Pseudo-MOS Learning: A Hybrid Full-to-No-Reference FIQA Framework

André Neto[1,2][0009−0001−0398−5859] and Nuno Gonçalves[1,2][0000−0002−1018−4945]

[1] University of Coimbra
[2] Institute of Systems and Robotics — Coimbra

Abstract. A persistent discrepancy exists between standard Image Quality Assessment (IQA) metrics and human perceptual judgments, typically quantified through Mean Opinion Scores (MOS). This gap poses a critical challenge for tasks where visual quality directly impacts performance, such as facial recognition and visual data transmission. In particular, assessing the perceptual quality of steganographically distorted facial images remains difficult, especially in the absence of reference images. To address this, we introduce a hybrid Full-to-No-Reference framework for Face Image Quality Assessment (FIQA), built upon a learning strategy based on pseudo-MOS. A full-reference fusion metric is first trained by regressing multiple classical IQA scores against human MOS on a subset of a facial dataset. This metric is then applied to the full dataset to generate pseudo-MOS labels. Using deep features extracted from a ResNet-18 model pretrained on ImageNet, we train a no-reference regressor capable of predicting perceptual quality. The proposed framework bridges full-reference supervision and no-reference inference, offering a scalable and accurate solution for FIQA in challenging conditions and paving the way for application-specific, data-driven IQA designs.

Keywords: Face IQA · Steganography · No-Reference IQA · Pseudo-MOS · Full-Reference Fusion · Deep Regression

1 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 in face recognition systems meet a quality standard that optimizes recognition performance [3]. Unlike general Image Quality Assessment (IQA), which considers visual attributes such as contrast, sharpness, and noise, FIQA focuses on assessing image quality in a manner that directly impacts face recognition accuracy [4].

IQA methods are divided into two types: subjective and objective. Subjective IQA relies on human evaluators who assign quality scores based on perceived visual quality—typically aggregated as Mean Opinion Scores (MOS). While considered the gold standard, subjective evaluation is costly, time-consuming, and

not scalable. In contrast, objective IQA methods use algorithmic models to predict perceptual quality automatically, often by pixel-wise comparison or learned features.

Objective IQA methods can be further categorized as Full-Reference (FR-IQA) or No-Reference (NR-IQA). FR-IQA requires access to an undistorted reference image and evaluates quality by comparing the distorted image to it. Although often accurate, FR-IQA is only applicable when a pristine reference is available. NR-IQA, on the other hand, estimates image quality without any reference, making it more suitable for real-world scenarios. However, its ability to generalize across content and distortion types remains limited.

Facial Image Quality Assessment (FIQA) falls within the broader IQA domain but is uniquely constrained by biometric applications, where reference images are rarely available. As such, most FIQA methods are inherently no-reference, relying on task-specific priors and learned representations to estimate quality.

A fundamental challenge in IQA lies in the discrepancy between objective quality metrics and human perceptual judgments. While classical metrics such as Peak Signal-to-Noise Ratio (PSNR) [5], Structural Similarity Index (SSIM) [6], and Visual Information Fidelity (VIF) [7] provide automated assessments of image quality, their correlation with subjective perception remains inconsistent across datasets [8]. This misalignment is particularly problematic in facial image analysis, where perceptual quality is influenced by both intrinsic distortions and observer-related biases.

A growing body of research highlights the presence of demographic and non-demographic biases in FIQA, whereby factors such as ethnicity, gender, and age influence perceived image quality [3, 9–11]. These biases often stem from dataset imbalances and inter-observer variability. For instance, face recognition accuracy tends to be lower for dark-skinned individuals, and female faces are frequently rated with lower quality scores in FIQA evaluations [2], underscoring the need for inclusive and perceptually aligned quality assessment approaches.

The International Civil Aviation Organization [12] (ICAO) and the ISO/IEC 19794–5 standard [13] 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.

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

Evidence from neuroscience further supports the complexity of facial image perception. The fusiform face area, a specialized region in the human brain, is selectively activated by face stimuli [19, 20]. This biological specialization makes FIQA particularly sensitive to both stimulus features (e.g., age, gender, ethnicity, attractiveness) and the demographic background of the observers.

An even greater challenge arises when evaluating the quality of steganographically distorted facial images. Steganography is the practice of concealing information within digital media, typically by subtly modifying pixel values in a way that is imperceptible to human observers [21]. Although visually unobtrusive, these alterations can degrade biometric features and compromise recognition performance. NR-IQA approaches, while not requiring references, are generally not designed to detect such imperceptible, task-relevant distortions.

To address these limitations, we propose a hybrid FIQA framework based on pseudo-Mean Opinion Scores (pseudo-MOS). Beginning with a small subset of facial images labeled with subjective MOS, we train a regression-based fusion model that integrates multiple classical FR-IQA scores into a single, perceptually aligned quality measure. This FR fusion model is then applied to a larger dataset to generate pseudo-MOS labels for images lacking ground-truth annotations. A NR regression model is subsequently trained on deep features extracted from a ResNet-18 pretrained on ImageNet, enabling perceptual quality prediction without the need for reference images.

Our approach bridges the gap between FR supervision and NR inference. It provides a scalable solution for IQA in images subjected to complex, low-visibility distortions such as steganography, and lays the groundwork for adaptable, domain-specific quality assessment models.

2 Related Work

To improve perceptual alignment, several fusion-based IQA methods have been proposed. Liu et al. [22] introduced a multi-method fusion (MMF) framework in which multiple FR-IQA scores are linearly combined through regression to better approximate human judgments. Similarly, Henniger et al. [23] developed a Random Forest model trained on handcrafted image features drawn from ISO face quality standards, improving predictive utility for biometric applications. Fusion models demonstrate that integrating complementary quality cues yields superior MOS correlation compared to standalone metrics [24].

Beyond full-reference settings, the scarcity of ground-truth subjective scores has prompted the development of weakly supervised approaches based on pseudo-labels. Chen et al. [25] proposed generating pseudo-MOS scores by averaging outputs from several FR-IQA metrics. RankIQA [26] employed synthetic degradations and relative ranking supervision to learn ordinal quality relationships. Wu et al. [27] used cascaded CNN regressors trained on pseudo-MOS to bootstrap NR-IQA learning. These approaches confirm the viability of pseudo-supervision in training deep IQA models when human MOS labels are limited or unavailable.

Recent progress in NR-IQA has been driven by the use of deep features from CNNs pretrained on large datasets. Kang et al. [28] showed that CNNs can directly regress quality scores from image patches. In the domain of FIQA, techniques such as SER-FIQ [29] exploit the dropout-based variability of face embeddings as a proxy for image quality, while MagFace [30] learns quality-aware facial representations by correlating embedding norms with recognition utility. QualFace [31] adapts deep face recognition models to conform with ICAO and ISO/IEC document standards, introducing an explicit quality estimation module for ID and travel document verification. These methods mark a shift from handcrafted quality indicators to learned feature-level representations tailored to face recognition tasks.

Other studies emphasize that image quality is inherently task-specific. In FIQA, quality is defined not by visual aesthetics but by its impact on face recognition performance. Standards such as ISO/IEC 19794–5 codify this operational perspective, specifying conditions for acceptable biometric image acquisition. Supervised models like FaceQnet [32] learn to predict face utility scores directly from embeddings. Datasets such as PIPAL [33], which include generative distortions, further highlight the need for context-specific IQA evaluation. Our work follows this trajectory by targeting steganographically degraded facial images, an emerging use case not addressed in current FIQA literature.

3 Methodology

3.1 Dataset

Our dataset is derived from the publically available Face Research Lab London (FRLL) [34] set, comprising 102 frontal ICAO-compliant facial images. Each image was encoded using four printer-proof steganographic methods, described ahead, each applied at nine different intensity levels, yielding a total of 3,672 distorted images.

The dataset was partitioned into four subsets, as follows:

- Labeled Set (15 identities, 540 images): a core set of demographically diverse subjects, shown in Fig. 1, annotated with MOS through human evaluation.
  - Training Subset (12 identities, 432 images): used to train the full-reference fusion model, regressing FR-IQA metrics to human MOS.
  - Disjoint Test Subset (3 identities, 108 images): held out from the framework and are used solely for final results comparisons, simulating out-of-distribution generalization.
- Unlabeled Set (87 identities, 3,132 images): no subjective scores were collected for these images. Instead, pseudo-MOS labels were generated using the FR fusion model, enabling large-scale weak supervision.
- No-Reference (NR) Training Pool (99 identities, 3,564 images): Combines the 12 labeled identities and the 87 pseudo-labeled identities, and is used to train the NR regression model on deep features extracted from distorted images.



Fig. 1: Labeled set selected from the FRLL dataset.

The printer-proof steganography methods used were based on Generative Adversarial Networks (GANs) [35] to encode and decode information, we can obtain various results depending on the method used:

- StegaStamp [36]: 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 [37]: 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 [38]: 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 [39]: focuses on high-level robust steganography, such as [36, 37], 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 [40] to normalize and balance frequencies of the inputs.

Each image in our dataset was evaluated approximately 30 times by human observers, providing a robust MOS dataset.

We followed ITU-R BT.500–15 [41] 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 rounded number of tests in each session was 70. Resulting in over 14,000 evaluations.

To conduct the sessions we created a webapp in Django, seen in Fig. 2, where the observers were asked to evaluate each image on a scale from 1 to 100 using a slider bar, with no time limit. 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.

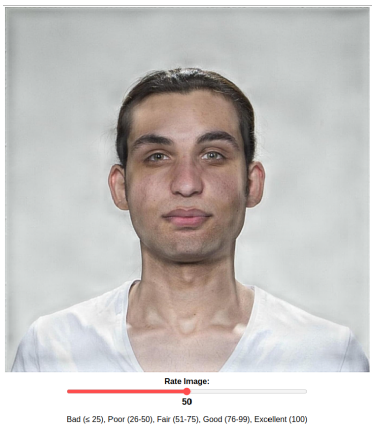

Fig. 2: Snippet of the custom webapp platform used to access overall image quality.

To address this, we propose a two-stage hybrid pipeline that first builds a full-reference (FR) quality estimator supervised by MOS, and then extends this supervision to a no-reference (NR) model via pseudo-MOS learning.

3.2 Full-Reference Analysis and Fusion

We compute 41 FR-IQA scores for each distorted image in the MOS-labeled subset. To identify which metrics align best with human perception, we calculate both the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC) [42] between each metric and the human MOS. This analysis serves to rank the metrics and select the top-k most perceptually relevant.

3.3 Pseudo-MOS Generation via FR-IQA Fusion

The selected top-k IQA metrics are standardized and used as input features for supervised regression models trained to approximate the human MOS. We evaluate Support Vector Regression [43], Random Forest [44], XGBoost [45], and

LightGBM [46] as candidate fusion models. Random Forest consistently achieved the highest PLCC and SRCC on the validation set for $k = 5$, and was selected as the fusion model. This trained regressor is then applied to the unlabeled portion of the dataset (3,132 distorted images), generating pseudo-MOS scores that serve as proxy ground-truth labels.

3.4 NR Regression Model

To enable NR quality prediction, we extract deep features from each distorted image using a ResNet-18 [47] model pretrained on ImageNet. These feature vectors are then used to train a regression model, again using Random Forest, tasked with predicting the pseudo-MOS labels obtained from the FR fusion step. The final model performs quality assessment using only the distorted input image, without requiring access to a reference image, thereby enabling NR-IQA in real-world scenarios.

4 Results

4.1 Correlation of Individual IQA Metrics with Human Perception

The correlation analysis between individual FR-IQA metrics and human-rated MOS revealed substantial variability in performance. As shown in Fig. 3, several metrics exhibit strong linear trends with MOS, while others are poorly aligned or even negatively correlated. To quantify this, we ranked all 41 metrics based on the average of their PLCC and SRCC.

Fig. 4 presents this ranking, highlighting that metrics such as PSNR, PSNR-B [48], and MS-SSIM [49] show the highest alignment with subjective opinion, while others, such as SRSIM [6], diverge significantly.

4.2 Full-Reference Fusion Performance

Using the 432 MOS-labeled images and $k = 5$ best metrics, we evaluated six fusion models.

As summarized in Table 1, the Random Forest regressor achieved the best overall performance across all four evaluation metrics, as seen in Fig. 5. Notably, it reached a PLCC of 0.8582 and SRCC of 0.8637, outperforming both linear and gradient boosting approaches.

The trained Random Forest model was then applied to 3,132 unlabeled images, generating pseudo-MOS scores that serve as soft ground truth for the no-reference model.

4.3 No-Reference Regression Model

We trained a NR regressor using features extracted from a ResNet-18 pretrained on ImageNet. These features, paired with the pseudo-MOS scores generated from the fusion model, served as training data for a second Random Forest model.

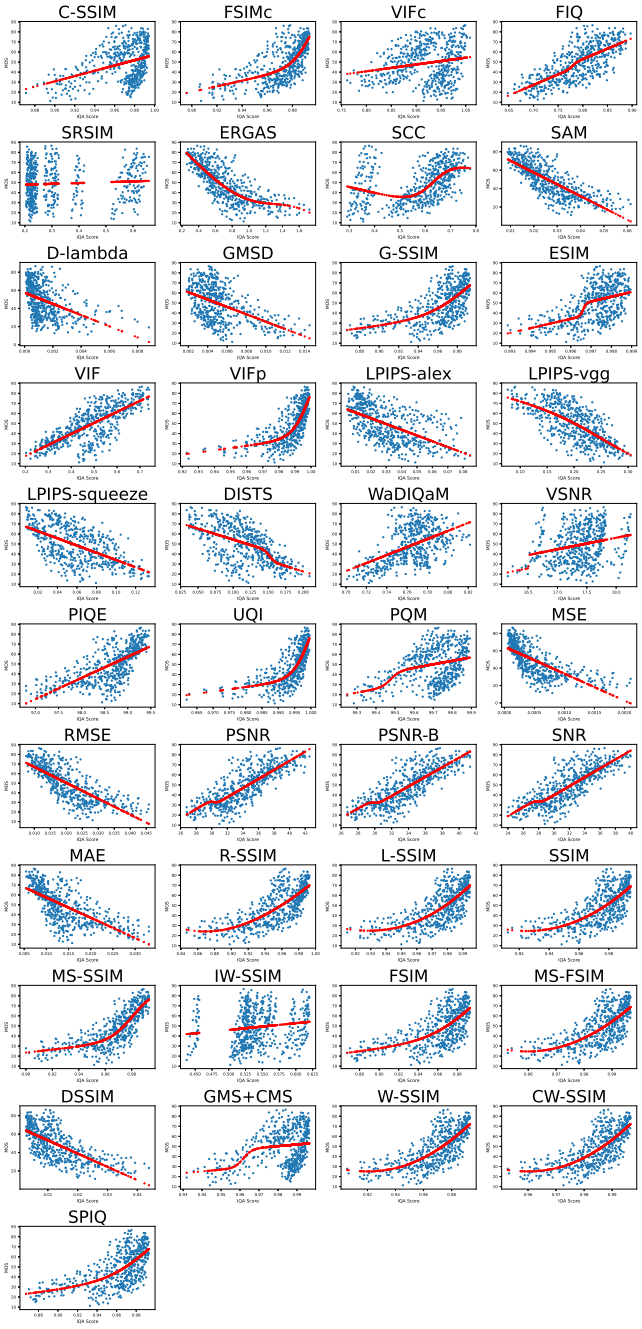

Fig. 3: Scatter plots illustrating the relationship between Mean Opinion Scores (MOS) and individual Full-Reference IQA metrics.

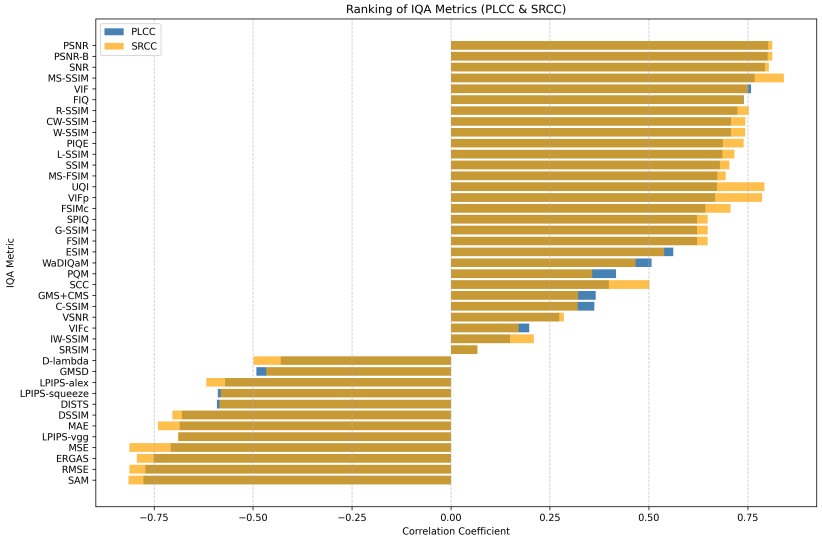

Fig. 4: Ranking of FR-IQA metrics by correlation with MOS (PLCC and SRCC).

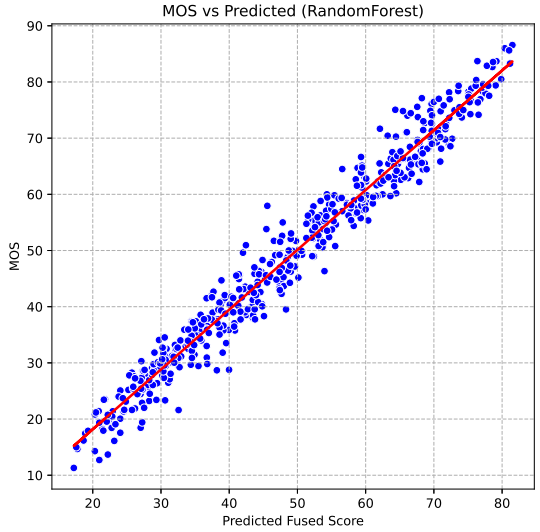

Fig. 5: MOS vs. Predicted fused Random Forest score

Table 1: Fusion performance on 432 labeled images (MOS ground truth).

Model	PLCC	SRCC	MSE	MAE
Linear Regression	0.8123	0.8248	108.83	7.98
Ridge	0.8125	0.8246	108.76	7.96
Random Forest	**0.8582**	**0.8637**	**84.48**	**6.95**
SVR	0.8207	0.8207	108.44	8.08
XGBoost	0.8560	0.8596	108.44	8.08
LightGBM	0.8560	0.8577	85.90	7.15

Fig. 6 shows the resulting alignment between the predicted NR quality scores and the fusion-based pseudo-MOS values. A clear linear trend is observed, with the model capturing both low and high quality regimes effectively.

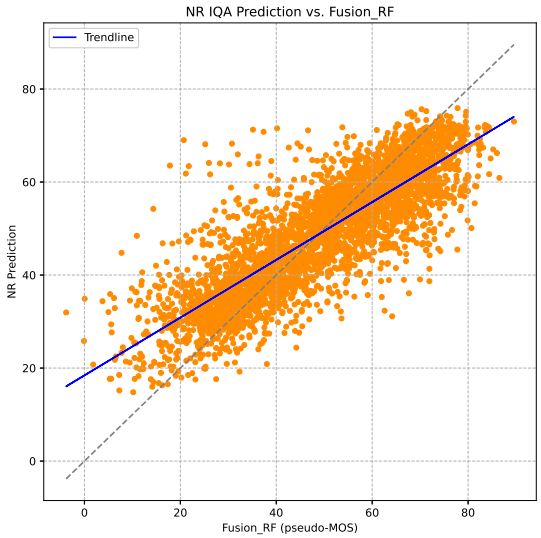

Fig. 6: Scatter plot of NR predictions vs. pseudo-MOS scores. The dashed line indicates identity; the blue line is the linear trend.

Quantitative results are reported in Table 2. Our proposed model, trained using deep features and pseudo-MOS supervision, achieves a PLCC of 0.8361 and SRCC of 0.8382, with an MSE of 90.48 and MAE of 7.16. In comparison, classical NR metrics such as NIQE and PIQE demonstrate weaker correlations with perceptual quality and substantially higher error. Even task-specific approaches like SER-FIQ and MagFace underperform, highlighting the robustness

of our weakly supervised strategy for NR-IQA on steganographically degraded facial images.

Table 2: Performance of our NR IQA model, ResNet18 + Random Forest (RF), compared to standard NR-IQA baselines, using pseudo-MOS as ground truth.

Method	PLCC	SRCC	MSE	MAE
Ours (ResNet18 + RF)	**0.8361**	**0.8382**	**90.48**	**7.16**
NIQE	0.7536	0.7431	6859.81	82.62
PIQE	0.2993	0.3114	3297.65	57.05
SER-FIQ	-0.1648	-0.1542	123.50	9.23
MagFace	-0.6095	-0.6362	4437.91	66.24

5 Conclusion and Future Work

We proposed a hybrid full-to-no-reference framework for FIQA that predicts image quality in the absence of reference images by leveraging a pseudo-MOS supervision strategy. Our method first trains a FR fusion model to regress human perceptual judgments on a labeled subset, generating pseudo-MOS labels for a larger unlabeled dataset. These labels are then used to train a deep NR regressor, enabling quality prediction from distorted images alone. This two-stage pipeline effectively bridges the gap between fully supervised FR-IQA and reference-free NR-IQA approaches.

Beyond the development of our ICAO-compliant NR IQA metric, the proposed framework offers a flexible foundation for constructing a variety of task-specific models. By enabling scalable, perceptually grounded supervision with limited ground-truth annotations, our approach can facilitate quality-aware training in applications such as GANs, forensic imaging, and domain-adapted biometric pipelines.

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. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2nd ed., 2002.
6. 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.
7. 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.
8. Z. Babnik and V. Struc, "Assessing bias in face image quality assessment," *ArXiv preprint*, 2022.
9. 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.
10. 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.
11. J. E. Tapia *et al.*, "Beauty score fusion for morphing attack detection," in *Proceedings of the British Machine Vision Conference (BMVC) Workshops*, 2024.
12. "Doc 9303 machine readable travel documents, part 3: Specifications common to all mrtds," September 2015.
13. "Information technology – biometric sample quality – part 5: Face image data," April 2010.
14. C. of Europe, "European convention on human rights, article 14: Prohibition of discrimination," 1950. Council of Europe Treaty Series No. 5.
15. U. Nations, "Universal declaration of human rights, article 7: Equality before the law," 1948. General Assembly Resolution 217 A.
16. 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.
17. E. Union, "Artificial intelligence act (ai act) - regulation (eu) 2024/1689," 2024. Official Journal of the European Union, L 1689, 12 July 2024.
18. 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.
19. 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.
20. D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception," *Annu. Rev. Neurosci.*, vol. 31, no. 1, pp. 411–437, 2008.
21. J. Fridrich, "Steganography in digital media: Principles, algorithms, and applications," *Steganography in Digital Media*, 01 2010.
22. T. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1793–1807, 2013.
23. O. Henniger *et al.*, "On the assessment of face image quality based on handcrafted features," in *Proceedings of BIOSIG*, 2020.
24. 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.

25. L. Chen, C. Pan, and Y. Fang, "Blind image quality assessment with pseudo-mos learning," *IEEE Signal Processing Letters*, vol. 28, pp. 1090–1094, 2021.
26. X. Liu *et al.*, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *ICCV*, 2017.
27. J. Wu *et al.*, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Transactions on Image Processing*, vol. 29, pp. 7414–7426, 2020.
28. L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR*, 2014.
29. P. Terhörst *et al.*, "Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *CVPR*, 2020.
30. Q. Meng *et al.*, "Magface: A universal representation for face recognition and quality assessment," in *CVPR*, 2021.
31. J. Tremoço, I. Medvedev, and N. Gonçalves, "Qualface: Adapting deep learning face recognition for id and travel documents with quality assessment," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–6, 2021.
32. K. Hernandez-Ortega *et al.*, "Faceqnet: Quality assessment for face recognition based on deep learning," in *ICB*, 2019.
33. L. Jin *et al.*, "Pipal: A large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020.
34. L. DeBruine and B. Jones, "Face research lab london set," *Open Science Framework*, 2017.
35. 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.
36. M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," June 2020.
37. 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.
38. 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.
39. 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.
40. Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2022.
41. 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.
42. F. Xiao and Z. Wang, "On the validity of common iqa metrics: A statistical perspective," *Signal Processing: Image Communication*, vol. 57, pp. 117–127, 2017.
43. D. Basak, S. Pal, and D. Patranabis, "Support vector regression," *Neural Information Processing – Letters and Reviews*, vol. 11, 11 2007.
44. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
45. T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 785–794, ACM, Aug. 2016.

46. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
47. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
48. 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.
49. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," 2003.