Advancing FIQA with Intrinsic Age Features: Introducing the U3FQ (Unified Tri-Feature Quality) Metric

Anonymous CVPR submission

Paper ID 12988

Abstract

Facial Image Quality Assessment (FIQA) plays a pivotal role in the enhancement of face matching and recognition systems. Traditional FIQA metrics tend to focus on subjective human visibility, which may not align with the features essential for accurate recognition. To bridge this gap, we propose the Unified Tri-Feature Quality Metric (U3FQ), a novel assessment framework that integrates three critical elements: age variance, facial expression impact, and congruence scores from four state-of-the-art recognition models. Unlike conventional metrics, U3FQ employs a multi-task learning paradigm with a ResNet model trained to perform the dual tasks of direct facial image quality assessment and facial age prediction simultaneously. This innovative approach achieves a mean absolute error (MAE) of ±2 years for both genders in age prediction while ensuring a robust facial quality metric. Our method refines congruence scores with quantitative modifiers that account for age discrepancies and expression intensities, ensuring a quality metric that more accurately predicts the likelihood of recognition success. We rigorously evaluated U3FQ against general IQA techniques—BRISQUE, NIQE, and PIQE—as well as specialized FIQA methodologies such as FaceQnet, SER-FIQ, and MagFace. The results confirm that U3FQ marks a substantial advancement in the field of FIQA, offering a holistic and theoretically sound quality assessment tool that is acutely relevant to a wide array of facial recognition scenarios.

1. Introduction

Facial Image Quality Assessment (FIQA) is instrumental in enhancing the performance of face recognition systems by appraising the quality of facial images. Traditional FIQA methods predominantly aim at predicting the biometric utility of standalone images, yielding a solitary quality score per image. Nonetheless, within the face recognition milieu, this approach encounters a conceptual dilemma termed the "quality paradox" mentioned in [14]. This paradox pertains to the necessity of denoting the precision or reliability of comparison scores for sample pairs encompassing the evaluated image, thereby imposing unique considerations on FIQA in the face recognition domain.

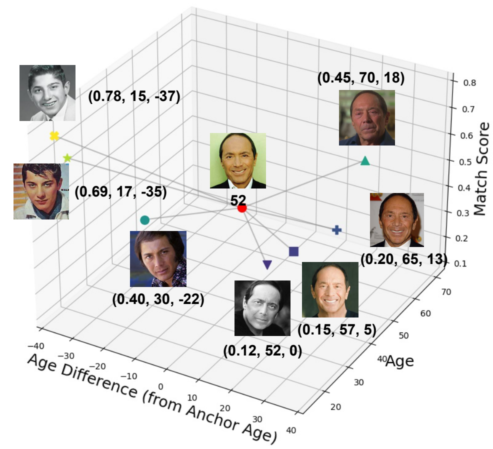

Figure 1. The efficacy of face matching systems is significantly impacted by the age variation between the images being compared. The triplet representation emphasizes the similarity distance, the age of the compared image, and the age difference in relation to the anchor image, with Image 6 as the reference.

The influence of facial attributes, such as age and expressions on face recognition, has been extensively studied. The research by Schlett et al. [14] investigates the intricacies of the quality paradox, underscoring the intricate manner in which FIQA methodologies are influenced. Significantly, while FIQA strategies utilizing conventional Image Quality Assessment (IQA) techniques may not directly correlate with face recognition utility, those predicated on veritable quality scores grapple with addressing the quality paradox when producing scores.

In our work, we introduce a novel paradigm, the Unified Tri-Feature Quality (U3FQ) metric, which resolves the quality paradox by integrating intrinsic age characteristics and facial expressions into FIQA.

Facial Expression Categorization

Acknowledging the pivotal role of facial expressions in the accuracy of face recognition, we implement a binary classification scheme for facial expressions—distinguishing between congruent and disparate expressions. This binary approach is intended to condense our examination and underscore the influence of expressions on quality assessment.

Age Disparity Integration

We incorporate age disparity as a core element of our FIQA metric, acknowledging its significant impact on the veracity of face recognition systems. This inclusion is crucial in devising a metric that captures the true nature of quality within a face recognition scenario.

In this paper, we introduce a groundbreaking learning-based method known as the Unified Tri-Feature Quality (U3FQ) metric, reconceptualizing Facial Image Quality Assessment (FIQA) through the prism of recognizability and quality estimation. Our approach diverges from conventional paradigms by employing match scores in a weakly supervised manner, serving as our quality metric. The salient contributions of our work are multifold:

- We are the first to systematically analyze the impact of facial age on the performance of recognition systems within the FIQA framework.
- Our method uniquely incorporates facial expression similarity into the assessment of recognition performance, enhancing the FIQA's relevance to practical applications.
- We propose a novel label-free prediction model for face image quality, leveraging the robust Huber loss function to train our quality regression network.
- Through extensive experiments, U3FQ demonstrates a significant advancement over current FIQA and IQA [4, 10] methodologies, outstripping them in terms of accuracy and generalization across benchmark datasets.

Our method's effectiveness is underscored by its superior performance in real-world scenarios, marking a substantial leap forward in the domain of biometric technology.

2. Related Works

In the field of face recognition, a surge of innovation has emerged in recent years through the adoption of unsupervised and semi-supervised learning methods. These methods are fundamentally aimed at enhancing the recognizability of face recognition systems. They do so by developing techniques that assess the quality of facial images and ensure that variations within the embeddings contribute to the robustness and reliability of the recognition systems.

Unsupervised learning approaches, which operate without labeled data, have been particularly notable. They harness intrinsic characteristics of the data to uncover patterns and quality metrics that are meaningful. For example, the Embedding Variability Analysis utilized in SER-FIQ and the Similarity Distribution Distancing strategy applied in SDD-FIQA exemplify how unsupervised learning can extract significant insights into the recognizability of facial images without the need for annotated data.

Similarly, semi-supervised learning techniques have proven to be vital. These methods make use of both annotated and unannotated data, providing a synergistic approach that utilizes available labels while also exploiting the abundance of unlabeled data. The Predictive Confidence Estimation approach of PCNet exemplifies the efficacy of semi-supervised methodologies in refining face recognition systems.

A detailed enumeration of these methodologies and their respective evaluation metrics and discussions is provided in Table 1. This table compiles a list of seminal works that have contributed to the advancement of face recognition technology, highlighting the range of strategies and their assessments that have been central to the evolution of this field.

3. Methodology

Our work introduces the Unified Tri-Feature Quality Metric (U3FQ) for Contextual Facial Image Quality Assessment (FIQA), a key advancement for the precision of biometric identification systems. This section outlines our integrated methodology for developing and refining the U3FQ metric, which accounts for match scores, age disparities, and facial expressions. We also discuss the machine learning and deep learning frameworks applied in our analysis.

- In Subsection 3.1, we provide a succinct overview of the theoretical foundations of U3FQ and its significance within the realm of FIQA. - Subsection 3.2 details the operational aspects of our model, including how age-related match score adjustments and expression-based calibrations are integrated using the AgeDB [11] dataset. - The design of our computational framework, which utilizes both Random Forest and ResNet models for quality score prediction, is presented in Subsection 3.3. - Finally, Subsection 3.4

Table 1. Overview of Recent Works Utilizing Face Recognition Models for Learning Methods

Reference Work	Methodology	Learning Paradigm	Evaluation Metrics and Contribution
SER-FIQ [15]	Embedding Variability Analysis	Unsupervised	Analyzes the robustness of face recognition systems through FNMR and EER, highlighting the variability of face embeddings as a quality measure.
SDD-FIQA [12]	Similarity Distribution Distancing	Unsupervised	Employs a novel metric of similarity distribution distance to assess the performance of recognition models, with a focus on the FMR.
CR-FIQA [3]	Relative Classifiability	Unsupervised	Introduces a method to estimate face image quality by learning sample relative classifiability, validated through FNMR and TAR plots.
Mag-Face [9]	ArcFace Enhancement	Unsupervised	Enhances the ArcFace model by incorporating a magnitude component, with a comprehensive analysis using FNMR plots.
FaceQnet [7]	Neural Network Quality Assessment	Supervised	Utilizes a CNN to predict face image quality, examining the correlation between predicted quality scores and actual recognition system performance.
Automatic FIQ [2]	CNN Approaches	Semi-supervised	Implements a CNN for facial image quality prediction, comparing FNMR and FMR to evaluate performance.
PCNet [16]	Predictive Confidence Estimation	Semi-supervised	Assesses the predictive confidence of face recognition models using a network designed for semi-supervised learning, focusing on FMR.
Optimization Based [6]	Quality Label Supervision	Unsupervised	Studies the effect of supervised quality labels on unsupervised learning models, analyzing FMR and AUC as key metrics.
FaceQAN [1]	Adversarial Noise Profiling	Unsupervised	Explores adversarial noise patterns to evaluate face recognition quality, using FMR and AUC curves for performance assessment.

offers an examination of various computational techniques for match score determination and addresses the handling of non-mated pairs, informed by data from the NIST MEDII dataset to enhance recognition accuracy.

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 ??. 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 facial feature changes 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.

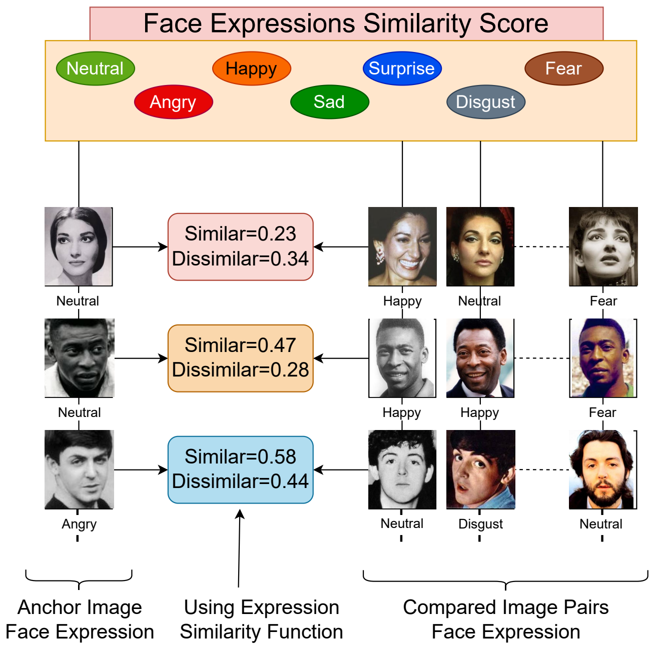

Figure 2. The effectiveness of face matching systems is markedly influenced by the age disparity between compared images. The triplet representation highlights the similarity distance, the age of the compared image, and the age difference relative to the anchor image, with Image 6 serving as the reference point.

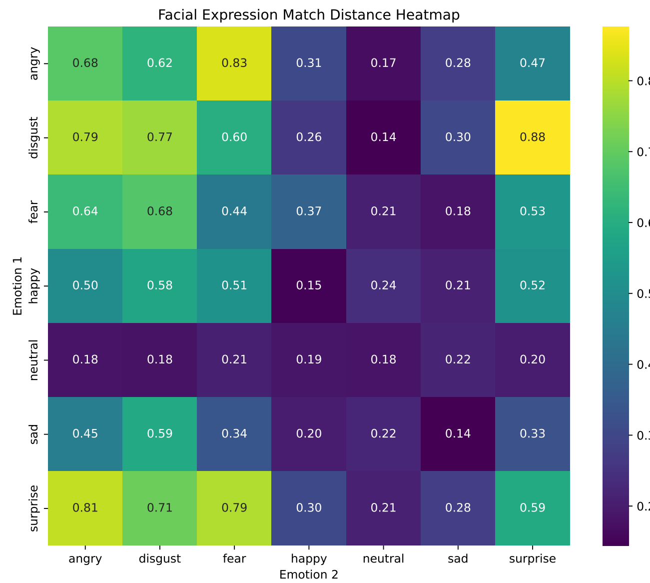

Figure 3. The differential impact of facial expressions on the match score, with weak emotions having a constant effect and strong emotions modifying the score proportionally to their intensity.

To empirically underpin this approach, we present Detection Error Tradeoff (DET) plots that demonstrate the variance in performance with different age groups for all four models: VGG-FACE, OpenFace, ArcFace, and FaceNet. Due to page limitations, these plots are included in the supplementary material. Here, we have added the DET plots from VGG-Face in Figure 4, which show the False Non-Match Rate (FNMR) for different age groups. These plots highlight the impact of age difference on the efficacy of face-matching systems, for different anchor age where 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 significantly influences recognition performance, as variations in expressions can distort critical facial features used in establishing a match. Consequently, this also affects the overall quality of recognition. Figure 3 illustrates the impact that discrepancies in facial expressions have on matching performance, as evidenced by the average match scores across diverse expression pairs.

3.2. Formulations and Optimization

Building on the observations from empirical evidence, we proceed to formulate the mathematical model that incorporates age difference penalties into the facial match score. The age difference penalty function is adapted as follows:

$$f(d,a) = \begin{cases} e^{\alpha(d+\beta)} & \text{if } d < 0 \text{ and } a \leq 30, \\ e^{\alpha(d+\beta)/\theta} & \text{if } d < 0 \text{ and } a > 30, \\ \gamma \cdot d & \text{if } d \geq 0, \end{cases} \quad (1)$$

where d represents the age difference between the anchor and the comparison image, a denotes the anchor's age, and α, β, and γ are parameters dictating the function's shape. The factor θ serves as a damping parameter that reduces the penalty for older anchors.

Our methodology also accounts for the subtle yet significant influence of facial expressions on the match score. This is achieved through the facial expression impact function $g(e)$, which distinguishes between 'weak' and 'strong' emotions, as detailed below:

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

where c is a constant factor for weak emotions, and d scales the expression score $\text{EXPR\_SCORE}(e)$ for strong emotions.

These formulations, alongside the empirical insights, collectively enhance the fidelity of the FIQA model's predictions. By incorporating the dynamics of human aging and expressions, we ensure that our facial recognition system is not only secure but also user-friendly, accommodating the complexities of human features and behaviors.

The algorithm detailed below outlines the process for computing the contextual quality score and estimating the age for a given input image using a ResNet model. The procedure leverages a feature vector that encompasses age, expression, and congruence score, which are derived from the input image and used to predict the quality score.

3.3. Architecture

The U3FQ algorithm initiates with the computation of the match score distance from an ensemble of models $M = \{M_1, M_2, M_3, M_4\}$. The distance metric, denoted as d, is derived from the pairwise discrepancies in the features extracted by each model for the given image I. The age difference function $f(d, a)$ is applied to adjust d based on the age of the anchor image a.

Simultaneously, the expression impact function $g(e)$ adjusts the congruence score depending on the facial expression e, where c is a constant factor for weak emotions and d is a scaling factor for strong emotions, coupled with the expression score $\text{EXPR\_SCORE}(e)$.

These functions are crucial as they capture the dynamic nature of facial recognition where age and emotional expression significantly impact the quality of facial features captured in an image.

The algorithm then constructs a feature vector F, encompassing the age a, expression e, and the aggregate congruence score U, calculated as a weighted sum of the match score distances adjusted by the age and expression multipliers. This vector is integral in capturing the nuanced elements that contribute to the facial image quality.

$$U \leftarrow \sum_{model \in M} d \cdot f(d, a) \cdot g(e)$$

$$F \leftarrow [a, e, U]$$

A set of stochastic embeddings are generated through the ResNet model RN across m iterations to provide robust estimates of the image quality Q and the subject's age. The embeddings are processed to yield a final quality score, reflecting the stability and robustness of the features in the presence of inherent variabilities in facial images.

This mathematical and algorithmic formulation of the U3FQ model demonstrates a robust mechanism for assessing facial image quality, providing insights into the complex interplay between age, expression, and recognition robustness. The model's efficacy is further corroborated through empirical evaluations, showcasing its potential to enhance the performance of biometric systems significantly.

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

Input: Single input image I, ResNet model RN, age a, expression e, match score distance models $M = \{M_1, M_2, M_3, M_4\}$

Output: U3FQ Score or Quality Score and Estimated Age of input Image.

$U \leftarrow 0$ ▷ Initialize U3FQ score
$A \leftarrow []$ ▷ Initialize list for age multipliers
$E \leftarrow []$ ▷ Initialize list for expression multipliers
for all $model \in M$ **do**
  $d \leftarrow \text{ComputeMatchScoreDistance}(I, model)$
  $\mathcal{A} \leftarrow f(d, a)$ ▷ Using the age difference function
  $A \leftarrow A + [\mathcal{A}]$
  if $\text{ExpressionsAreSame}(I, model)$ **then**
    $\mathcal{E} \leftarrow c$ ▷ Constant for same expression
  else
    $\mathcal{E} \leftarrow d * \text{EXPR\_SCORE}(e)$ ▷ Scaled for different expressions
  end if
  $E \leftarrow E + [\mathcal{E}]$
  $U \leftarrow U + d * \mathcal{A} * \mathcal{E}$
end for
$F \leftarrow [a, e, U]$ ▷ Feature vector containing age, expression, and U3FQ score
procedure U3FQ_ASSESSMENT(I, F, RN, $m = 100$)
  $QualityScores \leftarrow []$ ▷ Initialize list for quality predictions
  $Age \leftarrow []$ ▷ Initialize list for age predictions
  for $i \leftarrow 1$ to m **do**
    $quality, ageEstimate \leftarrow RN.\text{Predict}(I, F)$ ▷ Predict quality and age
    $QualityScores \leftarrow QualityScores + [quality]$
    $Age \leftarrow Age + [age]$
  end for
  $finalQuality \leftarrow \text{Average}(QualityScores)$
  $finalAge \leftarrow \text{Average}(AgeEstimates)$
  return $finalQuality, finalAge$
end procedure

3.4. Analytical Study

This subsection embarks on a discourse surrounding alternative methodologies that hold potential for future exploration. Considerations include the grouping of facial images by expression, the differential impacts of varying age differences, and other conceivable permutations that could influence the FIQA landscape.

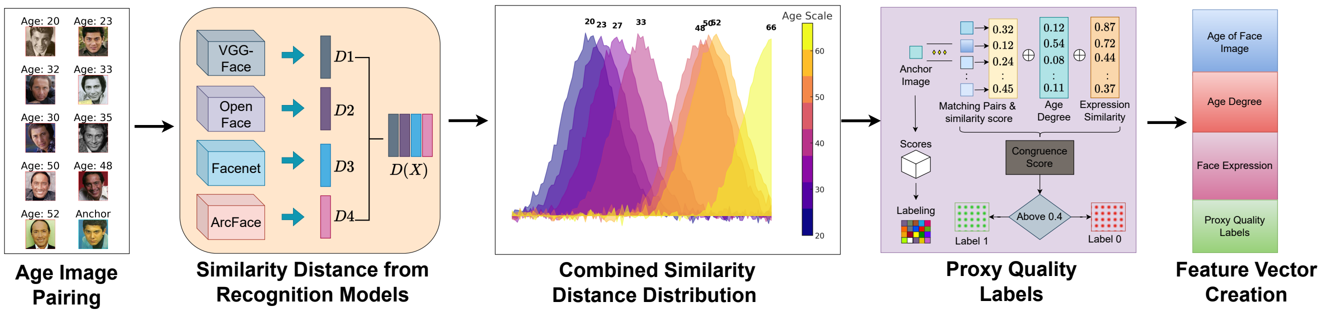

Figure 4. Generation of proxy ground truth quality label with face images adn their age difference for advancing FIQA.

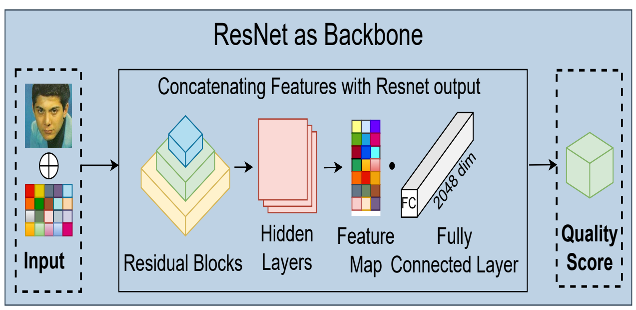

Figure 5. ResNet model as backbone for predicting Quality Score.

4. Experiments

4.1. Datasets

In our study, we utilize the AgeDB dataset [11], which comprises approximately 16,488 images representing a wide spectrum of age variations across different identities. This dataset serves as the cornerstone for our evaluations. For comprehensive assessment and benchmarking, we employ several esteemed datasets: LFW [8], ColorFeret [13], and Adience [5]. These datasets provide a diverse range of facial imagery that enables us to rigorously compare and analyze the performance of our proposed methodologies. Results are meticulously evaluated, showcasing the robustness of our approach in handling age-variant facial recognition tasks.

5. Results

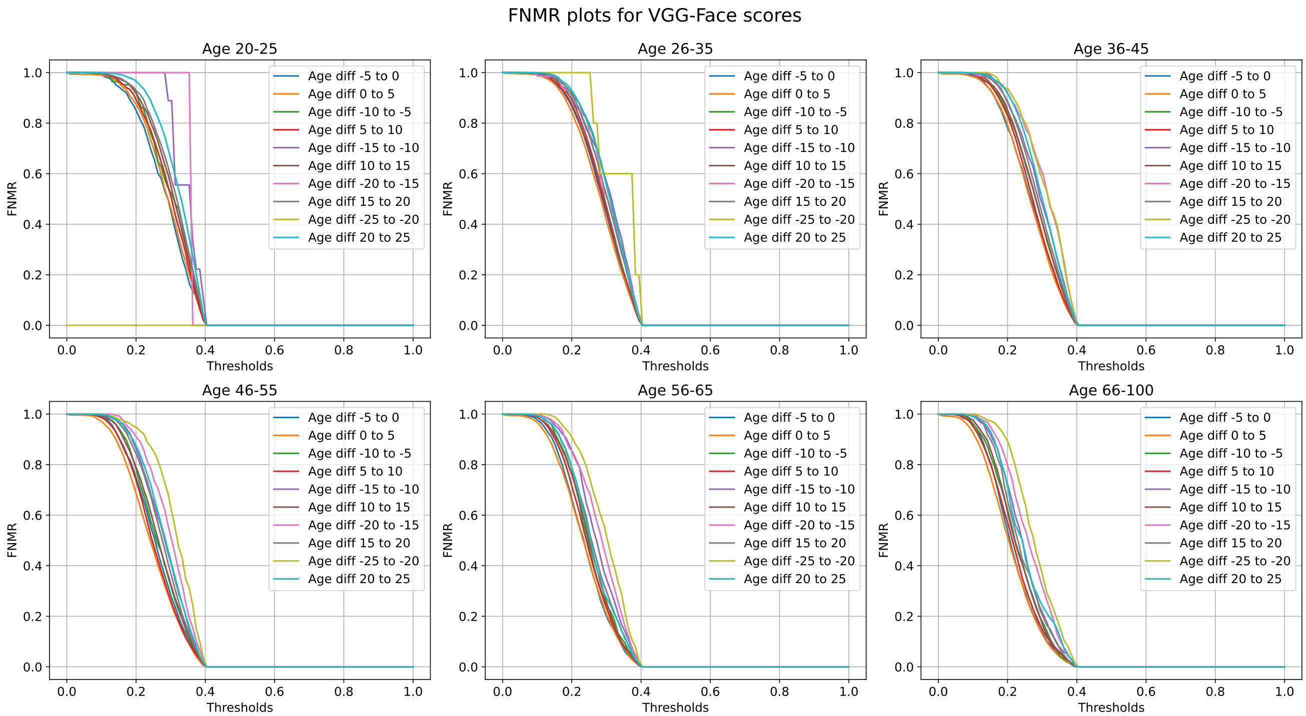

Figure 6. VGG-Face DET plots demonstrating FNMR across different age groups and age difference categories.

6. Conclusion

References

[1] Žiga Babnik, Peter Peer, and Vitomir Štruc. Faceqan: Face image quality assessment through adversarial noise exploration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 748–754. IEEE, 2022. 3

[2] Lacey Best-Rowden and Anil K Jain. Automatic face image quality prediction. *arXiv preprint arXiv:1706.09887*, 2017. 3

[3] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. Cr-fiqa: face image quality assessment by learning sample relative classifiability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5836–5845, 2023. 3

[4] Ruby W Chan and Peter B Goldsmith. A psychovisually-based image quality evaluator for jpeg images. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, volume 2, pages 1541–1546. IEEE, 2000. 2

[5] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014. 6

[6] IGD Fraunhofer. Optimization-based improvement of face image quality assessment techniques. 3

[7] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In

2019 International Conference on Biometrics (ICB), pages 1–8. IEEE, 2019. 3

[8] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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*, 2008. 6

[9] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021. 3

[10] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 2

[11] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 2, 6

[12] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. Sdd-fiqa: unsupervised face image quality assessment with similarity distribution distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7670–7679, 2021. 3

[13] P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000. 6

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

[15] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5651–5660, 2020. 3

[16] Weidi Xie, Jeffrey Byrne, and Andrew Zisserman. Inducing predictive uncertainty estimation for face recognition. *arXiv preprint arXiv:2009.00603*, 2020. 3