Enhancing Face Quality Assessment through Age and Expression Analysis

Prateek Jaiswal*
CVIT
International Institute of information technology
Hyderabad
Hyderabad, Telangana, IN
prateek.jaiswal@research.iiit.ac.in

Anoop Namboodiri
CVIT
IIIT Hyderabad
Hyderabad, Telangana, IN
anoop@iiit.ac.in

Abstract

In facial recognition systems, quality extends beyond conventional perceptual quality to features incorporating identity information. Most facial image datasets embark on factors such as illumination and pose, making current systems robust enough to these factors with impressive recognition performance. Still, it is imperative to acknowledge that age variation and emotional similarity significantly influence identity. Variations in these features might significantly deceive the FR systems. These features also serve as easy channels for adversarial attacks on FR systems that alter facial features, such as morphing. Hence, making FR systems sensitive to the variations introduced over the range of these features is critical. We propose that the Unified Tri-Feature Quality Metric (U3FQ) be incorporated. This novel assessment framework integrates three critical elements: age variance, facial expression similarity, and congruence scores from state-of-the-art recognition models such as VGG-Face, ArcFace, FaceNet, and OpenFace. The weighting U3FQ utilizes an advanced learning paradigm, employing a Regression Network model for facial image quality assessment. U3FQ was rigorously evaluated against general IQA techniques—BRISQUE, BLINDS-II, RankIQA, and specialized FIQA methodologies like PFE, SER-FIQA, and SDD-FIQA. Results are backed up with qualitative analysis on the effectiveness of the generated quality scores through DET plots of FNMR on different age ranges, expression matches heat maps, and Expected Verification Rate (EVRC) curves on various datasets.

CCS Concepts

• **Computing methodologies** → **Computer Vision, Biometrics**.

Keywords

Fingerprint Image Quality, Fingerprint Recognition System, Image Quality Assessment, Weakly Supervised Learning

ACM Reference Format:
Prateek Jaiswal and Anoop Namboodiri. 2024. Enhancing Face Quality Assessment through Age and Expression Analysis. In *Indian Conference on Computer Vision Graphics and Image Processing (ICVGIP 2024), December 13–15, 2024, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3702250.3702251

*Corresponding author

ICVGIP 2024, December 13–15, 2024, Bengaluru, India

ACM ISBN 979-8-4007-1075-9/24/12
https://doi.org/10.1145/3702250.3702251

1 Introduction

Facial Image Quality Assessment (FIQA) is integral to the enhancement of face recognition (FR) systems, especially when dealing with the diversity of image quality often encountered in real-world scenarios. The potential of FIQA as a predictor for Face Recognition (FR) performance has been the primary motivation driving researcher interest, making it the focal point of current studies. Traditional FIQA approaches[19, 46] primarily assess the standalone biometric utility of images. However, in the context of FR, this method faces a conceptual challenge known as the *"Quality Paradox,"* as discussed by Schlett et al. [37] can be seen in figure 1. This paradox highlights the need to accurately reflect the reliability of comparison scores for image pairs that include the assessed image, thus adding a layer of complexity to FIQA's role in face recognition performance.

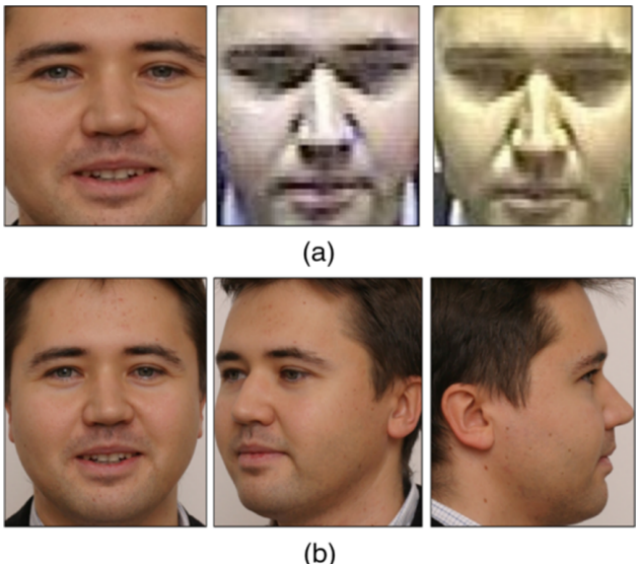

Figure 1: Displaying Image quality vs biometric quality. While the images (obtained from SCface database) in (a) are of poor image quality, the images in (b) may have lower biometric quality [8]

In recent advancements, FR techniques have shown remarkable results with high-quality frontal images and those of varying quality [10, 20, 29, 30, 33, 41, 42]. However, they still face significant hurdles in completely unconstrained environments [9, 45] where the quality of captured facial images cannot be guaranteed. FIQA methods strive to enhance the performance of FR systems in such settings by offering critical insights into the quality of input images. This input allows FR models to discern and possibly discard images of inferior quality that could lead to erroneous non-matches.

Modern FIQA (Face Image Quality Assessment) methods are generally categorized into two distinct styles: regression-based and model-based approaches. Regression-based methods[12, 23, 47] develop a direct mapping from the image space to quality labels, which are generated in a semi-automated manner. These labels often rely on comparison scores across matched image pairs or similarity scores between probe samples and reference images.

Conversely, model-based FIQA techniques[30, 42] integrate quality assessment directly within the face recognition (FR) model, evaluating quality based on the certainty or statistics derived from the generated facial features or embeddings. In face recognition models[5, 6, 13, 34, 38], match scores are determined by the distance between facial embeddings, which are optimized to differentiate between individuals (inter-class distance) and recognize the same individual across various images (intra-class distance).

The proposed approach extends this concept by incorporating the impact of age variations and emotional expressions on these distances. Understanding how aging and emotions alter facial features is vital, as they can significantly affect recognition accuracy. Additionally, the output of this process is a quality score, defined as the probability of finding the same image in the database. For instance, a quality score of 0.6 indicates a 60% probability of locating this image within the database.

By incorporating the effects of age and emotional expressions, our approach aims to enhance the reliability of facial recognition systems under various conditions, ensuring high accuracy despite changes in facial features.

This paper introduces the Unified Tri-Feature Quality (U3FQ) metric, a novel method in Facial Image Quality Assessment (FIQA). U3FQ integrates recognizability and quality estimation using a unique learning-based approach. Unlike traditional methods, it employs match scores in a weakly supervised manner as the primary quality indicator. The main contributions of our work include:

- Enhancing recognition reliability by integrating age and emotional expressions into FIQA.
- Introducing the U3FQ metric, which combines recognizability and quality estimation.
- Utilizing match scores weakly supervised as the core quality indicator.

2 Related Works

In this paper, we contextualize our work within the FIQA landscape, introducing the Unified Tri-Feature Quality (U3FQ) metric as a novel perspective in FIQA. Our approach, inspired by recent trends in unsupervised, semi-supervised, and regression-based learning, builds upon the advancements made by current state-of-the-art models such as MAGFace[30] and QMagFace[41]. While these models have significantly improved FIQA capabilities, our method uniquely integrates additional facial biometrics like age and expressions. This integration enriches the conventional FIQA framework, steering it towards more nuanced and holistic assessments.

2.1 Face Quality Assessment

There hasn't been a common standard for face quality in general, despite several advances in face quality assessment. The technical publications ICAO TR 9303 and ISO/IEC 39794-5, which attempt to define high-quality portrait-like photographs for use in official documents, are the most noteworthy works in this field. Nevertheless, these reports do not specify a particular quality metric; instead, they just offer recommendations for appropriate image capture. In contrast, the ISO/IEC 29794-4, which contains the NIST-developed NFIQ quality metric, provides a clear standard for the field of fingerprint identification.

For face recognition systems to function well, the quality of the facial image is crucial. Conventional techniques that evaluate quality based on picture features include Brisque[35], Niqe[31], and Piqe[43] as well as the ISO/IEC 19794-5 and ICAO 9303 standards.

2.1.1 ***Traditional methods in FIQA.*** Early face quality assessment methods relied on hand-crafted features to evaluate factors affecting recognition accuracy, such as pose, illumination, and blur. A notable example [24] calculated several quality measures using hand-crafted algorithms, combining them into two global measures: one for human perception and another for recognition accuracy.

This approach evolved into the Face Quality Index (FQI) [1], which combined quality factors from five image features to create a global accuracy-based measure. The authors simulated real-world variability by adding synthetic effects to original images.

BioLab-ICAO [16] introduced a method performing 30 individual tests for variability factors, returning a score for each. Unlike FQI, these tests aimed to ensure compliance with ISO/ICAO standards for Machine Readable Travel Documents. However, BioLab-ICAO did not combine individual scores into a single global measure, differentiating it from the FQI approach.

While hand-crafted and traditional learning-based image processing approaches can be advantageous due to their interpretability, as they are designed to measure specific image features such as blur, resolution, texture, and color, the results may not always be reliable. This is because these algorithms may not perform accurately in certain acquisition scenarios. Additionally, determining which of these features is most relevant for a particular task can be challenging.

2.1.2 ***Deep Learning Based methods.*** The advent of deep learning has revolutionized face quality assessment methodologies. A prime illustration is the work in [49], where researchers leveraged Convolutional Neural Networks (CNNs) to evaluate face image quality, focusing specifically on illumination conditions. This approach marks a significant shift from traditional hand-crafted feature extraction methods. Fundamentally aimed at significantly enhancing the recognizability of advanced face recognition systems. These innovative methods, exemplified by seminal works such as SER-FIQ[42], SDD-FIQA[33], PCNet[47], and [7], have effectively

demonstrated the efficiency of leveraging intrinsic data characteristics and a robust combination of both richly annotated and unannotated data. They underscore the substantial potential of utilizing advanced embedding variability analysis and sophisticated similarity distribution distancing strategies to comprehensively assess and evaluate facial image quality.

Drawing on the strengths of advanced computational techniques and human-perceivable facial attributes, the Unified Tri-Feature Quality (U3FQ) metric represents a sophisticated amalgamation of the finest elements in FIQA methodologies. U3FQ shares conceptual similarities with notable works like CR-FIQA [10], FaceQnet [23], and FaceQAN [4], but distinctively pushes the boundaries of conventional approaches. It incorporates a deeper, more nuanced integration of biometric analysis, transcending traditional computational assessments.

Acknowledging the importance of facial expressions, as highlighted in studies [11, 26, 39], U3FQ integrates these aspects into its framework. Additionally, it draws on the biometric significance of facial age features, as detailed in research [3, 14, 18, 40], demonstrating the impact of age characteristics on recognition.

3 Methodology

Our method for creating U3FQ combines deep learning frameworks that are intended to analyse and interpret facial data. The impact of age and expression on match scores can now be quantitatively measured thanks to this integration. This allows U3FQ to provide a thorough evaluation tool that extends standard metrics by taking advantage of the shifting traits of human faces.

3.1 Theoretical Background

Facial Age Difference: The efficacy of face matching systems is significantly influenced by the age difference between the anchor image and the comparison image, as illustrated in Figure 3. This influence varies notably with the anchor's age, necessitating a nuanced approach to modeling age difference penalties. For anchors aged between 20 and 30 years, negative age differences typically correlate with child images, which present a considerable challenge due to the substantial change in facial features that occur during maturation. Conversely, for anchors over 35 years of age, negative age differences represent younger adult images, where changes in facial features are less pronounced.

To empirically underpin this observation, we analyse Detection Error Tradeoff (DET) plots that demonstrate the variance in performance with different age groups for all four models: VGG-Face[44], OpenFace[2], ArcFace[13], and FaceNet[38]. Due to page limitations, these plots are included in the supplementary material where, we have added the DET plots from VGG-Face, OpenFace, ArcFace and FaceNet that shows the False Non-Match Rate (FNMR) for different age groups. These plots highlight that there is a pronounced increase in FNMR as the age difference becomes more negative. The trend gradually inverts with increasing anchor age, reflecting the maturation and stabilization of facial features over time.

The Influence of Facial Expressions: The similarity in facial expressions between two images notably influences recognition performance, as variations in expressions can distort critical facial features used in establishing a match. This impacts the overall quality of recognition. Figure 5 illustrates the impact that discrepancies in facial expressions have on matching performance, evidenced by average match scores across expression pairs.

Our methodology ensures a more refined and context-sensitive assessment of facial expression similarity, taking into account not just the physical resemblance but also the nuanced expressive context of each face. This approach leads to a more accurate and realistic evaluation of facial images, particularly relevant in dynamic real-world scenarios where facial expressions can vary significantly shown its calculation in Figure 4.

3.2 Formulations and Optimization

Building on observations from empirical evidence, we developed a mathematical model that incorporates a logistic adjustment to account for the non-linear impact of age differences on facial similarity scores. The adjusted match score function is defined as:

$$f(d,a) = \begin{cases} \frac{\Lambda}{1+e^{-\kappa(\xi-\xi_0)}} & \text{if } a \leq 30, \\ \frac{\Lambda}{1+e^{-\kappa(\xi-\xi_0)}} & \text{if } a > 30, \end{cases} \tag{1}$$

where:

- $\xi = \alpha d + \beta a + \gamma d^2$ for $a \leq 30$,
- $\xi = \delta d + \epsilon a + \zeta \log(\max(a,1)) + \eta da$ for $a > 30$,
- d represents the age difference between the anchor and the comparison image,
- a denotes the anchor's age,
- Λ is the curve's maximum value,
- κ is the logistic growth rate,
- ξ_0 is the x-value of the sigmoid's midpoint,
- Parameters α, β, γ, δ, ϵ, ζ, and η control the function's shape.

The basis of Equation (1) is to model the non-linear influence of age variations on facial similarity. Empirical analysis, as discussed in Section 3.1, shows that younger anchors (ages 20-30) display greater sensitivity to age differences, affecting match scores more significantly than older anchors. This is represented by the function's formulation, where the parameters α, β, and γ for younger anchors ensure a higher response to variations in age difference, while parameters δ, ϵ, ζ, and η for older anchors account for a more gradual impact.

The methodology also integrates the influence of facial expressions on match scores through the expression impact function $g(e)$:

$$g(e) = \begin{cases} c & \text{if } e \text{ is a weak emotion}, \\ d \cdot \text{EXPR\_SCORE}(e) & \text{if } e \text{ is a strong emotion}, \end{cases} \tag{2}$$

where c is a constant for weak emotions, and d scales the expression score EXPR_SCORE(e) for strong emotions.

Basis for Classifying Emotional Expressions as Weak. Expressions such as happy, sad, and fear are classified as "weak" in this context due to their relatively predictable and less disruptive nature on facial recognition models. While these emotions reflect genuine affective states, they do not substantially alter key facial landmarks or introduce significant variability in feature extraction. For instance:

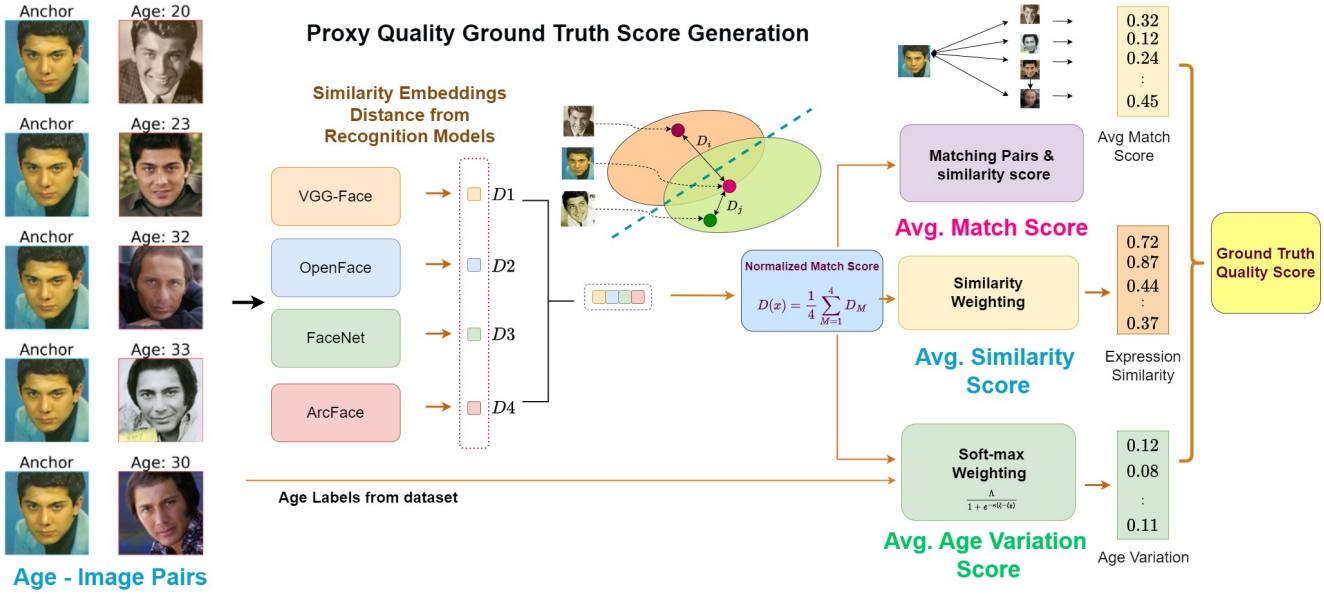

Figure 2: The figure presents a method for generating pseudo ground truth labels in face recognition by assessing age-related variations and expression similarity. It starts by calculating similarity distances between images of the same individuals at different ages using face recognition models. These distances are then normalized and combined with age and expression data to get Avg. Match Score, Avg. Similarity Scores and Avg Age Variation Scores. These Scores are combined based on weighting to provide a combined score that is used are for fine-tuning regression network, leading to a comprehensive quality score that encapsulates recognition accuracy, age differences, and expression similarities.

- **Happiness** involves minor changes such as slight eye narrowing and mouth curvature, which maintain recognizable facial structure.
- **Sadness** typically results in subtle changes like slight frowning or drooping eyelids, which do not obscure the face's core geometry.
- **Fear** can present with widened eyes and a slightly open mouth, but these adjustments are often less pronounced than expressions involving extensive facial distortions.

By contrast, strong emotions such as surprise or anger create more substantial, non-linear changes in facial features, making them more challenging for recognition systems and requiring adjusted scaling in quality assessments.

Integration of Metrics. Once $f(d, a)$ and $g(e)$ are computed, they are combined with the Average Match Score to derive the Average Age Variation Score and Average Emotion Similarity Score:

$$\text{Avg. Age Variation Score} = \text{Integration}(f(d, a), \text{Avg. Match Score}),$$
$$\text{Avg. Similarity Score} = \text{Integration}(g(e), \text{Avg. Match Score}).$$

These integrated scores provide a comprehensive measure of facial image quality, capturing the complex interplay between age, emotional expression, and overall image congruence. This approach ensures that our FIQA model accurately reflects variations due to aging and emotions, enhancing its robustness and applicability in real-world biometric systems.

Supplementary Documentation. Further computational details and in-depth calculations are provided in the supplementary materials. This additional documentation offers a detailed exploration of the U3FQ model's algorithmic processes, and readers interested in technical specifics and data analyses are encouraged to refer to it.

3.3 Architecture

3.3.1 Detailed Explanation of Distance Metrics and Embeddings. The U3FQ algorithm begins by calculating a match score distance d for an input image I using a set of face recognition models $M = \{M_1, M_2, M_3, M_4\}$. These models, which include VGG-Face, ArcFace, FaceNet, and OpenFace, each operate within unique embedding spaces. This diversity in embedding spaces means that the feature vectors derived from these models can differ significantly in terms of dimensionality, scale, and the type of facial features they emphasize.

1. Feature Extraction and Embedding Spaces. Each model M_i extracts a feature vector $\mathbf{v}_i$ from the input image I. These vectors represent different facets of facial features:

- **VGG-Face** focuses on general facial structures and details.
- **ArcFace** embeds features with a strong emphasis on angular distance, providing a robust metric for face verification.
- **FaceNet** optimizes embeddings to maximize Euclidean distance between different identities and minimize it for the same identity.
- **OpenFace** utilizes deep learning for embeddings that are optimized for real-world variations and recognition tasks.

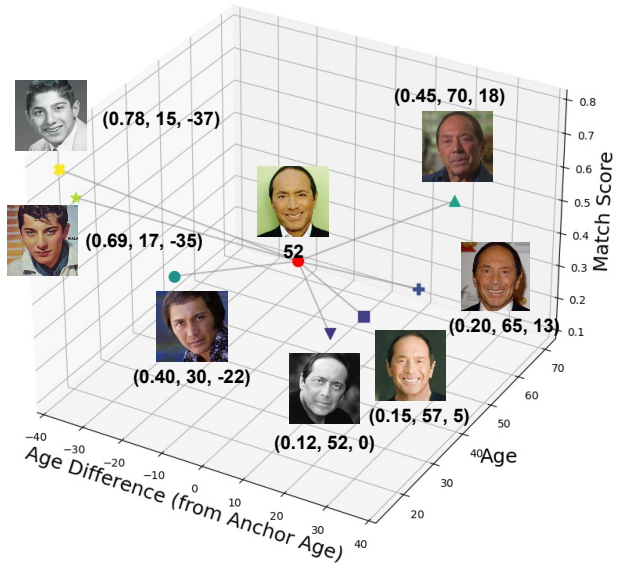

Figure 3: The efficacy of face matching systems is significantly impacted by the noticeable age variation between the images being compared. The comprehensive triplet representation emphasizes the similarity distance, the specific age of the compared image, and the notable age difference in relation to the anchor image, with Image 6 serving as the reference.

Due to these differences, the computed distances between images vary across models. For example, distances in the ArcFace space may be smaller due to its margin-based loss functions, while VGG-Face may have larger distances reflective of its learned features.

2. Normalization Process. Given the discrepancies in scale and interpretation across the embedding spaces, normalization is essential. Each computed match score distance d is adjusted using:

$$\text{Normalized}(d) = \frac{d - \mu_M}{\sigma_M}$$

where μ_M and σ_M represent the mean and standard deviation of distances for model M. This ensures comparability across models by aligning their distributions to a common scale.

The normalized match score distance is aggregated as:

$$\text{NAMS} = \frac{\sum_{\text{model} \in M} \text{Normalized}(d)}{|M|}$$

The universal threshold of 0.476, determined empirically, balances precision and recall across models, accommodating criteria such as ArcFace's threshold of 0.6 and VGG-Face's 0.4.

3. Computation of Age and Emotion Scores. The age and emotion scores are computed as follows:

- **Age Difference Score** $f_{\text{age}}(NAMS, a, d)$ modifies NAMS based on the known or predicted age a of the image. The weight of 0.7 emphasizes age's significant effect on recognition performance.

Algorithm 1 U3FQ: Unified Tri-Feature Quality Assessment for Contextual Facial Image Quality

Require: Single input image I, ResNet model RN, age a, expression e, match score distance models $M = \{M_1, M_2, M_3, M_4\}$
Ensure: U3FQ Score or Quality Score

$S \leftarrow 0$
MatchScore $\leftarrow 0$
NAMS $\leftarrow 0$ ▷ Normalized Average Matching Score
NAVS $\leftarrow 0$ ▷ Normalized Age Variation Score
NESS $\leftarrow 0$ ▷ Normalized Emotion Similarity Score
for all model $\in M$ **do**
  $d \leftarrow$ ComputeMatchScoreDistance$(I, model)$
  MatchScore $\leftarrow$ MatchScore + Normalize(d)
end for
NAMS $\leftarrow$ Average(MatchScore)$/(0.476)$
for all model $\in M$ **do**
  AgeDiffScore $\leftarrow$ AgeDiffScore + f_{age}(NAMS, a, d)
  EmotionSimScore $\leftarrow$ EmotionSimScore + f_{emotion}(NAMS, e)
end for
$S \leftarrow 0.1 \cdot$ NAMS $+ 0.7 \cdot$ NAVS $+ 0.2 \cdot$ NESS
procedure U3FQ_Assessment(I, RN, $m = 100$)
  QualityScores $\leftarrow$ []
  for $i \leftarrow 1$ to m **do**
    quality $\leftarrow RN$.Predict(I, S)
    QualityScores $\leftarrow$ QualityScores + [quality]
  end for
  finalQuality $\leftarrow$ Average(QualityScores) **return** finalQuality
end procedure

- **Emotion Similarity Score** $f_{\text{emotion}}(NAMS, e)$ adjusts the score according to facial expressions e, with stronger emotions having a higher scaling factor.

The final composite quality score S is:

$$S = 0.1 \cdot \text{NAMS} + 0.7 \cdot \text{NAVS} + 0.2 \cdot \text{NESS}$$

Empirical Validation and Robustness. The algorithm's robustness is validated using stochastic embeddings generated by the ResNet model RN over m iterations, providing an average quality score Q that reflects image stability and robustness under facial variations.

3.4 Regression Network and Quality Estimation

We have advanced and thoroughly refined an existing Convolutional Neural Network (CNN), originally pre-trained extensively for face recognition tasks, through a meticulous process of fine-tuning. This established approach of expertly adapting deep learning models to tasks closely akin to their initial training has been consistently and effectively demonstrated in numerous influential studies. Such versatile networks have been successfully repurposed for detecting a wide range of facial attributes distinct from identity, including gender, age, and race. In the specific context of comprehensive face quality assessment, it is firmly posited that a robust feature vector containing highly discriminative facial information should inherently encapsulate critical aspects of image quality.

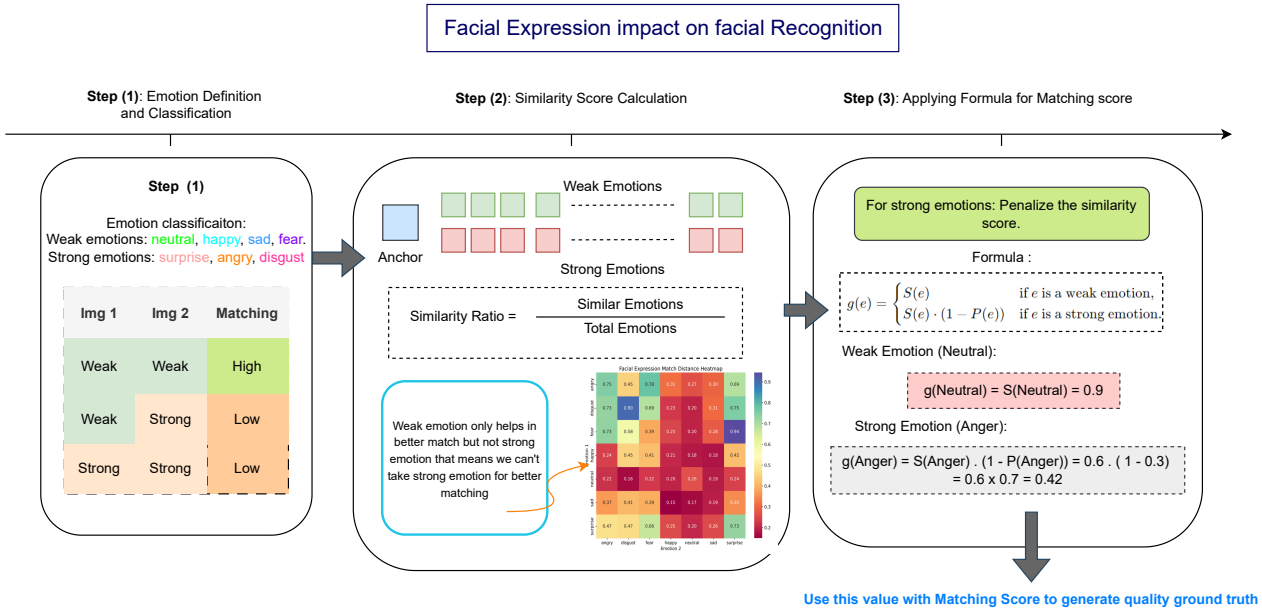

Figure 4: Calculating and Integrating Facial Expression Similarity with Face Similarity Distance

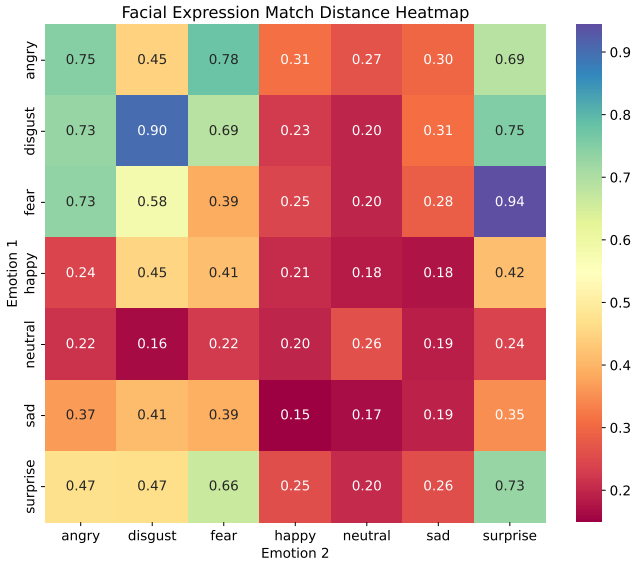

Figure 5: The differential impact of facial expressions on the match score is notable, with weak emotions having a relatively constant effect and strong emotions significantly modifying the score proportionally to their intensity.

For our specific adaptation, we selected the ResNet50 architecture as the foundational network. During the fine-tuning process, we removed the classification layers and augmented the network with fully connected layers, which were then fused with the existing feature vector. This amalgamation was subjected to a sigmoid activation function, designed to yield a quality score.

Crucially, we implemented a training strategy where the weights of the pre-existing layers were frozen, ensuring that only the newly integrated layers were subject to training. This training utilized the pseudo ground truth quality labels generated in the preceding step. The outcome of this refined model is a quality score, ranging from 0 to 1, which correlates with the performance of face recognition, offering a robust measure of the quality of facial images in terms of recognition efficacy.

4 Experiments And Results

Table 1: Summary of the Experimental Setup

Dataset	#Images	#IDs	Main Quality Factors (‡)		
			P-I	AV-E	N-D
AgeDB [32]	16,487	568	H	H	M
Adiance [15]	5,000	1,159	H	H	L
LFW [25]	5,000	1,135	M	H	H
MEDSII [17]	1,306	518	M	H	L

Notes:
† P-I: Pose and Illumination; AV-E: Age-Variation, Expression; N-D: Other Noise & Distortions - Scale.
‡ L: Low; M: Medium; H: High; Lr: Large; Values estimated subjectively by the authors.

In our comprehensive study, the AgeDB dataset, as cited in Moschoglou et al. [32], plays a critical role. This dataset, comprising

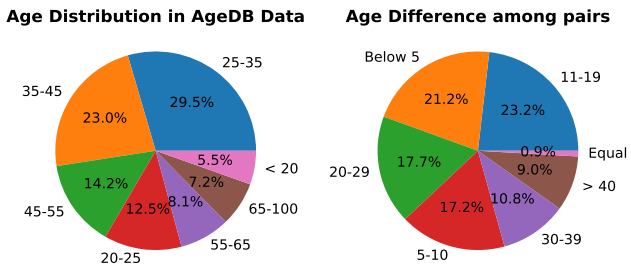

Figure 6: Distribution of Age-groups in AgeDB dataset.

16,487 images, serves as a foundational resource for examining age variations across different identities. A key visual element in our analysis is presented in Figure 6. This figure is composed of two informative pie charts. The first chart offers a detailed illustration of the age group distribution within the AgeDB dataset, providing a clear overview of the demographic composition. The second chart is particularly insightful, highlighting the age differences between pairs of images. This aspect is fundamental for understanding and improving identity matching in the context of age-related changes.

Figure 6 is pivotal in our study, illustrating age variation in different images of the same individual—a crucial element for evaluating age-invariant facial recognition systems. Additionally, it categorizes the age groups in the AgeDB dataset, highlighting the age diversity critical to our analysis. This visual representation is key to understanding the challenges in age-variant facial recognition, aiding in the development of more accurate systems.

As detailed in Table 1, key to our analysis, we generated approximately 279,000 pairs from AgeDB to cover a wider range of identities. For each identity, an average match score was computed from about 20 images. This approach allows for in-depth insights into age-related identity matching.

Additionally, we include the LFW [25] and Adience [15] datasets in the table, while MEDSII [17] is presented in the distribution but not included in the table. These datasets provide diverse facial images, enabling a comprehensive analysis and demonstrating the robustness of our methodologies in age-variant facial recognition.

4.1 Implementation Details and Setup

Our computational network is developed using the PyTorch framework, following the same implementation principles as described in [33], and operates on a machine equipped with four NVIDIA GeForce RTX 2080 Ti GPUs. For preprocessing, face images are uniformly aligned, scaled, and cropped to a resolution of 112×112 pixels utilizing the MTCNN algorithm, as detailed in [48]. During the training phase, all networks undergo optimization using the Adam optimizer, with a weight decay parameter set to 1×10^{-4}. The training process starts with an initial learning rate of 1×10^{-3}, which is subsequently reduced by a factor of 5×10^{-2} after every 5 epochs. This systematic adjustment in the learning rate ensures efficient convergence and optimal network performance.

In terms of runtime performance, the U3FQ algorithm demonstrated significant efficiency. The quality assessment for 16,000 images was completed in approximately 30 minutes, leveraging the power of the four NVIDIA GeForce RTX 2080 Ti GPUs for parallel processing. This runtime encompasses feature extraction, distance calculation, and metric computation, averaging to around 0.1125 seconds per image. Such performance highlights the algorithm's scalability and suitability for large-scale biometric applications, where processing time is a critical factor. The use of a multi-GPU setup ensures that feature extraction using models such as VGG-Face [44], FaceNet [38], ArcFace [13], and OpenFace [6] is executed concurrently, maximizing throughput and reducing computation time.

We compared U3FQ with various state-of-the-art Image Quality Assessment methods, including BRISQUE [31], BLINDSII [36], RankIQA [27], PFE [23], SDD-FIQA [33], and SER-FIQA [42]. Our experiments employed the aforementioned Face Recognition (FR) models for score computation. Additionally, we used MobileFaceNet

Table 2: AOC at FMR of 1×10^{-2} , 1×10^{-3} and 1×10^{-4}. The Blue color text indicates the best overall performance, whereas green represents the second best in comparison, and red signifies the lowest performance.

LFW				
Method	FMR@1e-2	FMR@1e-3	FMR@1e-4	Avg
BRISQUE [31]	0.0467	0.0900	0.1279	0.1127
BLINDS-II [36]	0.1944	0.2354	0.2765	0.2612
RankIQA [27]	0.1346	0.1120	0.1459	0.1435
PFE [23]	0.2035	0.2557	0.2905	0.2499
SDD-FIQA [33]	0.8101	0.7881	0.7784	0.7979
SER-FIQA [42]	0.5673	0.6534	0.7477	0.6701
QMAGFACE [41]	0.7956	0.8232	0.7734	8190
U3FQ (Ours)	0.8160	0.7653	0.7880	0.8035
Adience				
Method	FMR@1e-2	FMR@1e-3	FMR@1e-4	Avg
BRISQUE [31]	0.1845	0.2103	0.2412	0.2235
BLINDS-II [36]	0.1856	0.1546	0.1476	0.1710
RankIQA [27]	0.3412	0.2978	0.2876	0.3063
PFE [23]	0.3526	0.2768	0.2823	0.2870
SDD-FIQA [33]	0.5970	0.6423	0.5720	0.5996
SER-FIQA [42]	0.5123	0.5687	0.4562	0.4890
QMAGFACE [41]	0.6856	0.6232	0.6234	0.6390
U3FQ (Ours)	0.7036	0.6782	0.5610	0.6539
AgeDB				
Method	FMR@1e-2	FMR@1e-3	FMR@1e-4	Avg
BRISQUE [31]	0.2856	0.3235	0.3656	0.3123
BLINDS-II [36]	0.3781	0.3452	0.3708	0.3689
RankIQA [27]	0.3215	0.3076	0.2765	0.2887
PFE [23]	0.3892	0.3187	0.2956	0.3054
SDD-FIQA [33]	0.7292	0.7238	0.7563	0.7320
SER-FIQA [42]	0.6238	0.5982	0.6286	0.6129
QMAGFACE [41]	0.7156	0.7232	0.8234	0.7260
U3FQ (Ours)	0.7630	0.7432	0.7412	0.7520

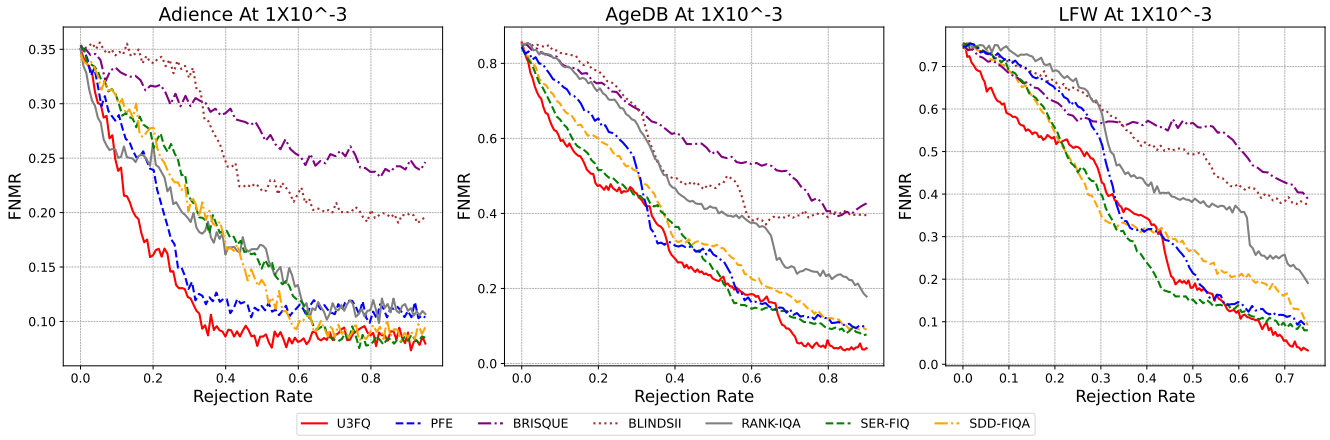

Figure 7: Effectiveness of Low-Quality Face Image Rejection in Face Verification: The EVRC (Expected Verification Rate Curve) Graphically Demonstrating FNMR (False Non-Match Rate) at a 1e-3 FMR (False Match Rate) Threshold Based on Predicted Quality Scores

as the backbone for our method, emphasizing its efficiency and real-world applicability.

4.2 Evaluation Metrics

In our study, the performance evaluation of the U3FQ was conducted by plotting the Error-Reject Curve (ERC). The ERC is a well-established method for representing Face Image Quality Assessment (FIQA) performance, as documented in the literature [21, 22]. It effectively demonstrates the impact of discarding a proportion of face images—specifically those of the lowest quality—on the face verification performance. This impact is measured in terms of the False Non-Match Rate (FNMR) [28] at a predetermined threshold, set at a constant False Match Rate (FMR) [28]. For our analysis, the ERC curves for all benchmarks were plotted at two fixed FMRs: 1e-3, as recommended for border control operations by Frontex, and 1e-4, details of which are included in the supplementary material. Additionally, we quantified the verification performance using the Area Over the Curve (AOC) of the ERC. This provides a comprehensive, aggregate performance across all rejection ratios.

In the evaluation of U3FQ, as detailed in Table 2 and Figure 7, the algorithm was compared against general Image Quality Assessment (IQA) methods such as BRISQUE, BLINDS-II, and RankIQA, and specialized Face Image Quality Assessment (FIQA) methods like PFE, SER-FIQA, and SDD-FIQA.

AUC (Area Under the Curve). The AUC quantifies classifier performance by measuring the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). AUC values closer to 1 indicate highly effective classification, reflecting accurate distinction between classes. In FIQA, a high AUC demonstrates the method's ability to rank images in line with true verification outcomes.

TAR (True Accept Rate). The TAR measures the proportion of correctly identified genuine matches at a given False Match Rate (FMR), critical in biometric verification. For example, a TAR at an FMR of 0.001 indicates the method's reliability under strict conditions. Higher TAR at low FMRs signifies better performance, showing consistent accuracy at stringent security thresholds.

Evaluation Insight. Using AUC and TAR provides a dual perspective: AUC shows overall classification capability, while TAR assesses performance at specific points relevant to biometric systems. This ensures a comprehensive evaluation of U3FQ's comparative strengths.

5 Conclusion

Through the Unified Tri-Feature Quality Metric (U3FQ), we propose a pivotal advancement in the domain of Facial Image Quality Assessment (FIQA). By integrating age variance and facial expression impact, U3FQ presents a novel and comprehensive method for evaluating facial images. This research emphasizes the significance of these biometric features in enhancing the accuracy and reliability of recognition models, thereby transcending the conventional FIQA metrics that predominantly rely on subjective human visibility assessments. Through rigorous evaluations on an extensive set of face quality image datasets and benchmark comparisons with state-of-the-art techniques, U3FQ has demonstrated its superiority in delivering relevant and precise quality assessments. Looking ahead, our future work aims to augment the predictive power of U3FQ with additional features such as illumination and pose to further refine the accuracy of reference quality labels, ensuring that U3FQ remains at the forefront of FIQA methodologies. We intend to broaden the scope and effectiveness of U3FQ making it an even more robust tool for assessing facial image quality in diverse and challenging recognition scenarios under new version of UXFQ.

References

[1] Ayman Abaza, Mary Ann Harrison, and Thirimachos Bourlai. 2012. Quality metrics for practical face recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 3103–3107.

[2] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* 6, 2 (2016), 20.

[3] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. 2018. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing* 2018, 1 (2018), 1–35.

[4] Žiga Babnik, Peter Peer, and Vitomir Štruc. 2022. Faceqan: Face image quality assessment through adversarial noise exploration. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 748–754.

[5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[6] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 59–66.

[7] Lacey Best-Rowden and Anil K Jain. 2017. Automatic face image quality prediction. *arXiv preprint arXiv:1706.09887* (2017).

[8] Samarth Bharadwaj, Mayank Vatsa, and Richa Singh. 2014. Biometric quality: a review of fingerprint, iris, and face. *EURASIP journal on Image and Video Processing* 2014 (2014), 1–28.

[9] Fadi Boutros, Naser Damer, Jan Niklas Kolf, Kiran Raja, Florian Kirchbuchner, Raghavendra Ramachandra, Arjan Kuijper, Pengcheng Fang, Chao Zhang, Fei Wang, et al. 2021. MFR 2021: Masked face recognition competition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–10.

[10] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. 2023. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5836–5845.

[11] Andrew J Calder and Andrew W Young. 2016. Understanding the recognition of facial identity and facial expression. *Facial Expression Recognition* (2016), 41–64.

[12] Kai Chen, Taihe Yi, and Qi Lv. 2021. Lightqnet: Lightweight deep face quality assessment for risk-controlled face recognition. *IEEE Signal Processing Letters* 28 (2021), 1878–1882.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[14] Natalie C Ebner. 2008. Age of face matters: Age-group differences in ratings of young and old faces. *Behavior research methods* 40 (2008), 130–136.

[15] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security* 9, 12 (2014), 2170–2179.

[16] Matteo Ferrara, Annalisa Franco, Dario Maio, and Davide Maltoni. 2012. Face image conformance to iso/icao standards in machine readable travel documents. *IEEE Transactions on Information Forensics and Security* 7, 4 (2012), 1204–1213.

[17] Andrew Founds, Nick Orlans, Whiddon Genevieve, and Craig Watson. 2011. NIST Special Databse 32 - Multiple Encounter Dataset II (MEDS-II). https://doi.org/10.6028/NIST.IR.7807

[18] Yun Fu, Guodong Guo, and Thomas S Huang. 2010. Age synthesis and estimation via faces: A survey. *IEEE transactions on pattern analysis and machine intelligence* 32, 11 (2010), 1955–1976.

[19] Javier Galbally, Sébastien Marcel, and Julian Fierrez. 2013. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE transactions on image processing* 23, 2 (2013), 710–724.

[20] Klemen Grm and Vitomir Struc. 2018. Deep face recognition for surveillance applications. , 46–50 pages.

[21] Patrick Grother and Elham Tabassi. 2007. Performance of biometric quality measures. *IEEE transactions on pattern analysis and machine intelligence* 29, 4 (2007), 531–543.

[22] Patrick J Grother, Patrick J Grother, Mei Ngan, and K Hanaoka. 2014. *Face recognition vendor test (FRVT)*. US Department of Commerce, National Institute of Standards and Technology.

[23] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. 2019. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics (ICB)*. IEEE, 1–8.

[24] Rein-Lien Vincent Hsu, Jidnya Shah, and Brian Martin. 2006. Quality assessment of facial images. In *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*. IEEE, 1–6.

[25] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[26] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* 13, 3 (2020), 1195–1215.

[27] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*. 1040–1049.

[28] A Mansfield. 2006. Information technology–biometric performance testing and reporting–part 1: Principles and framework. *ISO/IEC* (2006), 19795–1.

[29] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. 2021. Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security* 16 (2021), 4147–4183.

[30] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. 2021. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14225–14234.

[31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* 20, 3 (2012), 209–212.

[32] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. 2017. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 51–59.

[33] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. 2021. SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7670–7679.

[34] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

[35] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence* 22, 10 (2000), 1090–1104.

[36] Michele A Saad, Alan C Bovik, and Christophe Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing* 21, 8 (2012), 3339–3352.

[37] Torsten Schlett, Christian Rathgeb, Olaf Henniger, Javier Galbally, Julian Fierrez, and Christoph Busch. 2022. Face image quality assessment: A literature survey. *ACM Computing Surveys (CSUR)* 54, 10s (2022), 1–49.

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[39] Caifeng Shan, Shaogang Gong, and Peter W McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing* 27, 6 (2009), 803–816.

[40] Gillian Slessor, Deborah M Riby, and Ailbhe N Finnerty. 2013. Age-related differences in processing face configuration: The importance of the eye region. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 68, 2 (2013), 228–231.

[41] Philipp Terhörst, Malte Ihlefeld, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran Raja, and Arjan Kuijper. 2023. Qmagface: Simple and accurate quality-aware face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3484–3494.

[42] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. 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*. 5651–5660.

[43] Narasimhan Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani. 2015. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*. IEEE, 1–6.

[44] VGG. [n.d.]. The Visual Geometry Group (VGG) at the University of Oxford. ([n. d.]).

[45] Mei Wang and Weihong Deng. 2021. Deep face recognition: A survey. *Neurocomputing* 429 (2021), 215–244.

[46] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian C Lovell. 2011. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPR 2011 WORKSHOPS*. IEEE, 74–81.

[47] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. 2020. Inducing predictive uncertainty estimation for face recognition. *arXiv preprint arXiv:2009.00603* (2020).

[48] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[49] Lijun Zhang, Lin Zhang, and Lida Li. 2017. Illumination quality assessment for face images: A benchmark and a convolutional neural networks based model. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part III 24*. Springer, 583–593.