Optimizing Image Quality Assessment metrics for subjective perception controlling demographic bias

1st Nuno Gonçalves
Institute of Systems and Robotics
University of Coimbra
Coimbra, Portugal

***Abstract*—This paper examines the correlation between 41 objective Image Quality Assessment (IQA) metrics and Mean Opinion Scores (MOS) across 102 facial images, each modified using four steganography methods at nine encoding levels. As IQA metrics are widely used to approximate human perception, understanding their reliability in the context of facial image distortion is crucial for applications in biometrics, forensics, and secure communication. Using Pearson's and Spearman's correlation coefficients, we rank IQA metrics and evaluate fusion approaches, including PCA, regression, and machine learning. Our analysis also explores demographic (age, gender, ethnicity) and non-demographic (attractiveness) biases in MOS, highlighting significant perceptual variations across observer and subject groups.**

***Index Terms*—FIQA, IQA, MOS, steganography, demographic bias, non-demographic bias, fusion techniques.**

I. Introduction

The evaluation of image quality is essential in applications such as biometric authentication, multimedia processing, and medical imaging [1], [2]. In the specific domain of facial image quality assessment (FIQA), the goal is to ensure that images used for face recognition (FR) systems meet a quality standard that optimizes recognition performance [3]. Unlike traditional image quality assessment (IQA), which considers general visual attributes such as contrast, sharpness, and noise levels, FIQA focuses on assessing image quality in a manner that directly impacts face recognition accuracy [4].

A fundamental challenge in FIQA lies in the discrepancy between objective image quality metrics and human perception, typically quantified through mean opinion scores (MOS). MOS is an aggregate measure obtained from subjective evaluations by human observers [5]. While existing IQA metrics such as peak signal-to-noise ratio [6] (PSNR), structural similarity index measure [7] (SSIM), and visual information fidelity [8] (VIF) provide automated assessments of image quality, their correlation with MOS remains inconsistent across datasets [9]. This misalignment is particularly problematic in facial image analysis, where the perceptual quality of an image may be influenced by both intrinsic image distortions and extrinsic observer biases.

A significant body of research has identified demographic and non-demographic biases in FIQA, where factors such as ethnicity, age, and gender influence the subjective perception of image quality [3], [10]–[12]. These biases arise from the overrepresentation of specific demographic groups in training datasets, as well as perceptual differences across observers. Prior studies have shown that FR accuracy is lower for dark-skinned individuals due to dataset imbalances, while female faces tend to receive lower quality scores in FIQA evaluations [2]. The existence of these biases underscores the need for a more robust and inclusive approach to IQA that accounts for variations in human perception.

The International Civil Aviation Organization [13] (ICAO) and the ISO/IEC 19794–5 standard [14] establish guidelines for image quality in Machine-Readable Travel Documents (MRTDs). These guidelines ensure uniform image conditions (e.g., lighting, focus, and resolution) and consistency across datasets. While these regulations establish a technical baseline, they do not account for perceptual biases and demographic variability in FIQA.

FIQA metrics are optimized for deep-learning-based verification systems rather than regulatory compliance. In ICAO-compliant documents, such as passports and national ID cards, strict quality criteria—including lighting, background uniformity, and sharpness—must be met, irrespective of biometric matching accuracy. This distinction becomes even more critical when facial images are embedded with printer-proof steganographic security features, which may introduce visual patterns that degrade FIQA scores despite maintaining ICAO compliance. These observations suggest that current FIQA alone is insufficient for assessing document image quality, necessitating a more robust approach that integrates multiple IQA methodologies.

The ethical implications of these biases are profound. Political regulations, such as the European Convention on Human Rights (Article 14) [15], the Universal Declaration of Human Rights (Article 7) [16], the General Data Protection Regulation (GDPR, Article 22) [17], and emerging AI governance frameworks, such as the European Artificial Intelligence Act (2024) [18] and proposals in the USA [19], aim to prevent discriminatory decisions. Despite these efforts, biases persist, often introduced through the human observers who evaluate facial images for FIQA algorithms.

Research in neuroscience and psychology suggests that this phenomenon is rooted in the unique processing of facial images by the human brain. The fusiform face area (FFA), a specialized brain region, is particularly tuned to recognizing and evaluating faces [20], [21]. This specialization makes facial image quality assessment a complex task, influenced

by both the demographic and non-demographic characteristics of the faces (e.g., age, gender, ethnicity, attractiveness) and the demographics of the observers (e.g., age, gender, ethnicity).

To address these challenges, this study explores a fusion-based approach to improve the correlation between IQA metrics and MOS. By integrating multiple IQA metrics, the goal is to create an optimized metric that more accurately reflects subjective human assessments. The primary contributions of this work are threefold. First, we rank individual IQA metrics based on their correlation with MOS using Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SRCC). Second, we compare different fusion techniques, including principal component analysis (PCA), regression models, and machine learning-based approaches, to determine the most effective method for enhancing MOS predictability. Third, we analyze the impact of observer demographics on MOS ratings to evaluate the extent of bias in FIQA methodologies.

This study provides a comprehensive evaluation of 41 IQA metrics, highlighting their strengths and weaknesses in aligning with human perception. Our findings demonstrate that fusion-based IQA models significantly improve MOS predictability, with random forest-based fusion outperforming linear and PCA-based methods. The results emphasize the necessity of incorporating perceptual biases into IQA frameworks to develop more accurate and equitable image quality assessment methodologies.

II. Related Work

Significant research efforts have been devoted to improving the accuracy and fairness of facial image quality assessment (FIQA). The limitations of traditional image quality assessment (IQA) techniques in predicting perceptual quality have led to the development of specialized FIQA metrics tailored to face recognition applications [4], [9]. These methods attempt to bridge the gap between objective quality measures and human perception by integrating statistical and perceptual quality indicators.

Recent studies have demonstrated that many FIQA methods exhibit demographic biases, where variations in ethnicity, age, and gender impact quality predictions. Cavazos et al. [3] analyzed multiple face recognition systems and found that lower recognition accuracy is often associated with underrepresented demographic groups, highlighting the limitations of IQA models trained on imbalanced datasets. Similarly, Kabbani et al. [10] observed that FIQA scores tend to favor certain demographic profiles, resulting in biased quality evaluations that could affect real-world biometric verification.

In addition to demographic biases, prior work has investigated the correlation between IQA metrics and mean opinion scores (MOS) obtained from human evaluators. Studies such as those by Huang et al. [2] and Terhoerst et al. [4] have shown that traditional IQA metrics, including peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM), fail to capture perceptual distortions effectively. Consequently, novel deep-learning-based approaches, such as learned perceptual image patch similarity (LPIPS) and deep image structure and texture similarity (DISTS), have been proposed to improve alignment with subjective evaluations [9].

Fusion-based IQA models have emerged as a promising solution to enhance MOS predictability. Recent work by Robinson et al. [22] introduced a multi-metric fusion technique combining statistical, structural, and perceptual indicators to improve quality assessment consistency. Similarly, Kortylewski et al. [23] proposed the use of synthetic data augmentation to mitigate biases in IQA model training. These studies suggest that integrating multiple quality metrics through machine learning or statistical fusion can lead to more robust assessments that align better with human perception.

Despite these advancements, there remains a need for a systematic evaluation of IQA fusion models to determine their efficacy in reducing bias and improving MOS correlation. This work builds upon prior research by conducting a comprehensive comparison of fusion-based IQA approaches, assessing their performance using Pearson's linear correlation coefficient (PLCC) and Spearman's rank-order correlation coefficient (SRCC). By analyzing the effectiveness of different fusion strategies, we aim to contribute towards the development of fairer and more accurate FIQA methodologies.

III. Methodology

A. Dataset

For practical reasons we decided to use the publically available Face Research Lab London (FRLL) Set [24], an ICAO compliant dataset that consists of 102 neutral frontal facial images alongside metadata on attractiveness provided by over 2500 observers. We carefully selected 15 demographically diverse facial images out of the dataset, shown in Fig. 1. Each image was subsequently processed using four printer-proof steganography methods, each applied at nine encoding threshold levels. This approach introduced a diverse range of quality distortions, yielding a total of 555 images, including their original references.

Fig. 1. Sample images from the FRLL dataset.

The used printer-proof steganography methods are based on Generative Adversarial Networks (GANs) [25] to encode and

decode information, we can obtain various results depending on the method used:

- StegaStamp [26]: claims to be the first steganography model capable of decoding data from printed images. The authors show robust results in decoding data under physical transmission by developing novel strategies to add noise in the training process, printer noise simulation, and distortion for the training dataset.
- CodeFace [27]: encoder and decoder networks are trained using end-to-end GANs. It introduces a new security system for encoding and decoding facial images that are printed in common IDs and MRTDs.
- RiemStega [28]: proposes a new loss function that extends the loss function based on the L_2 distance between images to the Riemannian manifold of symmetric and positive definite matrices.
- StampOne [29]: focuses on high-level robust steganography, such as [26], [27], striking a balance between high-quality encoded images and decoding accuracy. It mitigates distortion-related issues like JPEG compression, camera sensors and printer's Gaussian noise by incorporating gradient transform, wavelet transform, and Depthwise [30] to normalize and balance frequencies of the inputs.

Each image in our dataset was evaluated approximately 30 times by human observers, providing a robust Mean Opinion Score (MOS) dataset.

We followed ITU-R BT.500 — 15 [5] recommendation, and adopted the Single Stimulus method. Around 200 different observers were carefully instructed on how to perform the test session, the average duration of the session was 22 minutes, and the average number of tests in each session was 70. Resulting in over 14,000 images being evaluated.

To conduct the sessions we created a webapp, seen in Fig. 2, where the observers conceded their demographic data (age, gender, ethnicity, etc.) and were asked to evaluate each image on a scale from 1 to 100 using a slider bar, for as long as they wanted. The rating scale was divided into five categorical levels: scores from 1 to 25 were classified as Bad, 26 to 50 as Poor, 51 to 75 as Fair, 76 to 99 as Good, and a score of 100 as Excellent.

We collected observer's demographic (age, gender, ethnicity, and country of origin) and non-demographic information (education level, device being used, and place where test was being performed).

B. Statistical Bias Analysis

Considering that the homogeneity of variances assumption is not met, we performed Welch's analysis of variance (ANOVA) to examine the impact of observer demographics — gender (140 male, 60 female), ethnicity (173 causasian, 8 latino, 8 black, and others), and age grouped into categories (under 18, 18 — 25, 26 — 40, 41 — 60, over 60) as well as subject demographic and non-demographic (attractiveness) characteristics on MOS. The statistical significance of each factor was evaluated using a threshold of $p < 0.01$, and effect size analysis was conducted using Eta Squared [31] (η^2) to quantify the relative contribution of each variable.

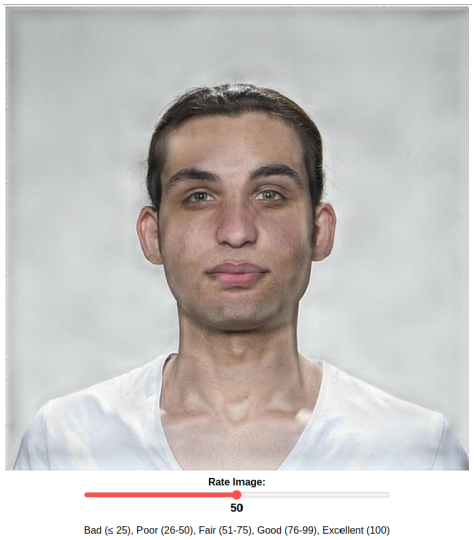

Fig. 2. Custom webapp platform used to access overall image quality.

1) Impact of Observer Demographics on MOS: According to table I, ethnicity, gender and age significantly impact MOS scores, suggesting that subjective evaluations of image quality are influenced by observer characteristics. Post-hoc Tukey's HSD [32] tests were conducted to determine specific group differences, identifying notable discrepancies across demographic subgroups. The presence of these biases underscores the need for IQA methodologies that incorporate perceptual fairness considerations.

TABLE I
WELCH'S ANOVA RESULTS FOR OBSERVER DEMOGRAPHICS.

Factor	Welch F	DF	p-value	Eta Squared
Observer Gender	6.97	1	0.00112	0.012
Observer Ethnicity	9.36	8	5.65e-13	0.027
Observer Age Group	33.00	4	$\cong 0$	0.045

2) Impact of Image Characteristics on MOS: Table II presents the influence of subject demographic and non-demographic characteristics on MOS. Results suggest that MOS is influenced by aesthetic preferences and other visual factors, reinforcing its inherently subjective nature.

TABLE II
WELCH'S ANOVA RESULTS FOR SUBJECT CHARACTERISTICS.

Factor	Welch F	DF	p-value	Eta Squared
Subject Gender	19.84	1	5.87e-06	0.018
Subject Ethnicity	8.85	3	7.38e-06	0.022
Subject Age Group	11.62	4	4.82e-35	0.033
Attractiveness	18.04	1	2.18e-05	0.015

3) Observer-Subject Interaction Effects on MOS: A two-way ANOVA was performed to analyze the interaction effects between observer demographics and subject characteristics.

TABLE III
TWO-WAY ANOVA INTERACTION EFFECTS.

Observer × Subject	F-value	DF	p-value	Eta Squared
Gender	2.31	1	0.0654	0.006
Ethnicity	1.51	24	0.0556	0.011
Age Group	2.87	16	1.94e-10	0.024

These findings confirm the existence of demographic bias in MOS ratings, making it crucial to account for such biases when using subjective scores in image quality assessment models. The effect size analysis using Eta Squared (η^2) provides insight into the relative contribution of each factor to MOS variance. Our results reinforce the need for better FIQA models to correct demographic biases.

C. Image Quality Metrics

A total of 41 different Image Quality Assessment (IQA) metrics were evaluated, spanning traditional signal-based measures (e.g., PSNR, SSIM), perceptual-based metrics (e.g., FSIM, VIF), and deep-learning-based approaches (e.g., LPIPS, DISTS). The correlation between these metrics and MOS was analyzed using Pearson's and Spearman's correlation coefficients.

Fig. 3 presents scatter plots comparing MOS values against individual IQA metrics. Each plot illustrates the relationship between subjective evaluations and objective IQA scores, highlighting the varying degrees of correlation among metrics. Some metrics, such as FSIM and MS-SSIM [33], demonstrate a strong monotonic relationship with MOS. These visualizations justify the need for a fusion-based approach to optimize MOS predictability.

D. Fusion-Based Image Quality Assessment

To improve the correlation between IQA metrics and MOS, we implemented a fusion-based approach integrating multiple quality measures. The fusion process involved weighting individual IQA metrics based on their predictive performance relative to MOS. We explored several fusion techniques, including:

- Principal Component Analysis (PCA): Used to reduce redundancy among IQA metrics and derive an optimized linear combination.
- Regression-Based Models: Linear regression and ridge regression were applied to determine the best combination of metrics that predict MOS effectively.
- Machine Learning Approaches: A Random Forest [34] (RF) model was trained using individual IQA metrics as input features, leveraging nonlinear interactions to improve MOS correlation. The dataset was split into 80% for training and 20% for testing. The model was configured with 100 trees (*n_estimators = 100*) and used the MSE criterion for optimization.

To assess the effectiveness of individual and fusion-based IQA metrics, we employed two correlation measures:

- Pearson's Linear Correlation Coefficient (PLCC): Evaluates the linear relationship between IQA scores and MOS.

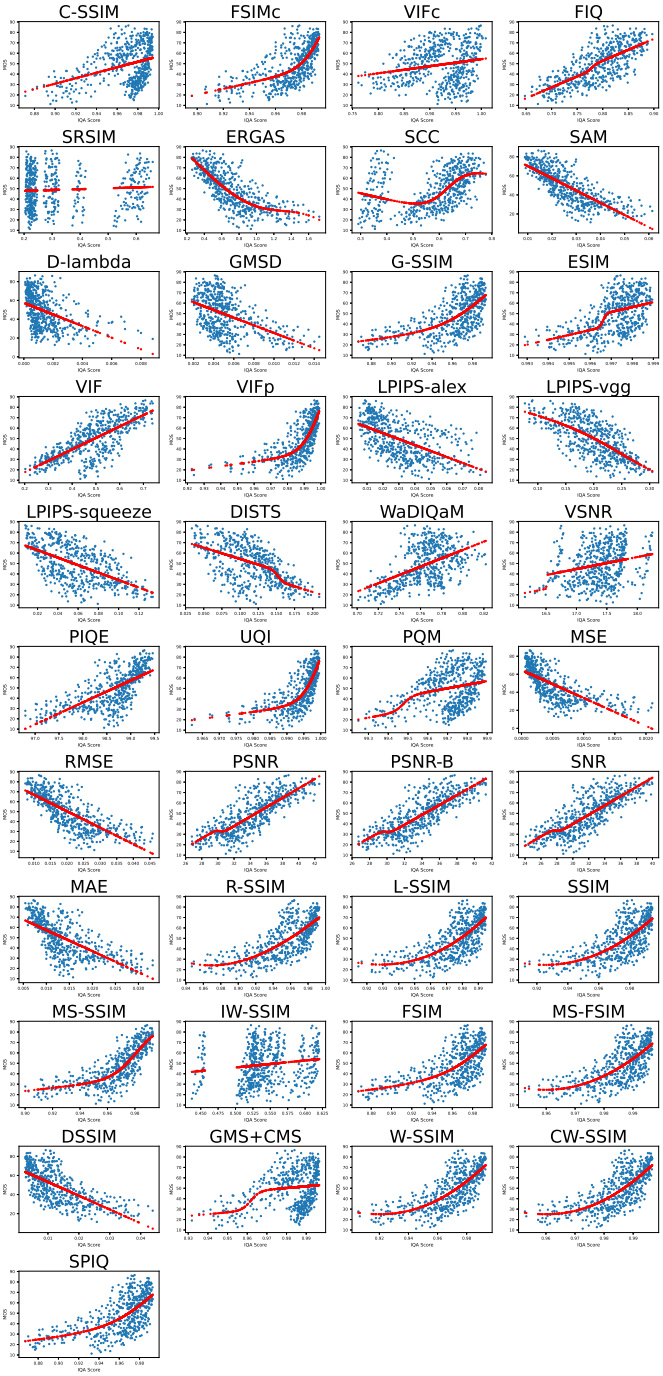

Fig. 3. Scatter plots illustrating the relationship between Mean Opinion Scores (MOS) and individual Image Quality Assessment (IQA) metrics.

- Spearman's Rank-Order Correlation Coefficient (SRCC): Measures the monotonic relationship between IQA scores and MOS, accounting for nonlinearity in observer perception.

E. Implementation Details

The experiments were conducted using Python. The dataset went through a post-screening process to normalize IQA

scores and eliminate outliers. Training and evaluation followed a five-fold cross-validation strategy to ensure robust model performance.

IV. Results

A. Ranking of Individual IQA Metrics

The correlation analysis between individual IQA metrics and MOS revealed significant variability in their predictive performance. Fig. 4 presents a ranked comparison of the 41 evaluated IQA metrics based on the arithmetic mean of the Pearson and Spearman correlation coefficients. Metrics such as MS-SSIM, SAM, and PSNR-B [35] demonstrated the highest correlation with MOS, suggesting that these metrics consistently align with subjective perception of distortions in facial images.

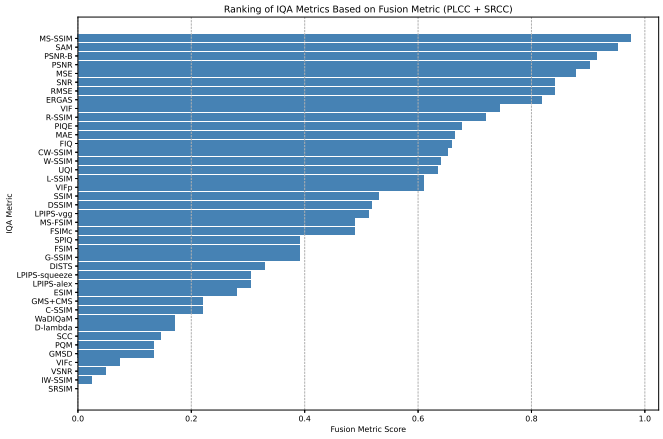

Fig. 4. Ranking of IQA Metrics Based on Fusion Metric (PLCC & SRCC).

B. Comparison of Fusion Models

To evaluate the effectiveness of different fusion techniques, we compared multiple approaches, including PCA-based fusion, ridge regression, and random forest-based models. Fig. 5 summarizes the performance of these models in terms of PLCC and SRCC, indicating that the random forest fusion method yielded the highest correlation with MOS.

The results indicate that machine learning-based fusion techniques outperform statistical fusion approaches such as PCA and regression. The random forest fusion model effectively captures nonlinear interactions between IQA metrics, enabling more accurate predictions of subjective image quality. PCA, while useful for dimensionality reduction, discards some potentially valuable information, leading to lower correlation coefficients. Regression-based models, constrained by their linear nature, also fall short in capturing complex dependencies among IQA metrics.

C. Correlation Between Fusion Scores and MOS

The application of fusion models significantly improved the correlation between IQA scores and MOS. Fig. 6 illustrates the relationship between the optimized fusion scores and MOS

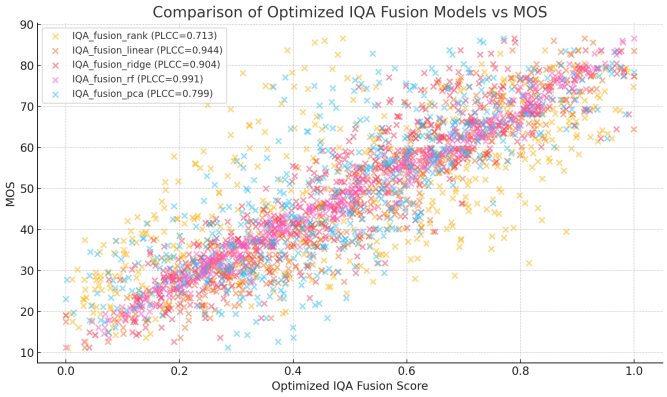

Fig. 5. Scatter plot of Fusion-Based IQA Scores vs. MOS.

values, demonstrating a substantial improvement in correlation coefficients. While deep learning models could further enhance performance, the Open Source Face Image Quality (OFIQ) framework prioritizes simple and fast implementations, making it the most suitable choice for this study.

These results underscore the advantage of integrating multiple IQA metrics to obtain a more reliable predictor of MOS. The success of the fusion model suggests that no single IQA metric is sufficiently comprehensive to capture the full range of perceptual quality factors, but rather, a combination of complementary metrics provides the most robust estimation.

Overall, the findings highlight the superiority of machine learning-based fusion strategies in FIQA. The improved MOS predictability achieved through fusion further supports the hypothesis that integrating multiple complementary metrics enhances the robustness of IQA systems. Additionally, these results provide insights into the design of more accurate quality assessment models that align better with human perception.

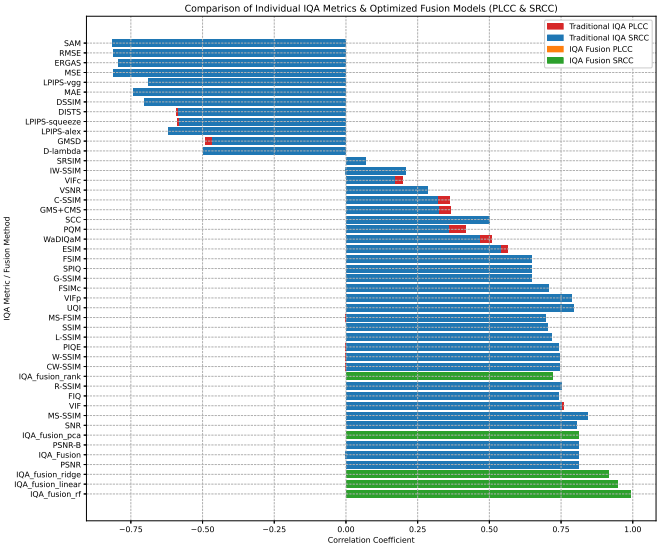

Fig. 6. Comparison of IQA metrics and Optimized Fusion Models (PLCC & SRCC)

V. Conclusion and Future Work

Beyond the evaluation of IQA metrics and fusion methods, this study analyzed the extent of observer demographic biases in MOS ratings. The ANOVA results presented in Section III indicate that observer characteristics, including gender, ethnicity, and country of origin, significantly influenced MOS. These findings suggest that FIQA methodologies should account for potential biases introduced by subjective assessments from diverse observer groups.

A promising direction for future research is the development of a fully automated, no-reference ICAO-compliant image quality assessment model. Our current approach relies on a fusion of full-reference IQA metrics, which requires access to pristine reference images for quality comparison. However, in real-world applications, such as passport verification and ID issuance, reference images are often unavailable. To address this limitation, we propose training a deep neural network (DNN) on the fusion metric, enabling the transition from a reference-based to a no-reference image quality model.

Developing a deep learning-based, no-reference ICAO-compliant IQA model represents a crucial step toward improving image quality evaluation in biometric and security applications. This approach would ensure robust, unbiased, and practical quality assessments for automated identity verification systems.

References

[1] H.-I. Kim, S. H. Lee, and M. R. Yong, "Face image assessment learned with objective and relative face image qualities for improved face recognition," *IEEE International Conference on Image Processing*, pp. 4027–4031, 2015.

[2] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2781–2794, 2020.

[3] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 101–111, 2021.

[4] P. Terhoerst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Face quality estimation and its correlation to demographic and non-demographic bias in face recognition," *IEEE International Joint Conference on Biometrics*, pp. 1–11, 2020.

[5] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," Recommendation BT.500-15, International Telecommunication Union, Radiocommunication Sector, 2023. Available at: urlhttps://www.itu.int/rec/R-REC-BT.500-15.

[6] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2nd ed., 2002.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[9] Z. Babnik and V. Struc, "Assessing bias in face image quality assessment," *ArXiv preprint*, 2022.

[10] W. Kabbani, K. Raja, R. Ramachandra, and C. Busch, "Demographic variability in face image quality measures," *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, 2024.

[11] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Face quality estimation and its correlation to demographic and non-demographic bias in face recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11, 2020.

[12] J. E. Tapia *et al.*, "Beauty score fusion for morphing attack detection," in *Proceedings of the British Machine Vision Conference (BMVC) Workshops*, 2024.

[13] "Doc 9303 machine readable travel documents, part 3: Specifications common to all mrtds," September 2015.

[14] "Information technology – biometric sample quality – part 5: Face image data," April 2010.

[15] C. of Europe, "European convention on human rights, article 14: Prohibition of discrimination," 1950. Council of Europe Treaty Series No. 5.

[16] U. Nations, "Universal declaration of human rights, article 7: Equality before the law," 1948. General Assembly Resolution 217 A.

[17] E. Union, "General data protection regulation (eu) 2016/679, article 22: Automated individual decision-making, including profiling," 2016. Official Journal of the European Union, L 119.

[18] E. Union, "Artificial intelligence act (ai act) - regulation (eu) 2024/1689," 2024. Official Journal of the European Union, L 1689, 12 July 2024.

[19] T. W. House, "Blueprint for an ai bill of rights: Making automated systems work for the american people," 2022. Office of Science and Technology Policy, United States.

[20] N. Kanwisher and G. Yovel, "The fusiform face area: a cortical region specialized for the perception of faces," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 361, no. 1476, pp. 2109–2128, 2006.

[21] D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 411–437, 2008.

[22] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?," *ArXiv preprint*, 2020.

[23] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[24] L. DeBruine and B. Jones, "Face research lab london set," *Open Science Framework*, 2017.

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018.

[26] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[27] F. Shadmand, I. Medvedev, and N. Goncalves, "Codeface: A deep learning printer-proof steganography for face portraits," *IEEE Access*, vol. 9, pp. 167282–167291, 2021.

[28] A. Cruz, G. Schardong, L. Schirmer, J. Marcos, F. Shadmand, and N. Gonçalves, "Riemstega: Covariance-based loss for print-proof transmission of data in images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2025.

[29] F. Shadmand, I. Medvedev, L. Schirmer, J. Marcos, and N. Gonçalves, "Stampone: Addressing frequency balance issue in printer-proof steganography." ISR Technical Report. To be submited soon., 2022.

[30] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2022.

[31] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York, NY, USA: Routledge, 2nd ed., 1988.

[32] J. W. Tukey, *Comparing Individual Means in the Analysis of Variance*, vol. 5. International Biometric Society, 1949.

[33] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *IEEE Asilomar Conference on Signals, Systems and Computers*, pp. 1398–1402, 2003.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] L. Ma, S. Wang, G. Shi, D. Zhao, and W. Gao, "Psnr-b: A novel peak signal-to-noise ratio based on edge preservation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 537–549, 2012.