two digits of the image file name) of M-PIE and human visual perception, brightness images (filename with "06~08," named Bri2) with high VAs were selected. Samples of darker images were gradually added to the previously selected samples. After the addition of some dark images, the VA changed very little, so we added more dark samples to simultaneously carry out the FRA-A test and obtained the results in Table 1. The identification of each type of test image is listed as follows: Bri2 (06~08), Bri3 (05~09), Bri4 (05~09 and 15~17), Bri5 (04~11 and 14~18), Bri6 (02~18), and Bri7 (0~19). Table 1 shows the VAs for these types of images, where Bri1 represents the two types of images Bri2 and Bri1* and Bri0 represents the three types of images Bri2, Bri1*, and Bri0*. The VA of Bri2 was the peak and was clearly higher than the VA of Bri1. The VA of Bri3 was very close to the result of Bri2 and exhibited a significant decrease compared to the VA of Bri4. Thus, Bri3 was chosen as the VHQI. Images of similar brightnesses, which are marked as Bri4* (15~17), Bri5* (04, 10~11, 14, and 18), Bri6* (02 and 03), and Bri7* (00, 01, and 19), were paired with the same person in the VHQI before testing. When comparing model performance, the larger the area under the receiver operating characteristic (ROC) curve (AUC) is, the better the model effect will be. Retaining the recognition algorithm unchanged, the AUC was proportional to the quality of the image. Figure 3 shows the ROC curves for different brightnesses under FRA-A. Because the luminance was the only distortion, there were several types of data with good and very similar recognition rates. To show the classification effect of each type of data more clearly, the vertical and horizontal coordinates were adjusted. Figure 3 shows that the recognition performance of the Bri3* images is the highest, and the recognition rate decreases gradually with brightening and dimming.

To study the influence of sharpness on the recognition rate, we synthesized four degrees of blurred images (Blu1~Blu4) with motion blur and tested the images. Blur was added by convolving an image with a kernel, which was obtained by an affine transformation of the rotation matrix generated by the size of the kernel (K) and the rotation angle (45°). The original image and the composite image are shown in Figure 4. Table 2 and Figure 5 show the test results. With the reduction in clarity, the recognition rate of each type of data decreased significantly, and Blu4 was completely unsuitable for recognition.

3.3. VHQI of per Subject. In this work, we selected the VHQI of each subject in the SC database with high definitions, suitable brightnesses, no occlusions, frontal poses and neutral expressions using face recognition algorithms, human vision systems, and traditional brightness calculation methods. The specific processes are as follows:

Low-quality face images with interference factors (a nonfrontal pose, an occlusion, and a nonneutral expression) from the original image set Q0 were excluded as much as possible through the human visual system to obtain image set Q1.

High-definition images denoted by Q2 were manually screened from image set Q1 by two persons. The specific screening principle was based on the absolute scale of the subjective evaluation method [50], as shown in Table 3.

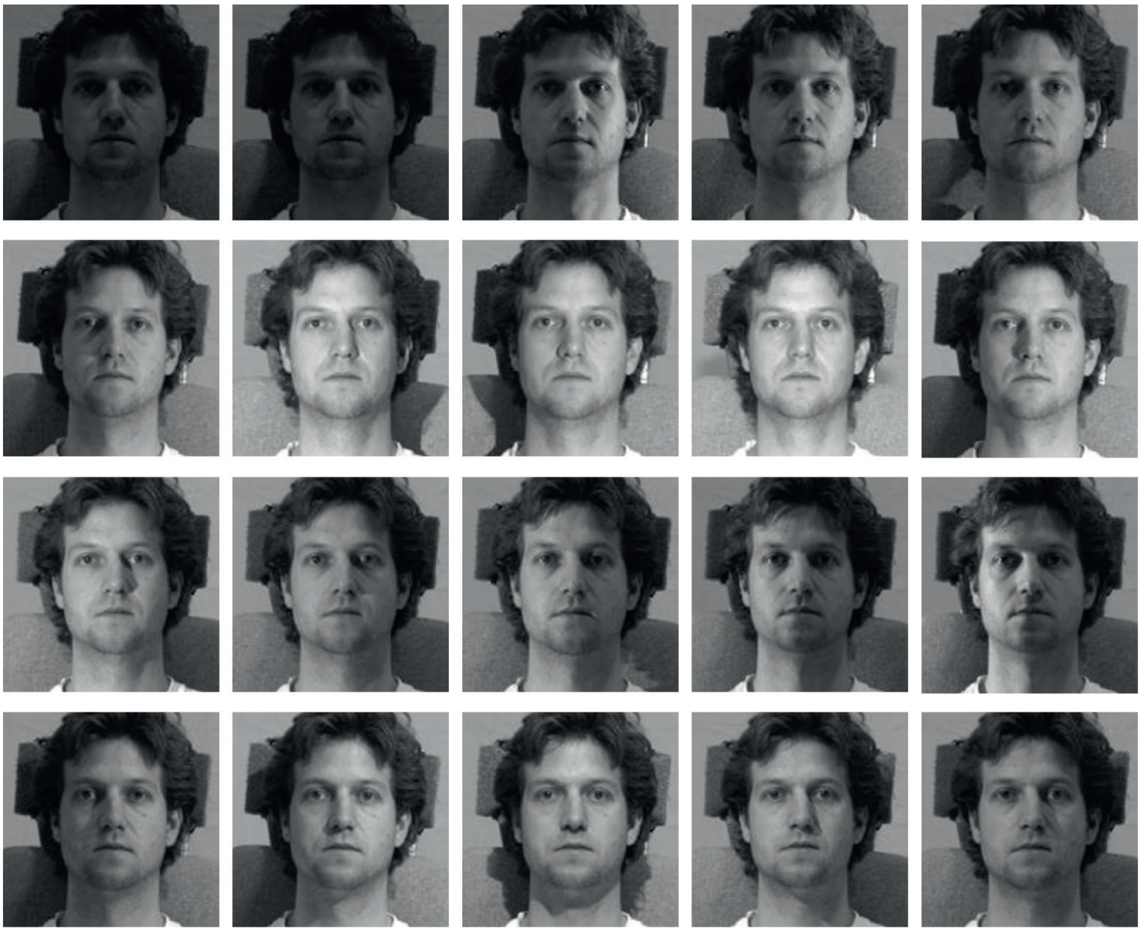


Figure 1: Images from the M-PIE captured under different light conditions.

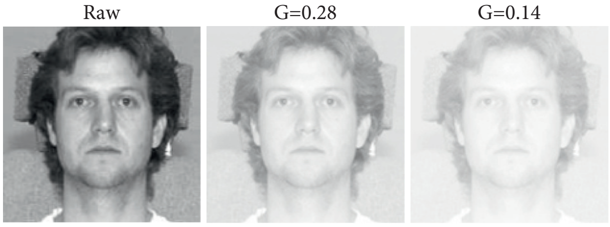


Figure 2: Examples of raw images from the M-PIE and image brightness tuning. The extended (E) images are made by gamma correction of the raw (R) image. The specific step is to scale each pixel of the original image to [0, 1] and then transform it by $E = R^G$. The specific gamma parameters 0.28 and 0.14 were selected to produce images with brightness intervals that can be discerned by the human visual system.

Table 1: Recognition rate of face images with different brightnesses.

Data	Bri0	Bri1	Bri2	Bri3	Bri4	Bri5	Bri6	Bri7
TAR (%) @ 1% FAR	96.82	99.21	100.00	99.96	99.85	99.73	99.28	98.89
TAR (%) @ 0.1% FAR	88.16	95.58	99.58	99.40	98.64	97.77	95.56	93.83
TAR (%) @ 0.01% FAR	75.80	89.31	98.67	97.80	95.71	93.30	88.52	84.89

TAR: true acceptance rate; FAR: false acceptance rate.

When images are rated as 5 points, they are classified into Q2.

Images of different brightnesses in Q2 were selected for testing to determine the appropriate brightness. We cropped the 96×96 face area from the center of the face image to reduce the background influence. The brightness of the face area was determined by the distribution of gray values. Assume that the number of pixels whose gray value

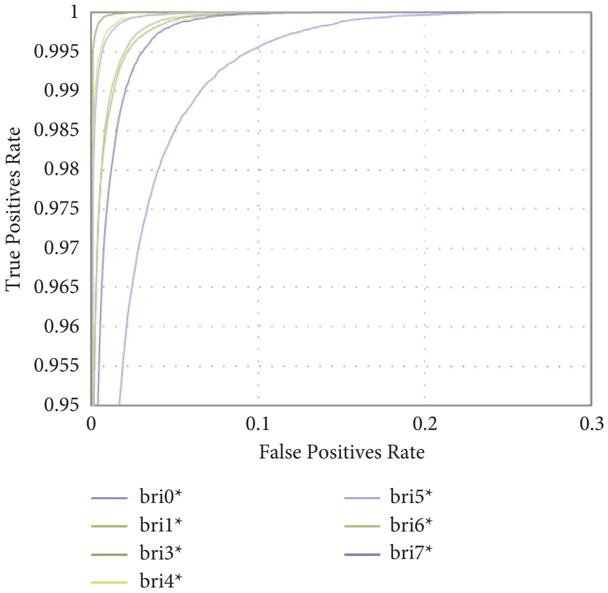


Figure 3: ROC curves for images with different brightnesses. It illustrates that under the same FRA, different recognition performances can reflect different image qualities. The figure shows that the AUC gradually decreases from Bri3∗ to Bri7∗ or Bri0∗, indicating that the image quality continues to decline with the change in brightness. To show the classification effect of each type of data more clearly, the vertical coordinates were adjusted.

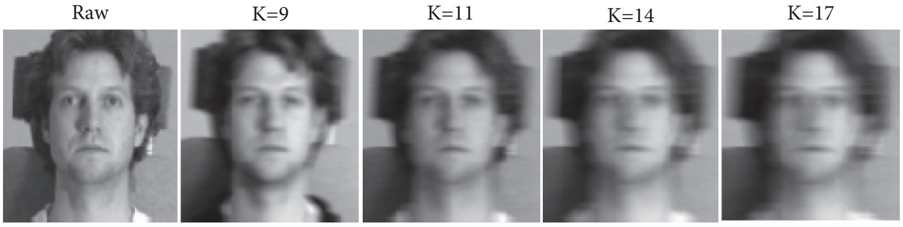


Figure 4: Examples of raw image and image sharpness tuning. The numbers at the top of the images are parameters.

Table 2: Recognition rate of face images with different sharpness degrees.

Data	Blu0	Blu1	Blu2	Blu3	Blu4
TAR (%) @ 1% FAR	99.97	97.20	91.02	47.47	18.87
TAR (%) @ 0.1% FAR	99.56	88.48	73.44	15.13	2.32
TAR (%) @ 0.01% FAR	98.79	76.67	54.58	3.66	0.14

was between v and v_1 (where v was less than v_1) was m, and the total number of pixels was n. P_i (m/n) was the proportion of pixels in the defined brightness interval. The values of v, v_1, and P_i were the brightness parameters to be determined. An initial brightness range $[k_0, k_1]$ and P_i were chosen manually. Then, we continuously adjusted the brightnesses of the images, assuming that the brightness adjustment range was $[k_{0-p}, k_{1+q}]$, where p and q were positive parameters. After each adjustment, a face recognition test was carried out on multiple images of the same person under a certain brightness, where n is the number of adjustments, and the VAs $\{\mathrm{va}_1, \mathrm{va}_2, \mathrm{va}_3, \ldots, \mathrm{va}_n\}$ were obtained when the FAR was equal to 0.01% with different brightnesses. Appropriate brightnesses appeared when the VA began to evidently change; that is, the brightness corresponding to the difference between the VA and all previous VAs was less than or equal to α, and the difference between the VA and all subsequent VAs was greater than or equal to β $(\alpha < \beta)$. The visual explanation is shown in Figure 6, in which VA and the changing trend of VA are hypotheses for interpretation. We obtained high-resolution images with good brightness (v is 90, v_1 is 200, and P_i is 0.65) and called them Q3.

The VHQI of each object was obtained by testing multiple images of the same object in Q3 and selecting the top 80% of image pairs that were ranked when the similarity score was higher than the corresponding threshold value at an FAR of 0.01%. As we expected, the SS of image pairs were basically higher than this threshold value. The flowchart for determining VHQI(s) is shown in Figure 7.

3.4. Establishment of Quality Labels. On the basis of establishing VHQI, we employed FRAs trained on a self-built database, including four million images and manual evaluation to establish quality labels for the SC database. The above experiments demonstrate that sharpness is more sensitive to recognition, so we used the recognition rate to assist in the classification of sharpness. Then, we classified the brightness of the data with different sharpness.

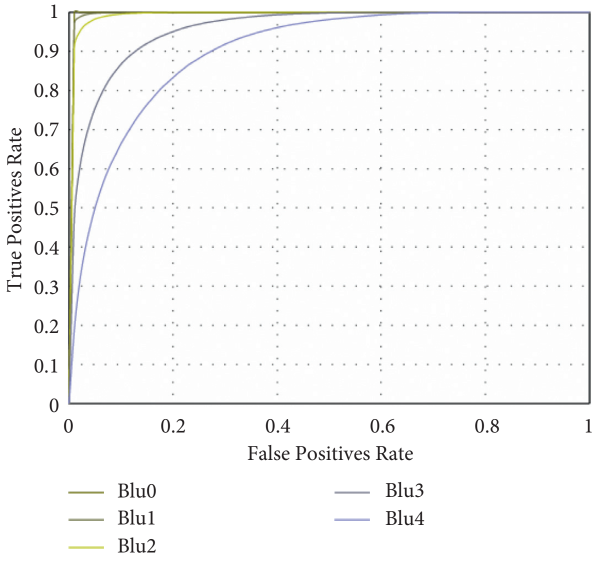


Figure 5: ROC curves for face images with different sharpness.

Table 3: Absolute evaluation scale for images.

Score/level	Quality scale
5 points/excellent	There is no sign that the quality of the image has deteriorated.
4 points/good	The image quality has deteriorated, but it does not interfere with viewing.
3 points/fair	It is clear that the image quality has deteriorated and is slightly obstructed for viewing.
2 points/poor	There is a hindrance to viewing.
1 point/bad	Images include a very serious hindrance to viewing.

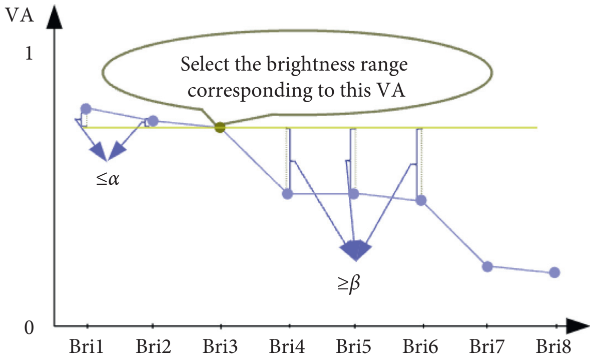


Figure 6: A visual explanation of the method for selecting the brightness range. The VA and the changing trend of VA are hypotheses for interpretation.

The dark (v was 0, v_1 was 80, and P_i was 0.75) and bright ranges (v was 150, v_1 was 255, and P_i was 0.75) were determined in a similar way. The differences among the above methods were that the selected images of different brightnesses were formed into positive samples with the standard images, and p and q were determined according to the recognition results.

Face image classification should have a corresponding significance in the category of biometric sample quality. The quality of biological samples can be divided into three categories. (i) low-quality samples (LQS) that cannot be used for identification or may produce poor identification results. If possible, these samples should be replaced with high-quality samples. (ii) Medium-quality samples (MQS) that may yield good certification results in most environments, but in requirements-based applications, it is necessary to include high-quality samples. (iii) High-quality samples (HQS) that can produce good certification results under any circumstances.

The face images without VHQIs were divided into three categories according to the SS, where each category represents the corresponding significance of the quality category for biometric samples. If a subject has multiple VHQI, the similarity score of an image built with a label is the average similarity value of all corresponding VHQIs. Images with SS below threshold 1 (T1) are defined as LQS. Images with SS above threshold 2 (T2) and below threshold 3 (T3) represent MQS. Images with SS above T3 are HQS.

The previous three thresholds were obtained by FRA-A and FRA-B. FRA-B is a commercial face matcher for self-identification channels. We assume that the terms A, B, and C are used to represent thresholds for FRA-A at 1%, 0.1%, and 0.01% FAR, respectively, and that L, M, and N are the corresponding FRA-B counterparts. To ensure that the established labels do not rely heavily on a single recognition algorithm, we used the threshold value combined with two algorithms to classify the image. Different SS may be obtained for the same pair of images with features extracted by different types of recognition. Therefore, the similarity score

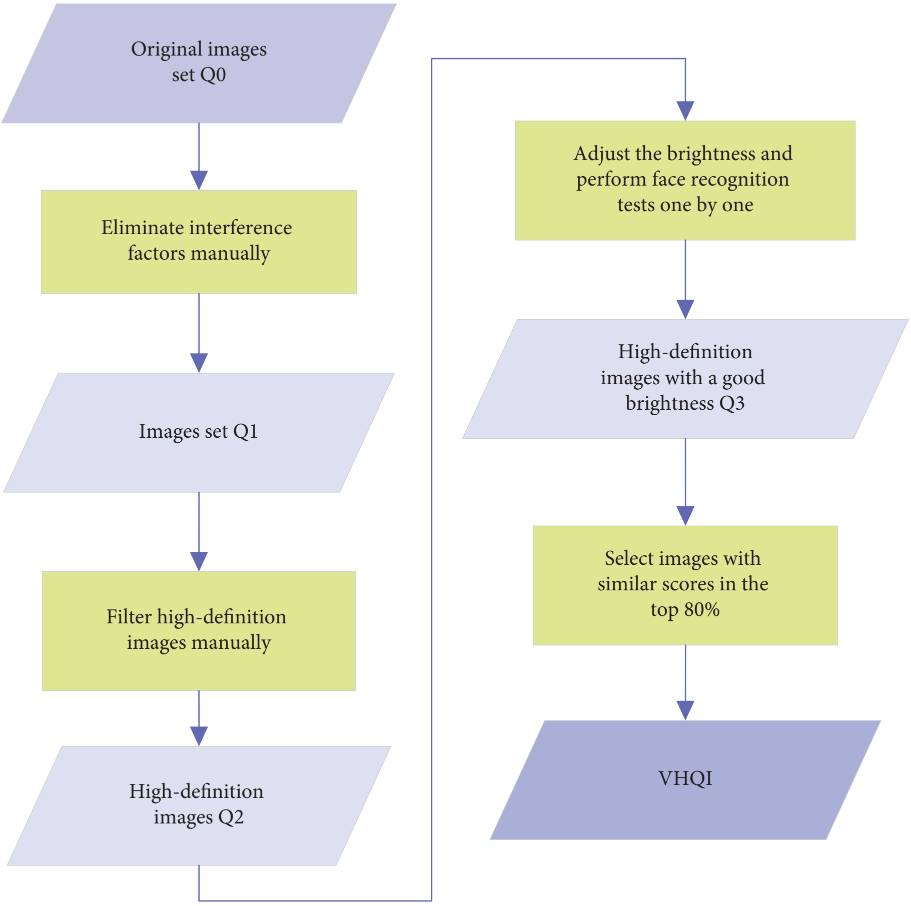

Original images set Q0
Eliminate interference factors manually
Images set Q1
Filter high-definition images manually
High-definition images Q2
Adjust the brightness and perform face recognition tests one by one
High-definition images with a good brightness Q3
Select images with similar scores in the top 80%
VHQI

Figure 7: The flowchart for determining VHQI.

calculated by FRA was transformed on the basis of the threshold (at 0.01% FAR) transformation of two recognition algorithms for the same value. If C is greater than N, $T1$, $T2$, and $T3$ can be represented by the following formulas:

$$\begin{aligned} T1 &= \min\left(A \cdot \frac{N}{C}, L\right), \\ T2 &= \max\left(B \cdot \frac{N}{C}, N\right), \\ T3 &= N. \end{aligned} \tag{2}$$

In this work, $T1$, $T2$, and $T3$ are 0.4, 0.54, and 0.65, respectively, after transformation. The similarity scores of image pairs obtained from the FRA-A were multiplied by N/C. The boundary between MQS and HQS was set at a certain interval to make the two types of samples more distinguishable. Each class image was screened by the human visual system with definition refinement criteria, as shown in Table 4. The L_1blur (very fuzzy), M_2blur (clear), and H_3blur (high clarity) images were selected from LQS, MQS, and HQS, respectively.

Finally, each of the above three categories was divided into three categories based on the brightness ranges defined above. Of the remaining images selected from the two categories bright and dark brightness, we selected the image with a certain brightness difference between bright and dark brightness as the appropriate brightness, which is consistent with the brightness level of the previous selection criteria. On the basis of establishing the VHQI, the flowchart for establishing these labels is shown in Figure 8. The face images were divided into nine categories. In the next section, M-PIE data were synthesized to simulate these nine types of data.

3.5. Network Structure. Based on the quality labels established for the dataset above, we attempted to train a classification model to predict image quality. Given that deep learning has been used to make great achievements in the field of computer vision, we also adopted this method to achieve the goal of this study. An important application of FIQA involves embedding it into a real-time face recognition system to improve the recognition or verification performance. FIQA has to be very efficient; otherwise, it would not make sense to use FIQA in real-time face recognition systems. This efficiency includes the model storage and prediction speed. The problem with model storage is that a large number of weight parameters induce high requirements on the device memory. The speed problem is mainly due to poor processor performance or high computational requirements. Lightweight classification networks became our primary choices for efficiency improvement. We

Table 4: Sharpness refinement criteria.

Image damage scale	Sign of sharpness
No damage observed	H_3blur (high clarity)
The image is damaged but is still pleasant/slightly unpleasant	M_2blur (clear)
The image is unpleasant/very unpleasant	L_1blur (very fuzzy)

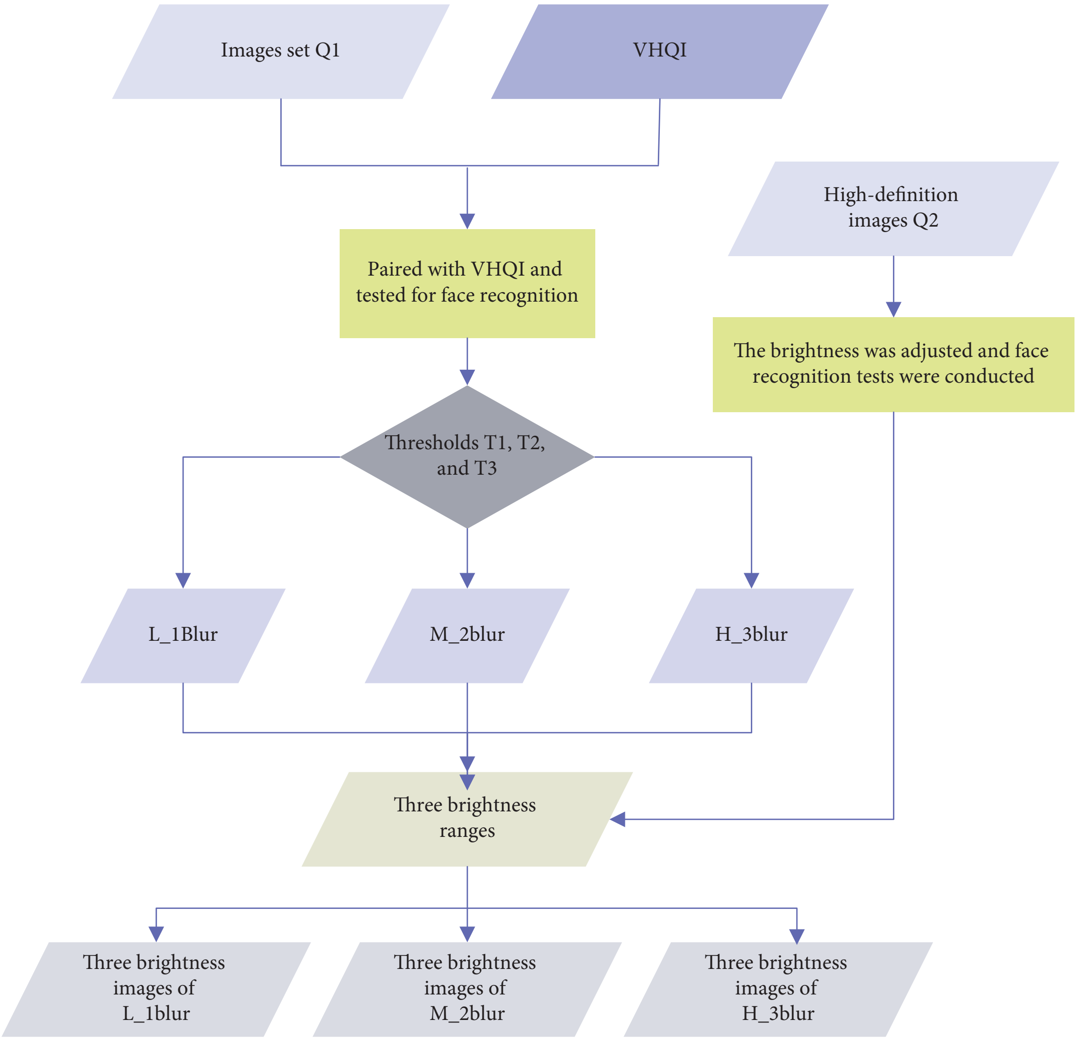

Images set Q1
VHQI
High-definition images Q2
Paired with VHQI and tested for face recognition
The brightness was adjusted and face recognition tests were conducted
Thresholds T1, T2, and T3
L_1Blur
M_2blur
H_3blur
Three brightness ranges
Three brightness images of L_1blur
Three brightness images of M_2blur
Three brightness images of H_3blur

Figure 8: The flowchart for establishing these labels.

adopted lightweight MobileNetV3 to predict FIQ. MobileNetV3 is created through a combination of network design and automated search algorithms, including network architecture search (NAS) and the NetAdapt algorithm. MobileNetV3 can achieve higher accuracy while reducing latency for classification.

MobileNetV3 is ameliorated from MobileNetV2 [51] and includes a resource-efficient block with inverted residuals and linear bottlenecks. These improvements were realized by redesigning expensive layers, introducing a new nonlinearity and adding a squeeze-and-excite (SE) submodule [52] . The initial set of filters decreased from 32 to 16, the last few layers of the network were removed, and the position was changed to maintain accuracy and reduce latency. The hard version of swish (h-swish) was proposed and used in the second half of the model to reduce the number of memories. Swish and h-swish are defined by the following formulas. The SE fixed at 1/4 of the number of channels was added after depthwise (DW) convolution:

$$\begin{aligned} \text{swish} \cdot x &= x \cdot \sigma(x), \\ h - \text{swish}\,[x] &= x \cdot \frac{\text{ReLU}(x+3)}{6}. \end{aligned} \tag{3}$$

Two MobileNetV3 models named MobileNetV3-Large and MobileNetV3-Small were created for high and low resource use cases, respectively. FIQA preferably has a faster response time, so MobileNetV3-Small was chosen for this work. In the literature [53], inspired by network pruning, Xu et al. proposed that IdleBlock targeting creates a larger receptive field and introduced a hybrid composition of IdleBlock with normal blocks that have constrained input and output dimensions. It was shown that hybrid composition networks with IdleBlocks are more efficient and able to both reduce computation and achieve real-world speed increases. The IdleBlock is implemented by a simple pruning method that involves concatenating a subspace ($C \cdot \alpha$ channels, α is between 0 and 1) of inputs including C channels and the rest subspace ($C \cdot (1 - \alpha)$) with transformations. An illustration featuring an inverted residual block (MBBlock) is shown in Figure 9. In this work, we replaced two MBBlocks by IdleBlock with half-pruned channels. The last two layers were replaced by fully connected layers. The architecture of the MobileNetV3-Small with IdleBlock is shown in Figure 10.

4. Results and Discussion

In this section, a series of experiments were conducted to verify the effectiveness of the proposed method. We demonstrated the performance of the classification model using various classification evaluation indicators. We report the rationality of established labels and the robustness of the proposed FIQA method for different FRAs. Finally, we conducted a cross-validation experiment on the LFW.

4.1. Synthesizing Data on the M-PIE. To more transparently explain the feasibility of the method for establishing the labels and the performance of the assessment method, we synthesized images with varying degrees of brightness and sharpness by adjusting brightness and implementing the blur methods mentioned above. Similar to the SC database, the selected data from the M-PIE contain frontal poses and neutral expressions. Three types of blurred images (L_1blur, M_2blur, and H_3blur) are obtained by setting the K parameter to 12 and 20. Appropriate, bright, and dark brightness ranges correspond to the brightnesses in Bri3, Bri1∗, and Bri7∗, respectively. Specifically, the Bri7∗ data are less than the data of the other two brightness images, so the Bri6∗ and number "18" data are dimmed to Bri7∗ brightness. We applied the method described in the previous section to establish labels and screened a total of approximately 15,000 images, including approximately 3,000 "3nor" images and 1,500 images from the other 8 categories. The 9 types of synthesized and labeled M-PIE images are shown in Figure 11.

4.2. Training Setup. In our implementation, hardware with 4 GeForce GTX 2080Ti GPUs was used for accelerated training, and the PyTorch deep learning framework was adopted under the Ubuntu 16.04 operating environment. A stochastic gradient descent with 0.9 momentum was chosen. The learning rate was initialized to 0.01, with a batch size of 256, and attenuated to $1e-5$ according to the adjustment strategy. The input image size was fixed to 96×96, and the input image was preprocessed as the input of the chosen FRAs. Eighty percent of the labeled SC and CMU datasets are used as training sets and the rest are used for testing. All models were trained with 100 epochs.

4.3. Classification Model Results

4.3.1. Classification Model Results on the SC Database. The easiest way to evaluate a classification model is to calculate the accuracy. The accuracy is the percentage of the number of correct predictions in the total samples. The accuracy rates of the trained model called FBSA_M (face brightness and sharpness model) and FBSA_M1 using MobileNetV3-Small on the SC database without and with IdleBlock were 89.83% and 90.87%, respectively. Since accuracy is not a comprehensive evaluation index, we also calculated the precision and recall. The precision is the probability that samples will actually be positive among all the samples predicted to be positive. The recall is defined as the proportion of the number of samples predicted to be positive to the true number of positive samples. The precision and recall of each class are shown in Table 5.

Table 5 shows that the three classes of L_1blur images have the best classification effects. The results from the three types of M_2blur images are relatively poor. The reason for these poor results may be that the similarity fraction interval for the three types of H_3blur images is small, making the model classification difficult. Furthermore, we illustrate the confusion matrix of the classification results in Figure 12 to show the class in which each sample was assigned. Overall, most images were correctly classified. Basically, the misclassified samples were grouped into adjacent categories that had the same degree of either blur or brightness. The number of samples that were incorrectly predicted to be L_1blur was extremely small, which indicates that the model is still effective when limiting the recognition of low-quality images (L_1blur).

4.3.2. Classification Model Results on the M-PIE. On the M-PIE, the classification accuracy of the model was 99.00% and 99.51% without and with IdleBlock (FBSA_M2), respectively. The model classification effect is very good, so we do not show the corresponding precision, recall, and confusion matrix of the model. The effect on the M-PIE is better than that on the SC database, probably because the M-PIE data were collected in a controlled environment and each kind of synthesized data had excellent consistency, resulting in easier classification.

4.4. FIQA Performance. The error-versus-reject curve [54] (ERC) proposed by Grother and Tabassi is often used to evaluate FIQA performance. In this method, FRAs are used to determine whether the image pairs match, and FIQA is used to predict the quality of the image for later filtering. First, an error rate is selected based on a fixed threshold of similarity scores. The error rate is then recalculated by removing images whose quality scores predicted by an FIQA model are below the ever-increasing quality threshold. Finally, the quality threshold is taken as the abscissa, and the

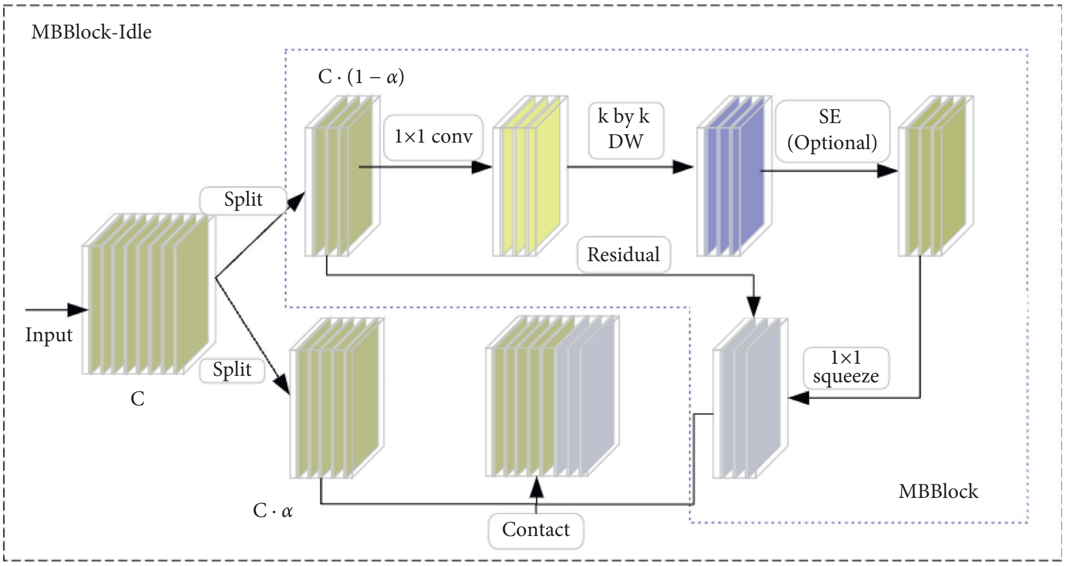


Figure 9: IdleBlock with MBBlock.

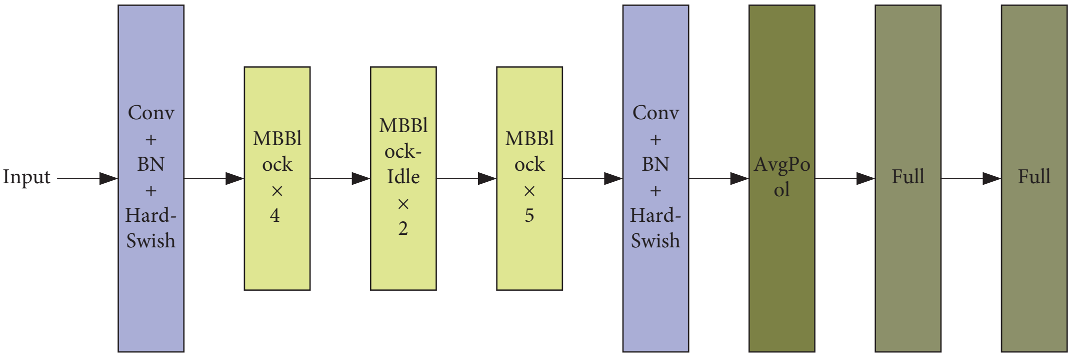


Figure 10: The MobileNetV3-Small with IdleBlock.

recalculated error rate is taken as the ordinate to draw a curve and obtain the ERC.

The quality labels established for images in this paper were category labels, and each category represented a different quality. Therefore, when drawing the ERC, we chose to remove a certain type of image rather than selecting quality thresholds and removing images below these thresholds. From this curve, the reasonability of our labels and the performance of model classification can be revealed. To verify that the quality prediction model we trained was still effective for other recognition algorithms, we chose four algorithms to verify the quality assessment effect. The four algorithms included a Light CNN-9 plus residual layer network (FRA-C), LightCNN-29 (FRA-D), IR-50 [55] (FRA-E), and IR-152 [55] (FRA-F). FRA-C was trained on the same training set as FRA-A. FRA-D, FRA-E, and FRA-F were trained on MS-Celeb-1M. We chose the false nonmatch rate (FNMR) as the error rate, with initial values of 0.20 and 0.35, to show the relationship between the prediction quality and recognition performance for all the FRAs mentioned in this paper.

4.4.1. Performance of FBSA_M1 on the SC Database. The resulting ERCs are shown in Figure 13 upon removing each type of image according to the predicted labels on the SC database. On the whole, the three categories of images (L_1blur, M_2blur, and H_3blur) had certain degrees of discrimination. When the threshold was equal to 0.2 FNMR, the L_1blur images exhibited a good distinction from the other 6 categories of images for all FRAs. These 3 categories of M_2blur are similar to the results of “3dark” and “3bri.” This similarity may be partly due to the mutual classification error between M_2blur, “3nor” and “3bri,” resulting in some of the predicted M_2blur categories being greater than the required threshold; likewise, the opposite is the case for “3bri” and “3nor.” The reason for this situation may also be partly because the numbers of these classes below the threshold are similar. However, the average score of M_2blur was lower than that of “3dark” or “3bri”, so when the threshold was 0.35, M_2blur and H_3blur were clearly distinguished. In the case of two thresholds, the FNMRs of 6 FRAs all decreased significantly after the removal of the three types of L_1blur images. This result indicates that these

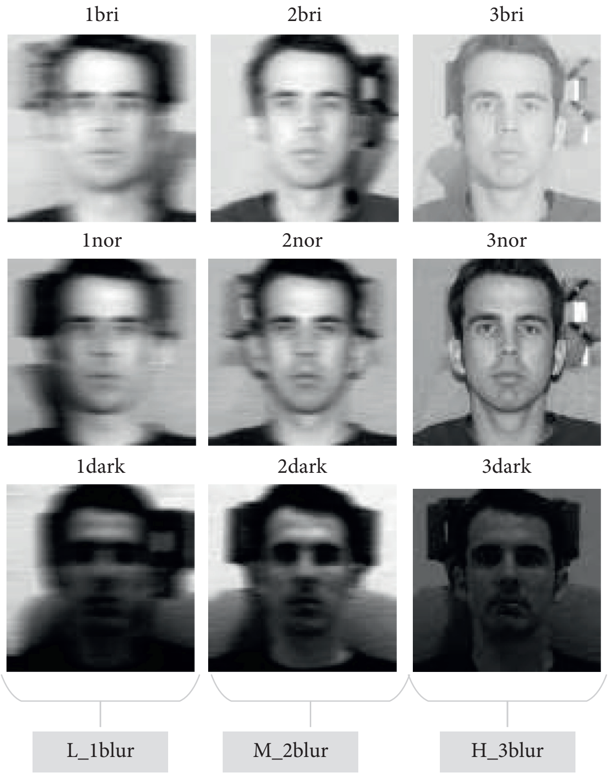


Figure 11: Data in the M-PIE and synthesized data. Nine categories of face images. From left to right, the sharpness increases, and from top to bottom, the brightness decreases.

Table 5: Precision and recall of the classification model on the SC database.

Class	Precision (%)	Recall (%)
1bri	96.43	91.01
2bri	80.00	82.19
3bri	87.30	91.67
1nor	86.25	90.79
2nor	85.59	82.79
3nor	87.07	84.17
1dark	100.00	97.50
2dark	92.19	98.30
3dark	99.15	96.67

three types of images hinder recognition, even for FRA-C, FRA-D, FRA-E, and FRA-F, which did not participate in the establishment of the quality labels. After the three types of H_3blur images were removed, the FNMR values increased to different degrees, meaning that these images can promote recognition. For all FRAs, the FNMRs reached their peaks after "3nor" was removed; this phenomenon is consistent with the definition of "3nor" images as VHQIs in this paper. The mentioned experiments show that the proposed method has a certain compatibility for generalization to other FRAs.

Table 6 reports the results for FRAs with and without a quality assessment module. Experiments without this module are called the baselines. Comparison methods include a general-purpose IQA method deep bilinear CNN (DBCNN) [23] and two FIQAs, i.e., FaceQnet [37] and MagFace [43]. We pretrained the synthetic distortion CNN (SCNN) in DBCNN on LFW and extended data (the cross-validation set in 4.5) and fine-tuned DBCNN. Meanwhile, to verify the effect of IdleBlock we added, we also conducted ablation experiments with FBSA_M. The three comparison methods input an image and predict the corresponding QS. Our dataset can be roughly divided into three types of quality images. An FIQA that predicts a result for a QS requires the application of a threshold to classify an image; however, this threshold is not easy to determine. Therefore, we divided the quality scores predicted by each comparison method into three categories, with the lowest quality representing the images filtered by FRAs.

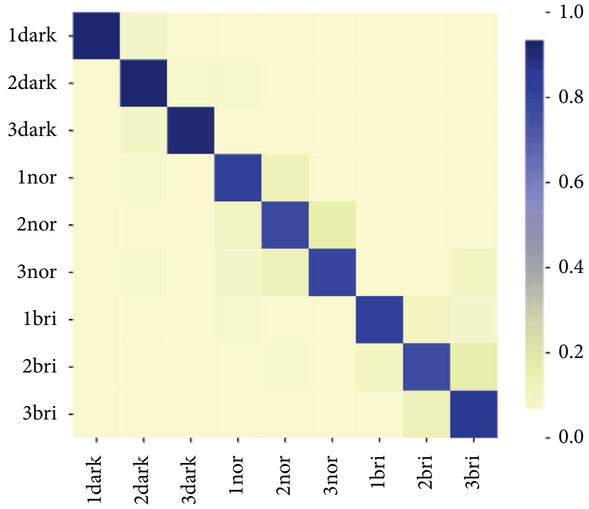
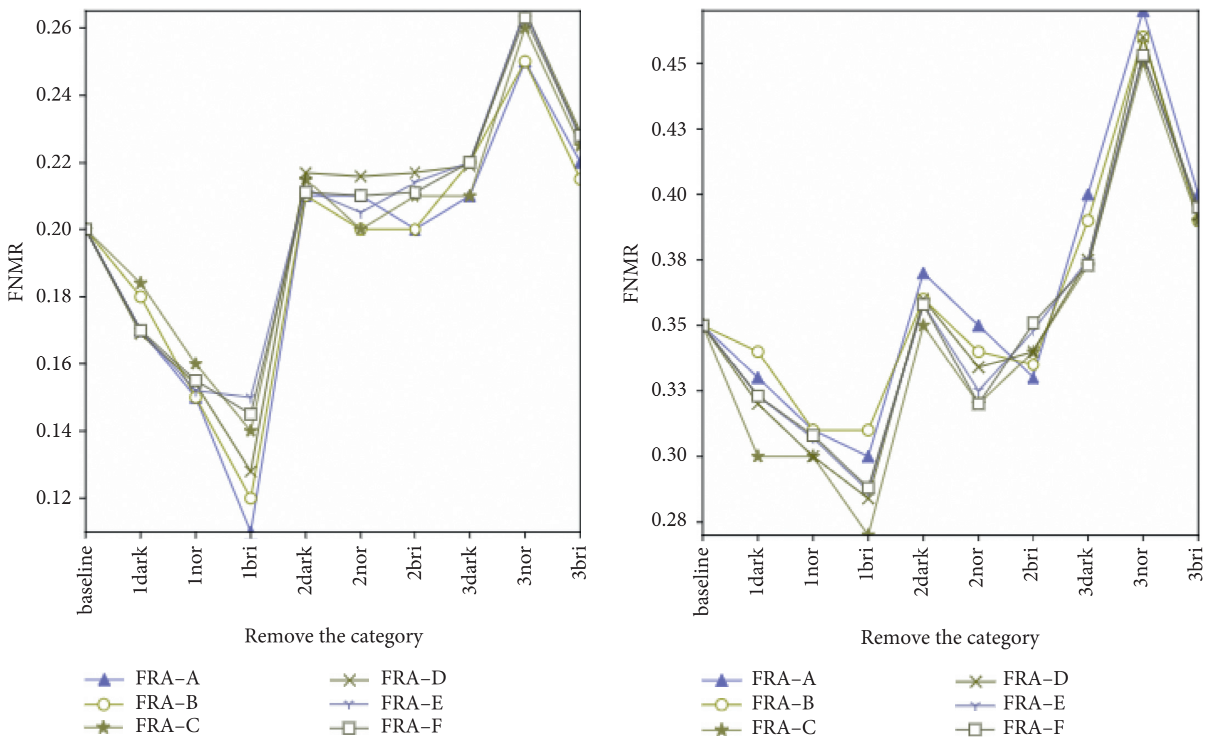


Figure 12: Confusion matrix of the classification results.

At fixed FARs, all FRA performances with FaceQnet, FBSA_M, and FBSA_M1 regarding poor-quality image rejections are improved. With FBSA_M1, the TAR maximally increases by 17.11%, 16.91%, and 11.37% for FRA-A at 1% FAR, 0.1% FAR, and 0.01% FAR, respectively. For most FR algorithms, the results of FBSA_M1 is better than those of FBSA_M. For the other FRAs that were not involved in the creation of labels, the recognition rates were improved by at least 7.61%, 15.45%, and 7.95% after the FBSA_M1 module was used to filter the low-quality images. Our FBSA_M1 is at least 2.8% better than FaceQnet at 1% FAR and is slightly worse at 0.1%. DBCNN is an excellent general-purpose IQA, but DBCNN may not learn the quality characteristics related to recognition in the SC database due to the difference between the pretrained data and the real data. MagFace has little effect, probably due to the small training set relative to the 5.8 M images in the original paper. These experiments show that the proposed FBSA_M1 can reject low-quality images to improve the recognition performance.

4.4.2. Performance of FBSA_M2 on the M-PIE. The resulting ERCs are shown in Figure 14 upon removing each type of image according to the predicted labels on the M-PIE. The ERCs are similar to the results for the SC database. Under the two thresholds, the FNMRs of all FRAs decreased significantly after L_1blur was removed; as expected, the results



(a)

(b)

Figure 13: ERCs of the FRAs for the SC database. ERC curves with different initial thresholds. (a) 0.20 FNMR and (b) 0.35 FNMR. The curves show the efficiency of removing face images in reducing FNMR. The ERC curve drawn in Figure 13 is slightly different from that expressed by the proposer. We recalculate FNMR after removing a class of images but not those below a certain quality threshold. Our ERCs can still show the effect of each type of image removed on recognition performance.

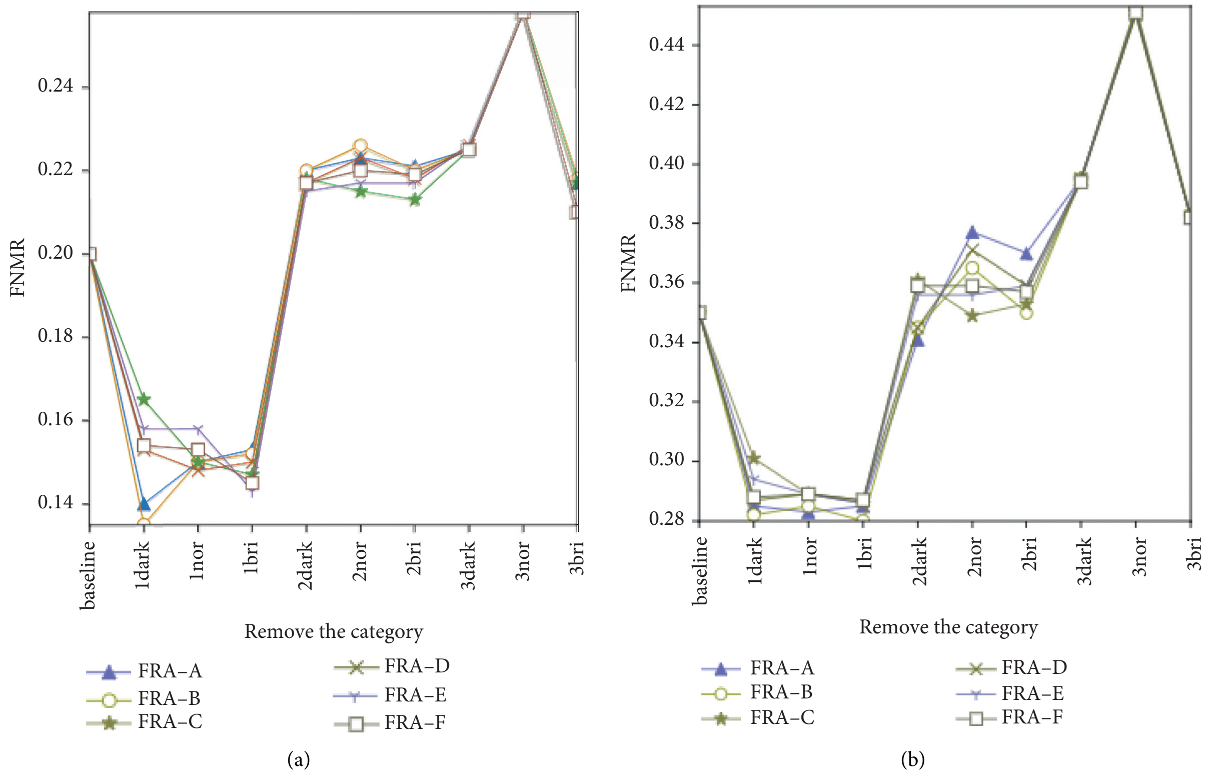
Table 6: Verification performance with and without DBCNN, FaceQnet, MagFace, FBSA_M, and FBSA_M1 quality assessment modules on the SC database.

FRA	(F)IQA	TAR (%) @ 1% FAR	TAR (%) @ 0.1% FAR	TAR (%) @ 0.01% FAR
FRA-A	Baseline	81.39	64.24	42.63
	DBCNN	82.14	61.98	36.79
	FaceQnet	91.61	78.19	**54.28**
	MagFace	80.04	62.46	42.54
	FBSA_M	97.31	80.97	54.10
	FBSA_M1	**98.50**	**81.16**	54.00
FRA-B	Baseline	83.26	64.99	37.57
	DBCNN	80.88	64.55	45.35
	FaceQnet	93.80	77.95	**48.41**
	MagFace	81.03	61.05	35.71
	FBSA_M	96.59	81.56	47.62
	FBSA_M1	**97.52**	**81.82**	47.40
FRA-C	Baseline	91.55	66.79	37.95
	DBCNN	82.14	56.83	34.04
	FaceQnet	96.36	78.61	**48.74**
	MagFace	86.19	65.23	38.94
	FBSA_M	98.65	82.15	47.13
	FBSA_M1	**99.16**	**82.41**	47.96

Table 6: Continued.

FRA	(F)IQA	TAR (%) @ 1% FAR	TAR (%) @ 0.1% FAR	TAR (%) @ 0.01% FAR
FRA-D	Baseline	81.37	65.82	40.31
	DBCNN	80.23	62.68	40.33
	FaceQnet	92.35	78.52	**51.72**
	MagFace	80.39	64.35	41.05
	FBSA_M	97.06	82.31	51.11
	FBSA_M1	**98.16**	**82.41**	50.88
FRA-E	Baseline	85.26	63.98	30.32
	DBCNN	84.46	61.45	30.85
	FaceQnet	91.47	75.72	**39.84**
	MagFace	85.23	62.77	31.78
	FBSA_M	97.31	80.05	38.38
	FBSA_M1	**97.86**	**79.43**	38.27
FRA-F	Baseline	86.15	69.75	35.20
	DBCNN	85.46	67.83	35.27
	FaceQnet	93.61	80.38	**45.29**
	MagFace	85.23	68.74	35.29
	FBSA_M	97.14	85.30	44.51
	FBSA_M1	**97.86**	**85.67**	44.51

Note. Having a quality prediction model means that the TARs are recalculated after the predicted low-quality images are discarded. L_1blur images (1bri, 1nor, and 1dark) predicted by our models and the lowest third of images predicted by three comparison method were discarded. The bold values mean that under different FARs, FIQA models are best for improving validation rates for different FRAs and are to give readers a faster understanding of the performance of the FIQA algorithms.



(a)

(b)

Figure 14: ERCs of multiple FRAs for the M-PIE and its extension images. ERC curves with different initial thresholds. (a) 0.20 FNMR. (b) 0.35 FNMR.

were the opposite after H_3blur was removed. In addition, without “3nor,” the FNMRs reached their maxima. M_2blur, “3dark,” and “3bri” had a general effect at the 0.2 FNMR threshold, but the effect was improved at the 0.35 FNMR threshold. Similarly, “3nor” was predicted to have the best quality. L_1blur, M_2blur, and H_3blur had obvious differences in recognition performance, indicating that the evaluation results of the quality evaluation model in this paper were strongly correlated with the recognition performance.

Table 7 summarizes the verification performance with and without different FIQAs on the M-PIE. With FBSA_M2,

Table 7: Verification performance with and without DBCNN, FaceQnet, MagFace, and FBSA_M2 on the M-PIE dataset.

FRA	(F)IQA	TAR (%) @ 1% FAR	TAR (%) @ 0.1% FAR	TAR (%) @ 0.01% FAR
FRA-A	Baseline	82.42	68.90	43.89
	DBCNN	87.69	73.64	59.07
	FaceQnet	92.41	81.46	**72.03**
	MagFace	87.49	74.35	65.25
	FBSA_M2	**100**	**94.42**	60.97
FRA-B	Baseline	80.46	63.36	46.15
	DBCNN	90.66	76.73	**64.54**
	FaceQnet	93.31	81.40	63.89
	MagFace	87.63	74.14	59.26
	FBSA_M2	**99.77**	**84.60**	64.01
FRA-C	Baseline	55.35	44.67	40.38
	DBCNN	**79.97**	58.72	47.07
	FaceQnet	71.86	**63.19**	**59.45**
	MagFace	65.92	57.44	55.03
	FBSA_M2	76.22	62.07	56.11
FRA-D	Baseline	44.73	40.17	34.36
	DBCNN	54.47	49.40	43.42
	FaceQnet	**62.29**	**57.37**	**50.55**
	MagFace	57.55	53.81	48.59
	FBSA_M2	62.14	55.81	47.74
FRA-E	Baseline	61.10	48.17	42.41
	DBCNN	73.28	57.40	51.49
	FaceQnet	76.53	64.82	**59.29**
	MagFace	69.89	59.75	55.77
	FBSA_M2	**83.45**	**66.90**	58.92
FRA-F	Baseline	62.06	47.62	42.73
	DBCNN	73.56	56.32	56.27
	FaceQnet	77.69	64.21	**60.34**
	MagFace	70.16	58.70	55.95
	FBSA_M2	**85.24**	**66.14**	59.36

The bold values mean that under different FARs, FIQA models are best for improving validation rates for different FRAs and are to give readers a faster understanding of the performance of the FIQA algorithms.

TAR increased by 23.18% for FRA-F, 25.52% for FRA-A, and 17.08% for FRA-A at 1% FAR, 0.1% FAR, and 0.01% FAR, respectively. For all FRAs, the TAR improved by at least 10%. Similar to the results on the SC database, our FBSA_M2 was significantly higher at 0.1% FAR, except for FRA-D, and was slightly lower than that of FaceQnet, except for FRA-A. With the increase in the training set data, DBCNN and MagFace had a significant effect, but the effect was not as good as that of our method for most FRAs. FaceQnet is slightly higher than our method for FRA-C and FRA-D, but overall, our quality prediction model was effective for filtering out low-quality images to improve the recognition rate.

4.5. Cross-Database Performance. We have verified that our method exhibits good generalizations for different FRAs. Experiments to test the generalization capability for another LFW dataset were also conducted. We used the human visual system to assist in the establishment of labels, so we also verified whether the evaluation results were consistent with the human visual system. We evaluate different degrees of luminance and sharpness factors of face images, and the degradation span of these two factors in LFW is small. Therefore, we adjusted the brightness and sharpness of the images by a method that was similar to extending the M-PIE. Most of the subjects in this dataset had only one image, and most of the images included nonpositive postures and nonneutral expressions, so there is a lack of required VHQI for image matching. This experiment was just a test to determine whether our model worked when images included multiple distortion factors. We selected data with more than two images of one subject for testing. Because the sharpness and brightness distortions were extended in the same way as the M-PIE, we used the quality model trained on M-PIE for prediction. The ERCs of all FRAs are plotted in Figure 15.

Figures 15(a) and 15(b) show that under the two initial thresholds, the FNMR can be greatly reduced after the removal of images such as L_1blur; likewise, the FNMR can be significantly increased after the removal of images such as H_3blur. These two initial thresholds could not be employed to separate “3dark” from M_2blur, so we set the initial threshold as 0.6 FNMR and drew ERCs (see Figure 15(c)) to verify the effectiveness of our model. In this way, the quality differences between the three types of images can be visualized more clearly. In summary, our model can distinguish three types of data (L_1blur, M_2blur, and H_3blur), and these three types of data are strongly correlated with the recognition performance.

The results for FRAs with and without the state-of-the-art method and the proposed method are shown in Table 8.

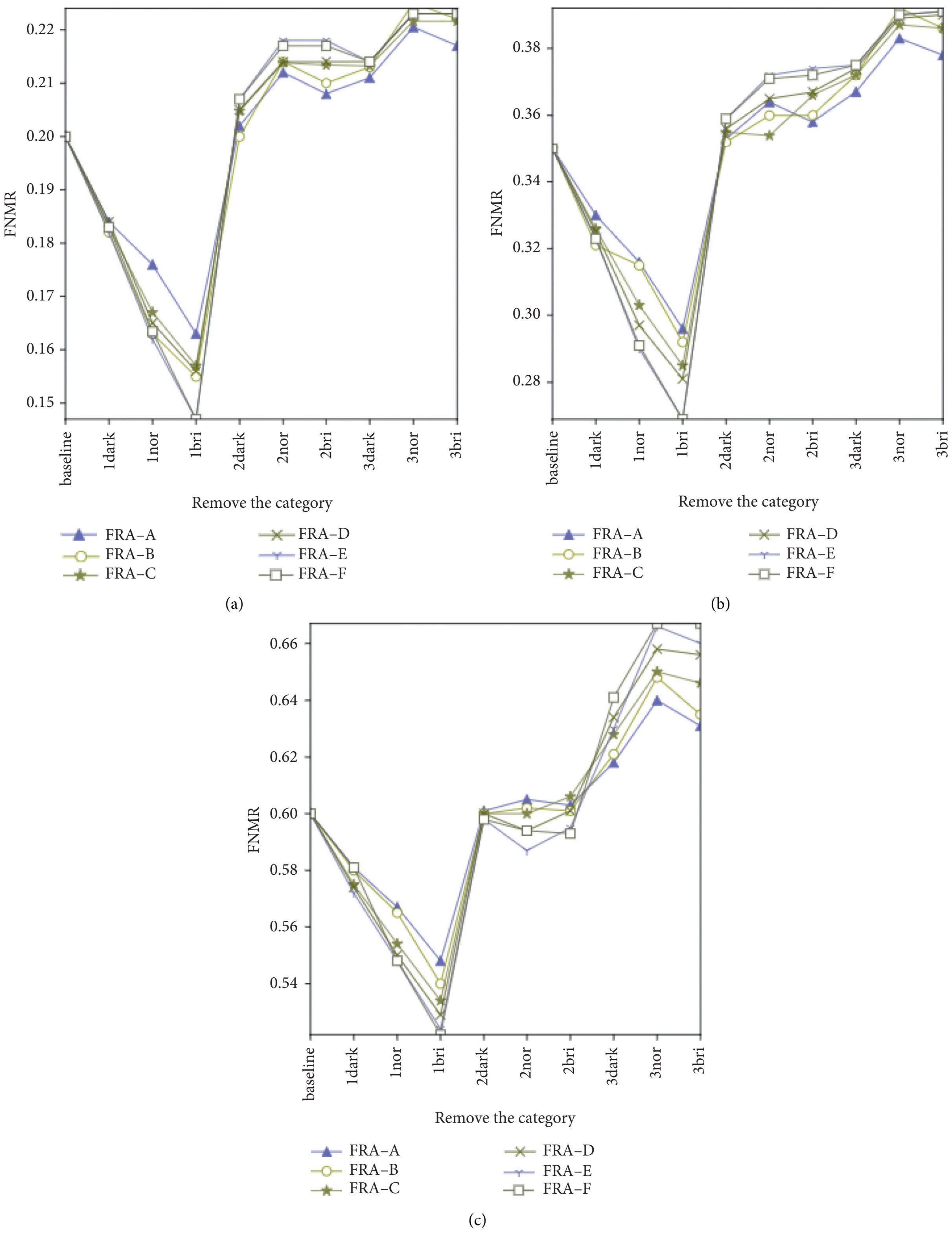


Figure 15: ERCs of multiple FRAs for the LFW and its extension images. ERC curves with different initial thresholds. (a) 0.20 FNMR. (b) 0.35 FNMR. (c) 0.60 FNMR.

Table 8: Verification performance with and without (F)IQA on the LFW.

FRA	(F)IQA	TAR (%) @ 1% FAR	TAR (%) @ 0.1% FAR	TAR (%) @ 0.01% FAR
FRA-A	Baseline	55.20	24.67	10.87
	DBCNN	50.41	25.15	13.15
	FaceQnet	60.92	28.33	12.25
	MagFace	53.14	23.93	10.79
	FBSA_M1	67.50	33.73	14.78
	FBSA_M2	**71.32**	**36.01**	**16.40**
FRA-B	Baseline	47.47	24.50	13.59
	DBCNN	52.44	25.53	13.43
	FaceQnet	50.28	25.57	13.43
	MagFace	45.90	22.62	11.51
	FBSA_M1	54.70	28.83	15.68
	FBSA_M2	**57.52**	**30.87**	**17.83**
FRA-C	Baseline	45.19	17.14	7.03
	DBCNN	50.67	18.95	7.25
	FaceQnet	51.44	19.84	7.80
	MagFace	42.60	16.69	6.96
	FBSA_M1	61.32	25.77	11.15
	FBSA_M2	**65.60**	**26.81**	**11.50**
FRA-D	Baseline	43.08	16.75	7.17
	DBCNN	48.51	18.64	7.44
	FaceQnet	50.04	19.74	8.12
	MagFace	42.09	15.73	6.14
	FBSA_M1	61.43	26.81	11.85
	FBSA_M2	**64.96**	**26.90**	**11.88**
FRA-E	Baseline	57.63	19.31	7.28
	DBCNN	64.52	21.67	7.62
	FaceQnet	65.49	23.09	8.23
	MagFace	56.33	17.73	6.28
	FBSA_M1	76.33	**32.75**	12.15
	FBSA_M2	**83.32**	31.98	**12.17**
FRA-F	Baseline	60.15	20.44	6.84
	DBCNN	66.98	22.83	7.02
	FaceQnet	67.93	24.42	7.68
	MagFace	58.97	18.81	5.82
	FBSA_M1	77.87	**34.68**	11.41
	FBSA_M2	**85.18**	33.78	**11.42**

The bold values mean that under different FARs, FIQA models are best for improving validation rates for different FRAs and are to give readers a faster understanding of the performance of the FIQA algorithms.

After the FIQAs were used to filter out low-quality images, the TARs improved for most FRAs. Our method exhibited better performance, with the highest improvements of 25.69% at 1% FAR, 13.34% at 0.1% FAR, and 5.53% at 0.01 FAR for FBSA_M2. FBSA_M2 exhibited a better performance relative to FBSA_M1, which may be attributed to the notion that the training data of FBSA_M2 were synthesized in the same way as this validation dataset. FBSA_M1 trained with data from practical application scenarios was still better than FaceQnet. Our method also produced accurate predictions of sharpness and brightness. DBCNN has some effect on all FRAs except FRA-A, and MagFace had little effect on cross-validation sets. Because IQA partially differs from FIQA, a good general-purpose IQA may not be suitable for FIQA tasks. At 0.01% FAR, the TARs had a low recognition rate. This result is because the images in the LFW contained various factors of distortion and were of worse quality after brightness and sharpness degradation, resulting in very few data that exceeded the 0.01% FAR threshold. Through our quality evaluation model, recognition performance can be effectively improved by rejecting low-quality images for different FRAs and datasets. The proposed method has better robustness.

Based on the assessment results of the model, we randomly selected images from each prediction class and displayed them in Figure 16. Based on this visualization, it is clear that the model exhibited an accurate judgment of extreme brightness and sharpness. We arranged the images based on the brightness and sharpness of adjacent classes in the training data, such as “1bri” and “2bri” and “1bri” and “1nor.” Therefore, some of the data are completely outside the training data for our model, which may lead to ambiguity in classification. Rowden and Jain [5] and Khodabakhsh et al. [56] concluded that human assessment strongly

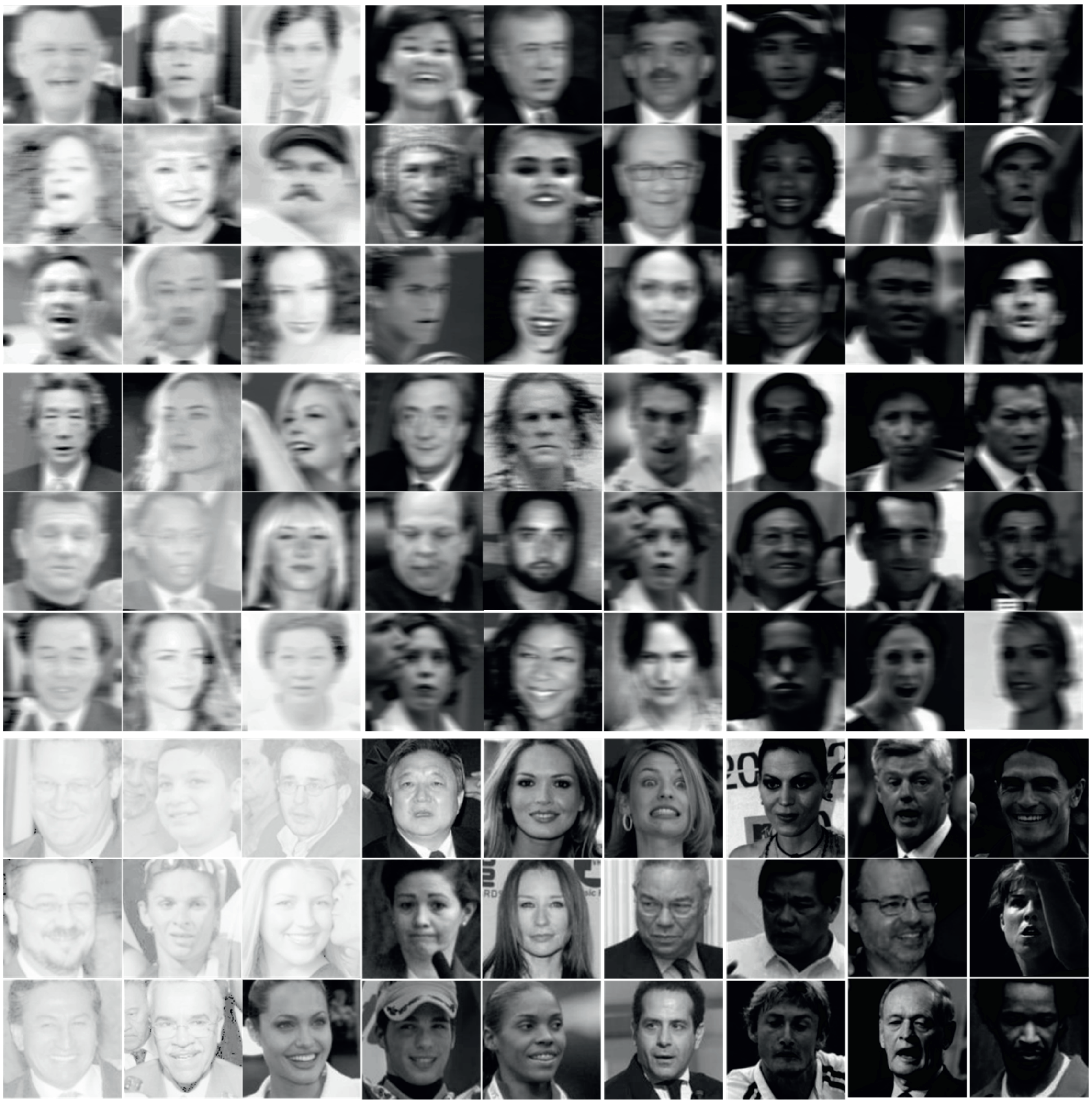


Figure 16: Classification results for LFW. The brightness dims from left to right. For example, the first row corresponds to "1bri," "1nor," and "1dark." The sharpness increases from top to bottom. For example, the first column corresponds to "1bri," "2bri," and "3bri."

correlates with FRA performance. It can also be concluded that each type of predicted image is correlated with the recognition performance defined in this paper.

5. Conclusions

We proposed a method of establishing FIQ labels based on brightness and sharpness that are strongly correlated to recognition and trained a model to predict quality. Overall, our model can accurately classify and distinguish images of different qualities, even for other FRAs that are not involved in the label creation and model training processes. We can also accurately evaluate the quality of FRAs mentioned in this paper on the cross-validation set. Note that an improvement in the classification accuracy of the model is needed to make further progress in the future. In addition, more factors affecting identification could be considered for adaptation to more varied application scenarios. In the future, the use of FIQA to improve the performance of image research projects is worth discussing.

Data Availability

Three datasets, M-PIE, LFW, and a custom dataset, are used in this paper, among which the first two are public datasets and the last one is owned by a technology limited company (Wisesoft Co., Ltd., Chengdu, Sichuan, China.). The custom dataset cannot be made publicly available because public availability would compromise privacy and we do not have permission to share the data. To replicate our method for other researchers, we also use the same method to perform experiments on a publicly available dataset (M-PIE), which interested readers can readily access. The M-PIE and LFW used can be found at http://multipie.org and http://vis-www.cs.umass.edu/lfw/, respectively.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key Research and Development Program of China, Research and Application Demonstration of Key Technology of Legal Identity Management (2016YFC0801100).

References

[1] O. A. Aghdam, B. Bozorgtabar, H. K. Ekenel, and J. Thiran, "Exploring factors for improving low resolution face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, pp. 2363–2370, Computer Vision Foundation/IEEE, Long Beach, CA, USA, June 2019.

[2] K. Cao, Y. Rong, C. Li, X. Tang, and C. C. Loy, "Pose-robust face recognition via deep residual equivariant mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 5187–5196, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, June 2018.

[3] S. Fekri-Ershad, "Gender classification in human face images for smart phone applications based on local texture information and evaluated kullback-leibler divergence," *Traitement du Signal*, vol. 36, no. 6, pp. 507–514, 2019.

[4] I. J. S. Biometrics, *Iso/iec 29794-1:2016 Information Technology - Biometric Sample Quality - Part 1:framework*, International Organization for Standardization, Geneva, Switzerland, 2016.

[5] L. B. Rowden and A. K. Jain, "Learning face image quality from human assessments," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 12, pp. 3064–3077, 2018.

[6] J. O. Hernandez, J. Galbally, J. Fiérrez, and L. Beslay, "Biometric quality: review and application to face recognition with faceqnet," 2020, https://arxiv.org/abs/2006.03298.

[7] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fiérrez, and C. Busch, "Face image quality assessment: a literature survey," 2020, https://arxiv.org/abs/2009.01103.

[8] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain, "Face recognition performance: role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[9] F. F. Alonso, J. Fierrez, D. Ramos, and J. Gonzalez-Rodriguez, "Quality-based conditional processing in multi-biometrics: application to sensor interoperability," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 6, pp. 1168–1179, 2010.

[10] G. Zhang and Y. Wang, "Asymmetry-based quality assessment of face images,"vol. 5876, pp. 499–508, in *Proceedings of the Advances in Visual Computing, 5th International Symposium, ISVC 2009*, vol. 5876, pp. 499–508, Springer, Las Vegas, NV, USA, November 2009.

[11] J. Yu, K. Sun, F. Gao, and S. Zhu, "Face biometric quality assessment via light CNN," *Pattern Recognition Letters*, vol. 107, pp. 25–32, 2018.

[12] R. Gross, I. A. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-pie," 2008. Gross Ralph, Matthews Iain, and Cohn Jeff et al. "Multi-PIE."" in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition*, vol. 28, no. 5, pp. 607–813, IEEE, Amsterdam, Netherlands, September 2008.

[13] G. B. Huang, M. Mattar, T. Berg, and E. M. Learned, "Labeled faces in the wild: a database for studying face recognition in unconstrained environments," Technical Report, University of Massachusetts., MA, USA, 2007.

[14] I. J. S. Biometrics, *ISO/IEC TR 29794-5:2010 Information Technology - Biometric Sample Quality - Part 5: Face Image Data*, International Organization for Standardization, Geneva, Switzerland, 2010.

[15] A. Howard, R. Pang, H. Adam et al., "Searching for mobilenetv3," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pp. 1314–1324, IEEE, Seoul, Korea (South), October 2019.

[16] S. Ahn, Y. Choi, and K. Yoon, "Deep learning-based distortion sensitivity prediction for full-reference image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, pp. 344–353, Computer Vision Foundation/IEEE, Nashville, TN, USA, June 2021.

[17] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "Pieapp: perceptual image-error assessment through pairwise preference," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 1808–1817, Computer Vision Foundation/IEEE Computer Society, Salt Lake City, UT, USA, June 2018.

[18] S. Golestaneh and L. J. Karam, "Reduced-reference quality assessment based on the entropy of DWT coefficients of locally weighted gradient magnitudes," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5293–5303, 2016.

[19] Z. Wan, K. Gu, and D. Zhao, "Reduced reference stereoscopic image quality assessment using sparse representation and natural scene statistics," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2024–2037, 2020.

[20] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 3474–3486, 2021.

[21] Y. Ma, W. Zhang, J. Yan, C. Fan, and W. Shi, "Blind image quality assessment in multiple bandpass and redundancy domains," *Digital Signal Processing*, vol. 80, pp. 37–47, 2018.

[22] K. Li, D. Shi, Y. Zhang, Q. M. J. Wu, X. Luan, and D. Song, "Cascnet: No-reference saliency quality assessment with cascaded applicability sorting and comparing network," *Neurocomputing*, vol. 425, pp. 231–242, 2021.

[23] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[24] F. T. Zohra and M. L. Gavrilova, "Adaptive face recognition based on image quality," in *Proceedings of the International 2017 Conference on Cyberworlds, CW 2017, Chester, United Kingdom*, pp. 218–221, IEEE Computer Society, Chester, UK, September 2017.

[25] L. B. Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.

[26] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proceedings of the in IEEE Conference on Computer Vision and*

Pattern Recognition, CVPR Workshops, pp. 74–81, IEEE Computer Society, Colorado Springs, CO, USA, June 2011.

[27] X. Gao, S. Z. Li, R. Liu, and P. Zhang, "Standardization of face image sample quality,"vol. 4642, pp. 242–251, in *Proceedings of the Advances in Biometrics, International Conference, ICB 2007*, vol. 4642, pp. 242–251, Springer, Seoul, Korea, August 2007.

[28] J. Sang, Z. Lei, and S. Z. Li, "Face image quality evaluation for ISO/IEC standards 19794-5 and 29794-5,"vol. 5558, pp. 229–238, in *Proceedings of the Advances in Biometrics, Third International Conference, ICB 2009*, vol. 5558, pp. 229–238, Springer, Alghero, Italy, June 2009.

[29] H. Kim, S. Lee, and Y. M. Ro, "Face image assessment learned with objective and relative face image qualities for improved face recognition," in *Proceedings of the IEEE International Conference on Image Processing, ICIP 2015*, pp. 4027–4031, IEEE, Quebec City, QC, Canada, September 2015.

[30] K. Nasrollahi and T. B. Moeslund, "Face quality assessment system in video sequences," vol. 5372, pp. 10–18, in *Proceedings of the Biometrics and Identity Management, First European Workshop, BIOID 2008*, vol. 5372, pp. 10–18, Springer, Roskilde, Denmark, May 2008.

[31] P. S. Wasnik, K. B. Raja, R. Raghavendra, and C. Busch, "Assessing face image quality for smartphone based face recognition system," in *Proceedings of the 5th International Workshop on Biometrics and Forensics, IWBF 2017*, pp. 1–6, IEEE, Coventry, United Kingdom, April 2017.

[32] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 90–94, 2015.

[33] L. Zhang, L. Zhang, and L. Li, "Illumination quality assessment for face images: a benchmark and a convolutional neural networks based model,"vol. 10636, pp. 583–593, in *Proceedings of the Neural Information Processing - 24th International Conference, ICONIP 2017*, vol. 10636, pp. 583–593, Springer, Guangzhou, China, November 2017.

[34] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1–9, IEEE Computer Society, Boston, MA, USA, June 2015.

[35] F. Yang, X. Shao, L. Zhang, P. Deng, X. Zhou, and Y. Shi, "DFQA: deep face image quality assessment,"vol. 11902, pp. 655–667, in *Proceedings of the Image and Graphics - 10th International Conference, ICIG 2019*, vol. 11902, pp. 655–667, Springer, Beijing, China, August 2019.

[36] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: a dataset and benchmark for large-scale face recognition,"vol. 9907, pp. 87–102, in *Proceedings of the Computer Vision - ECCV 2016 - 14th European Conference*, vol. 9907, pp. 87–102, Springer, Amsterdam, The Netherlands, October 2016.

[37] J. Hernandez-Ortega, J. Galbally, J. Fiérrez, R. Haraksim, and L. Beslay, "Faceqnet: quality assessment for face recognition based on deep learning," in *Proceedings of the 2019 International Conference on Biometrics, ICB 2019*, pp. 1–8, Crete, Greece, June 2019.

[38] 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 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018*, pp. 67–74, IEEE Computer Society, Xi'an, China, May 2018.

[39] L. Zhang, X. Shao, F. Yang, P. Deng, X. Zhou, and Y. Shi, "Multi-branch face quality assessment for face recognition," in *Proceedings of the 19th IEEE International Conference on Communication Technology, ICCT 2019*, pp. 1659–1664, IEEE, Xi'an, China, October 2019.

[40] N. Zhuang, Q. Zhang, C. Pan et al., "Recognition oriented facial image quality assessment via deep convolutional neural network," *Neurocomputing*, vol. 358, pp. 109–118, 2019.

[41] B. F. Klare, B. Klein, E. Taborsky et al., "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark A," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 1931–1939, IEEE Computer Society, Boston, MA, USA, June 2015.

[42] P. Terh¨orst, 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pp. 5650–5659, IEEE, Seattle, WA, USA, June 2020.

[43] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," 2021, https://arxiv.org/abs/2103.06627.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[45] L. B. Rowden and A. K. Jain, "Automatic face image quality prediction," 2017, https://arxiv.org/abs/1706.09887.

[46] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, 2018.

[47] W. J. Scheirer, P. J. Flynn, C. Ding et al., "Report on the BTAS 2016 video person recognition evaluation," in *Proceedings of the 8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2016*, pp. 1–8, IEEE, Niagara Falls, NY, USA, September 2016.

[48] A. Blanton, K. C. Allen, T. Miller, N. D. Kalka, and A. K. Jain, "A comparison of human and automated face verification accuracy on unconstrained image sets," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016*, pp. 229–236, IEEE Computer Society, Las Vegas, NV, USA, June 2016.

[49] R. Hsu, J. Shah, and B. Martin, "Quality assessment of facial images," in *Proceedings of the Biometric Consortium Conference*, Baltimore, MD, USA, September 2006.

[50] P. Mohammadi, A. E. Moghadam, and S. Shirani, "Subjective and objective quality assessment of image: a survey," 2014, https://arxiv.org/abs/1406.7799.

[51] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 4510–4520, IEEE Computer Society, Salt Lake City, UT, USA, June 2018.

[52] M. Tan, B. Chen, R. Pang et al., "Mnasnet: platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 2820–2828, Computer Vision Foundation/ IEEE, Long Beach, CA, USA, June 2019.

[53] B. Xu, A. Tulloch, Y. Chen, X. Yang, and L. Qiao, *Hybrid Composition with Idleblock: More Efficient Networks for Image Recognition*, https://arxiv.org/abs/1911.08609, 2019.

[54] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

[55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 770–778, IEEE Computer Society, Las Vegas, NV, USA, June 2016.

[56] A. Khodabakhsh, M. Pedersen, and C. Busch, “Subjective versus objective face image quality evaluation for face recognition,” in *Proceedings of the 3rd International Conference on Biometric Engineering and Applications, ICBEA 2019*, pp. 36–42, ACM, Stockholm, Sweden, May 2019.