Pseudo-MOS Learning: A 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, Portugal
[2] Institute of Systems and Robotics, Coimbra, Portugal
andre.neto@isr.uc.pt, nunogon@deec.uc.pt

Abstract. A persistent gap remains between standard Image Quality Assessment (IQA) metrics and human perceptual judgments, typically quantified via Mean Opinion Scores (MOS). This poses a key challenge in applications where perceived quality impacts performance, such as facial recognition. We introduce a new no-reference Face Image Quality Assessment (FIQA) metric, developed within a Full-to-No-Reference learning framework. The process begins with a full-reference fusion model trained to regress classical IQA scores against human MOS on a labeled subset. This model is used to generate pseudo-MOS scores for the full dataset. These labels then supervise a no-reference deep regressor based on ResNet-18 features, producing a perceptually aligned metric that estimates quality directly from distorted facial images. We tested our approach in the context of steganographically degraded facial images, showing its effectiveness in scenarios involving subtle distortions and limited human annotations.

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

1 Introduction

Image Quality Assessment (IQA) plays a critical role in domains such as biometric authentication, multimedia processing, and medical imaging [26, 34]. It broadly refers to the estimation of visual quality based on attributes like contrast, sharpness, noise, and the presence of artifacts. Within this domain, Facial Image Quality Assessment (FIQA) focuses specifically on facial images, where quality is not assessed in terms of visual aesthetics, but rather in terms of its impact on the performance of face recognition systems [5, 52].

IQA methods fall into two categories: subjective and objective. Subjective methods use human ratings to measure perceived quality, often summarized as Mean Opinion Scores (MOS) [28]. These scores are reliable but expensive to collect and not scalable. Objective methods rely on algorithms to estimate quality, either by comparing to a reference image or by analyzing features of the image itself.

Objective methods can be divided into full-reference (FR) and no-reference (NR). FR methods compare a distorted image to a clean reference. They are

often accurate but can only be used when a reference is available. NR methods estimate quality without a reference and are more practical in real-world settings, though they often struggle to generalize across distortions and content [1].

FIQA is a subdomain of IQA focused on facial images. It plays a key role in biometric applications such as identity verification, where reference images are typically unavailable, and is also relevant in non-biometric scenarios like surveillance and forensic analysis. Consequently, most FIQA methods are no-reference NR, relying on task-specific priors or learned representations to estimate image quality [22].

A key challenge in IQA is the gap between objective metrics and human perception. Classical metrics such as PSNR [17] (Peak Signal-to-Noise Ratio), SSIM [58] (Structural Similarity Index), and VIF [49] (Visual Information Fidelity) provide automatic quality estimates but often show weak correlation with human ratings across datasets [1]. This issue is more pronounced in facial images, where perceived quality is shaped by both image distortions and biases from the observer.

Studies have shown that FIQA is affected by both demographic and non-demographic biases. Perceived quality can vary with ethnicity, gender, or age, often due to dataset imbalance and observer subjectivity [5,30,52]. For example, darker skin tones tend to produce lower recognition accuracy, and female faces are often rated with lower quality scores [26]. These effects highlight the need for more inclusive and perceptually aligned quality metrics.

The International Civil Aviation Organization [42] (ICAO) and the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) 19794–5 standard [27] 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.

These biases raise ethical concerns. Legal frameworks, such as the European Convention on Human Rights (Article 14) [13], the Universal Declaration of Human Rights (Article 7) [41], the General Data Protection Regulation (Article 22) [55], the European Artificial Intelligence Act (2024) [56] and the United States Bill of Rights [25], aim to prevent discriminatory decisions. Still, biases persist, often introduced through human observers involved in labeling.

Neuroscience shows that face perception relies on the fusiform face area, a brain region specialized for facial stimuli [32, 54]. This biological specialization makes FIQA particularly sensitive to both stimulus features (e.g., age, gender, ethnicity, attractiveness) and the demographic background of the observers.

Steganographically distorted facial images pose a harder problem. Steganography hides data by slightly changing pixel values, often in ways that escape human detection [14]. Recent printed-proof techniques go further by ensuring that hidden data can survive physical printing and scanning noises, making them useful for secure document encoding. While these changes are visually minimal, they

can damage biometric features and reduce recognition accuracy [9,47,48,50]. NR methods are usually not designed to catch these small but critical degradations.

To handle these limitations, we propose a FIQA framework based on pseudo-MOS. We first use a small set of facial images labeled with overall quality scores to train a fusion model that combines FR metrics into a single predictor. This model generates pseudo-MOS for the rest of the dataset. Using these labels, we then train a NR deep regressor based on a ResNet-18 [20] pretrained on ImageNet [12], allowing it to estimate perceptual quality without needing a reference image.

Our approach bridges the gap between FR supervision and NR inference. It offers a scalable solution for evaluating images with subtle distortions, like steganography, and supports the development of quality assessment models tailored to domain-specific tasks. In doing so, it combines the accuracy of FR metrics with the practicality of NR models in a single IQA pipeline.

2 Related Work

Several fusion-based approaches have been proposed to better align IQA metrics with human perception. Liu et al. [35] introduced a multi-method fusion framework in which multiple FR-IQA scores are linearly combined through regression to better approximate human judgments. Similarly, Henniger et al. [21] developed a Random Forest model trained on handcrafted image features drawn from ISO face quality standards, improving predictive utility for biometric applications. These works show that fusing complementary quality cues improves correlation with MOS compared to single-metric methods

In the absence of subjective labels, several methods have adopted weakly supervised strategies based on pseudo-labels. Chen et al. [6] generated pseudo-MOS scores by averaging multiple FR-IQA scores. RankIQA [36] used synthetic degradations and ranking-based supervision to learn ordinal quality relationships. Wu et al. [60] trained cascaded CNN regressors on pseudo-MOS to support NR-IQA training. These methods show that pseudo-labeling can guide deep quality models when ground-truth MOS is limited.

Recent NR-IQA methods leverage deep features from CNNs pretrained on large datasets. Kang et al. [31] showed that CNNs can directly predict image quality from patches. In FIQA, SER-FIQ [51] (Stochastic Embedding Robustness for Face Image Quality) estimates quality by measuring the consistency of face embeddings under dropout. MagFace [39] links embedding magnitude to recognition performance to learn quality-aware features. FaceQnet [22] estimates how well a face image will perform in recognition tasks, using a regression model trained on features from a pre-trained network. QualFace [53] adapts face recognition networks for document images and adds a quality estimation branch aligned with ICAO and ISO/IEC standards. These approaches replace handcrafted indicators with learned representations optimized for face recognition.

Other studies emphasize that image quality is inherently task-specific. In FIQA, quality is defined not by visual aesthetics but by its effect on recogni-

tion performance. Standards such as ISO/IEC 19794–5 codify this operational perspective, specifying conditions for acceptable biometric image acquisition. Datasets such as PIPAL [29] and TID2013 [44], 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 [11] (FRLL) 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:

- MOS set (15 identities, 540 images): a core set of demographically diverse subjects, shown in Fig. 1, with subjective MOS annotations. It is split into:
  - MOS train set (12 identities): used to train the FR fusion model, regressing FR-IQA metrics to human MOS.
  - MOS test set (3 identities): held out from the framework and used only for final evaluation.
- Pseudo-MOS set (87 identities, 3,132 images): no subjective scores were collected for these images. Pseudo-MOS are generated for this set using the trained fusion model.
- NR train set: includes both the MOS train set and the pseudo-MOS set. It is used to train the NR regressor.

The printer-proof steganography methods used were based on Generative Adversarial Networks (GANs) [23] to encode and decode information, illustrated in Fig. 2. We can obtain various results depending on the method used:

- StegaStamp [50]: 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 adding printer noise in the training process.
- Code Face [47]: 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 [10]: 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.



(a) MOS train set **(b)** MOS test set

Fig. 1. Reference images from the MOS set, comprising the selected subjects from the FRLL dataset.

– StampOne [48]: focuses on high-level robust steganography, such as [47, 50], it balances 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 and wavelet transform to normalize and balance frequencies of the inputs.



(a) StegaStamp **(b)** Code Face **(c)** RiemStega **(d)** StampOne

Fig. 2. Steganographically distorted facial images from each method.

We followed the ITU-R BT.500–15 [28] recommendation and adopted the Single Stimulus (SS) method. The test was implemented using a custom Django web application seen in Fig. 3. Prior to the test session, participants signed an informed consent form and filled out a registration form providing demographic and environmental information such as age, gender, education, country of origin

and ethnicity, and others. Each image was shown individually, with no time limit. Ratings were submitted using a labeled slider, and automatic saving ensured session robustness.

Each image in the MOS set was evaluated approximately 30 times by human observers, resulting in over 14,000 ratings. We had around 200 participants, each session lasted about 22 minutes and included roughly 70 evaluations. Following the session, outlier observers were identified and removed using both Kurtosis-based and correlation-based post-screening methods described in ITU-R BT.500–15 [28], resulting in the exclusion of four participants.



Fig. 3. Django-based webapp created for the Single Stimulus test.

3.2 Debiasing of Subjective Scores

To correct for demographic bias in the subjective scores, we followed a procedure inspired by prior work on bias correction in perceptual tasks [8], where we applied a residualization method based on linear modeling. An ordinary least squares (OLS) regression was fit to the MOS values, using observer and image attributes, and their pairwise interactions as categorical predictors. The fitted bias components were subtracted from the original scores, and the residuals were mean-centered to preserve the global score distribution. As shown in Table 1, several factors exhibit statistically significant effects on the MOS prior to residualization, notably observer and subject ethnicity. After applying the residualization procedure, these effects disappear, as confirmed by an ANOVA test showing no significant impact from any individual factor. The corrected MOS labels are then used as ground truth in all supervised stages of the pipeline to ensure fairness and reduce the influence of socially conditioned priors.

3.3 Correlation of FR-IQA Metrics with Human Perception

We compute 40 FR-IQA scores for each distorted image in the dataset and compare them against the corresponding MOS, as seen in Fig. 4. Several metrics

Table 1. ANOVA [46] results for observer and image attributes. Before debiasing, several factors show statistically significant effects on MOS, p-value < 0.05. After residualization, all main effects show no significant impact, confirming the effectiveness of the debiasing procedure.

Factor	p-value	p-value (residualized)
Observer gender	0.022	0.9930
Observer ethnicity	8.44×10^{-4}	1
Subject gender	1.60×10^{-3}	0.9722
Subject ethnicity	7.36×10^{-3}	1
Observer gender $\times$ Subject gender	0.6417	0.6417
Observer ethnicity $\times$ Subject ethnicity	0.0582	0.0582

exhibit strong linear trends with MOS, while others are poorly aligned or even negatively correlated. For a detailed description of these metrics, we refer the reader to [1].

3.4 Fusion of FR-IQA Metrics for Pseudo-MOS Estimation

To identify which metrics align best with human perception, we compute both the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC) [61] which respectively quantify the linearity and monotonicity of the relationship between metric scores and MOS. To determine the appropriate number of metrics to retain for fusion, we applied Singular Value Decomposition (SVD) and the Picard criterion [18]. Metrics are first ranked by the average of their PLCC and SRCC with MOS. After normalizing the feature matrix, SVD revealed that six components capture 95% of the total variance, as shown in Fig. 5a, indicating an optimal dimensionality of $k = 6$.

To validate this truncation point, we examined the Picard plot in Fig. 5b, which compares singular values with the target projections. The stable ratio in the tail confirms that six components provide a good balance between expressiveness and stability. This supports a compact, informative subset of FR-IQA metrics.

After selecting the top $k = 6$ FR-IQA metrics, we trained a diverse set of supervised regressors to map these features to subjective MOS scores. The models included both linear and non-linear types:

- Linear models: Linear Regression [19], Ridge Regression [24], Bayesian Ridge [3], ElasticNet [63].
- Kernel-based: Support Vector Regression [2] (SVR).
- Ensembles: Random Forest [4], Extra Trees [16], Gradient Boosting [15], HistGradientBoosting [43].
- Boosted Trees: XGBoost [7], LightGBM [33], CatBoost [45].

Each regressor was trained using five-fold cross-validation with an exhaustive grid search over predefined hyperparameter spaces. The best model, CatBoost,

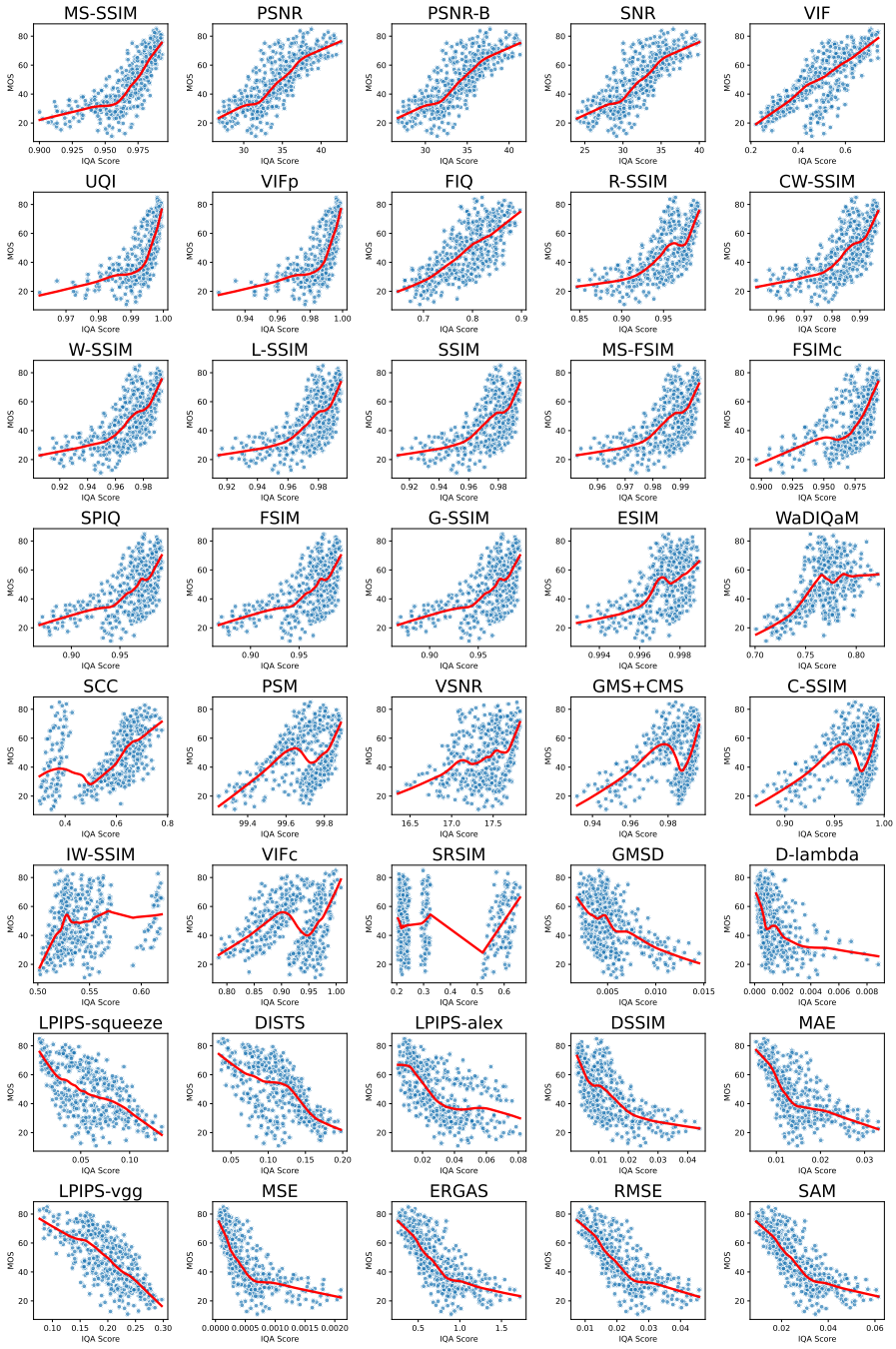
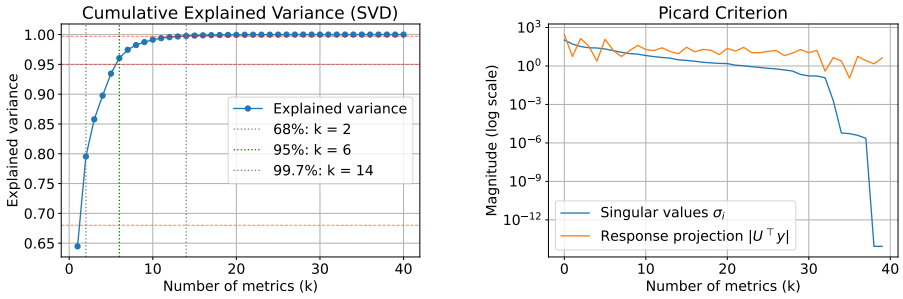


Fig. 4. Scatter plots illustrating the relationship between MOS and 40 individual full-reference IQA metrics. The red line represents a smoothed local trend curve.



(a) Cumulative variance explained by SVD components.

(b) Magnitude of singular values and projected response.

Fig. 5. SVD-based analysis of the FR metric space. The cumulative explained variance (a) guides the choice of dimensionality, while the Picard criterion (b) illustrates stability near $k = 6$.

optimal configuration was: depth = 8, iterations = 500, learning rate = 0.05, and L_2 regularization = 1. This trained regressor is then applied to the unlabeled portion of the dataset (3,132 distorted images), generating pseudo-MOS.

3.5 Training a No-Reference IQA Model from Pseudo-MOS

To enable NR quality prediction, we train a deep regression model end-to-end using pseudo-MOS scores as targets. The architecture is based on a ResNet-18 backbone [20] pretrained on ImageNet [12], with its final classification layer replaced by a lightweight multi-layer perceptron (MLP) regressor [3]. All layers are fine-tuned during training to learn perceptual quality representations specific to our task. The training set consists of distorted facial images paired with either pseudo-MOS, from the FR fusion, or real MOS labels, when available. The model is optimized using MSE and evaluated on the disjoint MOS test set of labeled images. Once trained, the model infers image quality solely from the distorted input, enabling NR assessment aligned with human perception.

4 Results

4.1 Full-Reference Fusion Metric Performance

To interpret feature contributions, we used SHAP [37] (SHapley Additive exPlanations) values derived from CatBoost's internal structure. SHAP values quantify how much a feature pushes the model's output away from the expected value. As shown in Fig. 6, MS-SSIM [59] (Multiscale SSIM) and UQI [58] (Universal Quality Index) had the largest positive impact on predictions, while PSNR-B [38] (PSNR-Blocking), VIF [49], PSNR [62] and SNR [17] (Signal-to-Noise Ratio) contributed with moderate influence. The color encoding further highlights how high and low feature values affect the model output.

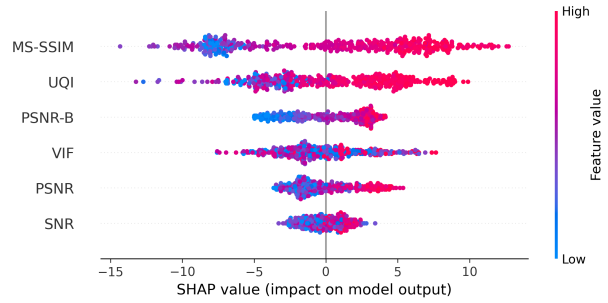


Fig. 6. SHAP summary plot for $k = 6$, generated from the best CatBoost model.

When comparing the fusion metric with the individual FR metrics, Table 2 shows that the fusion metric outperformed all individual metrics in terms of PLCC, SRCC, MSE and MAE. This demonstrates the effectiveness of our approach in combining multiple FR metrics to improve overall performance.

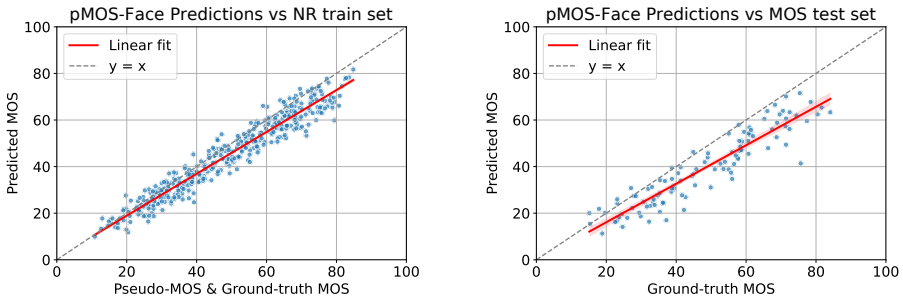
Table 2. Performance of our Fusion metric and the selected FR-IQA metrics. Higher PLCC and SRCC indicates better correlation with MOS. Lower MSE and MAE indicate more accurate quality estimates.

Metric	PLCC	SRCC	MSE	MAE
Fusion (CatBoost)	**0.8881**	**0.8893**	**65.07**	**6.160**
MS-SSIM	0.7712	0.8489	2542	47.27
PSNR	0.8009	0.8132	420.7	16.57
PSNR-B	0.7996	0.8126	434.8	16.84
SNR	0.7931	0.8047	493.9	18.08
VIF	0.7723	0.7671	2584	47.75
UQI	0.6866	0.8085	2540	47.24

4.2 No-Reference Regression Model Performance

We refer to our NR-IQA model as pMOS-Face (pseudo-MOS for Face), reflecting its use of pseudo-MOS labels generated from the FR fusion model and its evaluation on faces. This framework can be adapted to other datasets, provided that a small set of images with human labels is available for training the FR fusion model. Fig. 7 shows two evaluation scenarios: in Fig. 7(a), the model's predictions are compared against the pseudo and ground-truth MOS used during training, yielding a PLCC of 0.9285, SRCC of 0.9345, MSE of 120.2, and MAE of 9.158. Fig. 7(b) shows the predictions compared against the MOS test set, resulting in a PLCC of 0.9679, SRCC of 0.9691 and MSE of 35.34 and MAE

of 4.743. These results confirm both generalization to unseen human labels and alignment with the pseudo-supervision used for training.



(a) Training plot. **(b)** Test plot.

Fig. 7. Evaluation of our NR-IQA model. Panel (a) shows performance on the training set with pseudo and ground-truth MOS; panel (b) shows performance on the disjoint test set with ground-truth MOS.

The results indicate that our model is more effective in predicting perceptual quality than classical NR-IQA metrics such as NIQE [40] and PIQE [57], as well as task-specific approaches like SER-FIQ [51] and MagFace [39]. This highlights the robustness of our weakly supervised strategy for NR-IQA on steganographically degraded facial images. Quantitative results are reported in Table 3.

Table 3. Performance of our pMOS-Face compared to standard NR-IQA baselines.

Metric	PLCC	SRCC	MSE	MAE
pMOS-Face (ours)	**0.9285**	**0.9345**	**120.6**	**9.158**
NIQE	0.7536	0.7431	6859	82.62
PIQE	0.2993	0.3114	3297	57.05
SER-FIQ	-0.1648	-0.1542	123.5	9.230
MagFace	-0.6095	-0.6362	4437	66.24

When considering a more generalized use, where observer bias is not accounted for, our NR-IQA model shows a major improvement over the baselines. This is particularly relevant for task specific applications, where the model can be trained on a small set of images with human labels and then applied to larger datasets without the need for additional supervision. This approach allows for a more scalable and efficient solution not only for FIQA but also for other domains where reference images are not available.

5 Conclusion and Future Work

We proposed a 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 (10% of the dataset), generating pseudo-MOS labels for a larger unlabeled dataset. These forged 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 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.

6 Acknowledgments

The author would like to thank Dr. Shahrukh Athar for his support during the initial state-of-the-art review. His comprehensive list of implemented IQA metrics from his work were invaluable in kick-starting this study. The author also gratefully acknowledges the Institute of Systems and Robotics (ISR) for providing the resources and research environment that made this work possible. Lastly, we gratefully acknowledge everyone who contributed to the MOS labeling process, as their efforts were crucial in establishing the foundation for our research. This study has received funding from the EU Horizon Europe for the ACHILLES project under Grant Agreement No. 101189689, and FCT – Fundação para a Ciência e a Tecnologia, I.P., under the project UIDB/00048/2020 (DOI 10.54499/UIDB/00048/2020).

References

1. Athar, S., Wang, Z.: A comprehensive performance evaluation of image quality assessment algorithms. IEEE Access **7**, 140030–140070 (2019). https://doi.org/10.1109/ACCESS.2019.2943319
2. Basak, D., Pal, S., Patranabis, D.: Support vector regression. Neural Information Processing - Letters and Reviews **11** (2007)
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
4. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
5. Cavazos, J.G., Phillips, P.J., Castillo, C.D., O'Toole, A.J.: Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(1), 101–111 (2021). https://doi.org/10.1109/TBIOM.2020.3027269
6. Chen, L., Pan, C., Fang, Y.: Blind image quality assessment with pseudo-mos learning. IEEE Signal Processing Letters **28**, 1090–1094 (2021)

7. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: ACM SIGKDD. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
8. Clapés, A., Oliu, M., Farrús, M., Escalera, S.: From apparent to real age: Gender, age, ethnicity and error perception. In: CVPR Workshops. pp. 0–0 (2018)
9. Cruz, A., Schardong, G., Schirmer, L., Marcos, J., Shadmand, F., Gonçalves, N.: Riemannian loss for image similarity-based comparison applied to printer-proof steganography (2023), submitted to 37th Conference on Neural Information Processing Systems
10. Cruz, A., Schardong, G., Schirmer, L., Marcos, J., Shadmand, F., Gonçalves, N.: Riemstega: Covariance-based loss for print-proof transmission of data in images. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
11. DeBruine, L., Jones, B.: Face research lab london set. Open Science Framework (2017), https://osf.io/98q3w/
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
13. of Europe, C.: European convention on human rights, article 14: Prohibition of discrimination (1950), https://www.echr.coe.int/Documents/Convention_ENG.pdf, council of Europe Treaty Series No. 5
14. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications (2010). https://doi.org/10.1017/CBO9781139192903
15. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**(5), 1189–1232 (2001)
16. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning **63**(1), 3–42 (2006)
17. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, 2nd edn. (2002)
18. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM (1998)
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2nd edn. (2009)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
21. Henniger, O., et al.: On the assessment of face image quality based on handcrafted features. In: BIOSIG (2020)
22. Hernandez-Ortega, K., et al.: Faceqnet: Quality assessment for face recognition based on deep learning. In: ICB (2019)
23. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018)
24. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
25. House, T.W.: Blueprint for an ai bill of rights: Making automated systems work for the american people (2022), https://www.whitehouse.gov/ostp/ai-bill-of-rights/, office of Science and Technology Policy, United States
26. Huang, C., Li, Y., Loy, C.C., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(11), 2781–2794 (2020). https://doi.org/10.1109/TPAMI.2019.2914680
27. International Organization for Standardization: Information technology - biometric sample quality - part 5: Face image data (April 2010)

28. 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: https://www.itu.int/rec/R-REC-BT.500-15
29. Jin, L., et al.: Pipal: A large-scale image quality assessment dataset for perceptual image restoration. In: ECCV (2020)
30. Kabbani, W., Raja, K., Ramachandra, R., Busch, C.: Demographic variability in face image quality measures. In: International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–6 (2024). https://doi.org/10.1109/BIOSIG61931.2024.10786726
31. Kang, L., Ye, P., Li, Y., Doermann, D.: Convolutional neural networks for no-reference image quality assessment. In: CVPR (2014)
32. Kanwisher, N., Yovel, G.: The fusiform face area: A cortical region specialized for the perception of faces. Philosophical Transactions of the Royal Society B: Biological Sciences **361**(1476), 2109–2128 (2006)
33. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. In: NeurIPS. vol. 30 (2017)
34. Kim, H.I., Lee, S.H., Yong, M.R.: Face image assessment learned with objective and relative face image qualities for improved face recognition. In: IEEE International Conference on Image Processing (ICIP). pp. 4027–4031 (2015)
35. Liu, T., Lin, W., Kuo, C.C.J.: Image quality assessment using multi-method fusion. IEEE Transactions on Image Processing **22**(5), 1793–1807 (2013)
36. Liu, X., et al.: Rankiqa: Learning from rankings for no-reference image quality assessment. In: ICCV (2017)
37. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS) **30** (2017)
38. Ma, L., Wang, S., Shi, G., Zhao, D., Gao, W.: Psnr-b: A novel peak signal-to-noise ratio based on edge preservation. IEEE Journal of Selected Topics in Signal Processing **6**(6), 537–549 (2012). https://doi.org/10.1109/JSTSP.2012.2204259
39. Meng, Q., et al.: Magface: A universal representation for face recognition and quality assessment. In: CVPR (2021)
40. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013). https://doi.org/10.1109/LSP.2012.2227726
41. Nations, U.: Universal declaration of human rights, article 7: Equality before the law (1948), https://www.un.org/en/about-us/universal-declaration-of-human-rights, general Assembly Resolution 217 A
42. Organization, I.C.A.: Doc 9303 machine readable travel documents, part 3: Specifications common to all mrtds (September 2015)
43. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
44. Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al.: Image database tid2013: Peculiarities, results and perspectives. Signal processing: Image communication **30**, 57–77 (2015)
45. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: Unbiased boosting with categorical features. In: NeurIPS. vol. 31 (2018)
46. Ross, A., Willson, V.L.: One-way anova. Basic and Advanced Statistical Tests pp. 21–24 (2017)

47. Shadmand, F., Medvedev, I., Goncalves, N.: Codeface: A deep learning printer-proof steganography for face portraits. IEEE Access **9**, 167282–167291 (2021). https://doi.org/10.1109/ACCESS.2021.3132581
48. Shadmand, F., Medvedev, I., Schirmer Silva, L., Sena Marcos, J., Gonçalves, N.: Stampone: Addressing frequency balance in printer-proof steganography. In: CVPR Workshops. pp. 4367–4376 (2024). https://doi.org/10.1109/CVPRW63382.2024.00440
49. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing **15**(2), 430–444 (2006)
50. Tancik, M., Mildenhall, B., Ng, R.: Stegastamp: Invisible hyperlinks in physical photographs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
51. Terhörst, P., et al.: Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In: CVPR (2020)
52. Terhörst, P., Kolf, J.N., Damer, N., Kirchbuchner, F., Kuijper, A.: Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In: IEEE International Joint Conference on Biometrics (IJCB). pp. 1–11 (2020). https://doi.org/10.1109/IJCB48548.2020.9304865
53. Tremoço, J., Medvedev, I., Gonçalves, N.: Qualface: Adapting deep learning face recognition for id and travel documents with quality assessment. In: International Conference of the Biometrics Special Interest Group (BIOSIG). pp. 1–6 (2021). https://doi.org/10.1109/BIOSIG52210.2021.9548309
54. Tsao, D.Y., Livingstone, M.S.: Mechanisms of face perception. Annual Review of Neuroscience **31**(1), 411–437 (2008)
55. Union, E.: General data protection regulation (eu) 2016/679, article 22: Automated individual decision-making, including profiling (2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj, official Journal of the European Union, L 119
56. Union, E.: Artificial intelligence act (ai act) - regulation (eu) 2024/1689 (2024), https://eur-lex.europa.eu/eli/reg/2024/1689/oj, official Journal of the European Union, L 1689, 12 July 2024
57. Venkatanath, A., Praneeth, R., Babu, C.S., Gopalakrishna, M.A., Nair, A.S., Prabhakar, A.: Blind image quality evaluation using perception based features. National Conference on Communications (NCC) pp. 1–6 (2015). https://doi.org/10.1109/NCC.2015.7084923
58. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
59. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: Asilomar Conference on Signals, Systems and Computers. vol. 2, pp. 1398–1402 (2003)
60. Wu, J., et al.: End-to-end blind image quality prediction with cascaded deep neural network. IEEE Transactions on Image Processing **29**, 7414–7426 (2020)
61. Xiao, F., Wang, Z.: On the validity of common iqa metrics: A statistical perspective. Signal Processing: Image Communication **57**, 117–127 (2017). https://doi.org/10.1016/j.image.2017.05.001
62. Yoo, J.C., Ahn, C.: Image matching using peak signal-to-noise ratio-based occlusion detection. IET Image Processing **6**(5), 483–495 (2012)
63. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B **67**(2), 301–320 (2005)