UNIVERSIDADE D
COIMBRA

André Guilherme dos Santos Neto

METRICS OF AUTOMATIC IMAGE QUALITY ASSESSMENT BASED ON HUMAN PERCEPTION

A COMPARATIVE STUDY AND A PROPOSAL OF A NEW METRIC

VOLUME 1

Dissertação no âmbito do Mestrado em Engenharia Eletrotécnica e de Computadores orientada pelo Professor Doutor Nuno Miguel Mendonça da Silva Gonçalves e apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Fevereiro de 2025

UNIVERSIDADE D
COIMBRA

Metrics of automatic Image Quality Assessment based on Human Perception — A comparative study and a proposal of a new metric

Supervisor:

Nuno Miguel Mendonça da Silva Gonçalves

Jury:

João Pedro de Almeida Barreto

Nuno Miguel Mendonça da Silva Gonçalves

Vítor Manuel Mendes da Silva

Dissertation submitted in partial fulfillment for the degree of Master of Science in Electrical and Computer Engineering.

Coimbra, February 2025

Acknowledgments

Resumo

Mantém-se uma discrepância persistente entre as métricas padrão de Avaliação da Qualidade de Imagem (IQA) e os juízos perceptivos humanos, geralmente quantificados através dos Mean Opinion Scores (MOS). Esta divergência constitui um desafio central em aplicações onde a qualidade percebida afeta o desempenho, como no reconhecimento facial. Apresentamos uma nova métrica de Avaliação da Qualidade de Imagens Faciais (FIQA) sem referência, desenvolvida no âmbito de uma estrutura de aprendizagem do tipo Full-to-No-Reference. O processo tem início com um modelo de fusão com referência completa, treinado para fazer a regressão de métricas IQA clássicas em função dos MOS humanos num subconjunto anotado. Este modelo é então utilizado para gerar pseudo-MOS para o conjunto completo de dados. Estes rótulos supervisionam depois um regressor profundo sem referência baseado em características extraídas da ResNet-18, originando uma métrica alinhada com a perceção humana que estima a qualidade diretamente a partir de imagens faciais degradadas. Testámos a nossa abordagem no contexto de imagens faciais afetadas por esteganografia, demonstrando a sua eficácia em cenários com distorções subtis e anotações humanas limitadas.

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.

"O que acontece no Mundo é que toda a gente que nasce, nasce de alguma maneira poeta. Inventor de qualquer coisa que não havia no Mundo ainda, antes deles nascerem. E inteiramente individual. Cada um poeta que é!"

Agostinho da Silva

Contents

List of Acronyms

List of Figures

List of Tables

1 Introduction

1.1 Contextualization

Image quality refers to the degree to which a visual representation meets perceptual or functional expectations. It is typically associated with the presence or absence of distortions, artifacts, or degradations that affect how an image is perceived by humans or processed by machines [1], [2], [3]. High-quality images preserve structural, textural, and color information in a way that aligns with human visual preferences or supports reliable computer vision performance [4], [5].

Image Quality Assessment (IQA) is the process of quantifying image quality in a systematic and reproducible way. It plays a central role in optimizing image acquisition [6], compression [7], transmission [8], and restoration pipelines [9]. IQA methods aim to provide reliable quality estimates that correlate well with human perception [10], [11]. Due to the complexity of human vision and its subjective nature, building computational models that accurately reflect perceived quality remains a challenging task.

IQA methods are generally categorized as subjective or objective. Subjective assessment involves human observers who rate the perceived quality of images, typically following standardized protocols such as ITU-R BT.500 [12]. While subjective methods are considered the gold standard due to their direct alignment with human perception, they are time-consuming, expensive, and not scalable. In contrast, objective methods rely on computational models to estimate image quality automatically, aiming to approximate human judgment with high consistency and low cost [4], [13].

Subjective image quality assessment relies on human evaluations to generate ground-truth perceptual scores, often in the form of Mean Opinion Scores (MOS) or Difference Mean Opinion Scores (DMOS). These scores are typically collected under controlled conditions following international standards such as ITU-R BT.500 [12] or ITU-T P.910 [14]. Subjective assessment captures nuances of human vision that are difficult to model algorithmically, making it essential for validating and benchmarking objective IQA models [13], [15]. However, it is inherently limited by inter-observer variability, cultural or demographic bias, and the logistical costs associated

with large-scale studies [16]. Crowdsourcing platforms like Amazon Mechanical Turk [17] have recently enabled more scalable data collection, though at the expense of environmental control and consistency.

Objective IQA methods are commonly divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). FR-IQA assumes access to an undistorted reference image and compares it to the distorted version using perceptual models [10], [18], [19]. RR-IQA methods extract partial information from the reference image, enabling a compromise between performance and practicality [20], [21]. NR-IQA, also known as blind IQA, operates without any reference and is considered the most challenging, requiring models to infer quality based solely on the distorted image [22], [23].

Facial Image Quality Assessment (FIQA) is a specialized subfield of IQA that aims to quantify the quality of facial images with respect to their utility in downstream biometric or analytic tasks. Unlike general-purpose IQA, FIQA must account for both perceptual attributes (e.g., blur, noise) and task-specific considerations such as facial pose, occlusion, expression, and alignment [24], [25], [26]. High-quality face images are crucial for ensuring the accuracy and fairness of face recognition systems, particularly in security, forensics, and surveillance domains [27], [28]. Consequently, FIQA models often incorporate features from deep face recognition networks to align quality predictions with identity discrimination performance [29]. This task-oriented nature of FIQA makes it fundamentally different from traditional perceptual quality assessment.

FIQA systems have been shown to produce different quality scores depending on a person's age, gender, or skin tone [29], [30]. These differences often happen because the training data includes more examples from certain groups and fewer from others. For example, images of people with darker skin or non-frontal poses are often less common in training sets, which leads to lower quality scores for those individuals [31]. This can cause face recognition systems to perform worse for some groups than others, raising serious concerns about fairness.

The International Civil Aviation Organization [32] (ICAO) and the ISO/IEC 19794–5 standard [33] 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) [34], the Universal Declaration of Human Rights (Article 7) [35], the General Data Protection Regulation (Article 22) [36], and emerging AI governance frameworks, such as the European Artificial Intelligence Act (2024) [37] and

proposals in the USA [38], 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 [31], [39]. 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 [40]. 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.

A particularly relevant subclass of steganography for identity documents is printer-proof steganography, which embeds information in a way that remains intact through the print-scan process [41], [42]. These techniques aim to survive real-world transformations such as color shifts, compression, and physical degradation, while maintaining visual fidelity. However, even when such methods are visually imperceptible, they can induce subtle frequency-domain changes or spatial artifacts that disrupt face recognition systems. Assessing image quality in this context requires NR-IQA methods that are sensitive not only to perceptual degradation but also to recognition-relevant cues.

1.2 Problem Statement

Despite significant advances in FIQA, current approaches struggle to capture task-specific degradations introduced by modern steganographic techniques. Existing methods often rely on handcrafted features [43], supervised quality learning [44], [45], [46], or rank-based formulations [47]. However, these frameworks are typically trained on standard datasets and are not optimized to handle distortions that preserve visual fidelity while disrupting utility. Moreover, demographic and perceptual biases embedded in training data [48] raise concerns about the fairness and generalizability of FIQA systemsn. This is especially problematic for applications involving MRTDs, where regulation mandates consistent quality but does not address ethical or perceptual variance [32], [33]. Therefore, there is a critical need for NR-IQA methods that can detect subtle, high-impact degradations in facial images, particularly those arising from adversarial

steganography while accounting for demographic fairness and interpretability.

Traditional IQA frameworks tend to generalize across domains, but real-world applications demand task-aware assessment strategies. In the context of facial biometrics, image quality must be evaluated not only in terms of visual fidelity but also in terms of its impact on recognition performance and fairness. This motivates the development of application-specific IQA pipelines that consider context-dependent factors, such as demographic variability and task utility. At the same time, the perceptual dimension of image quality remains central. Human observers are still used as the ground truth in subjective assessments [49], [50], yet NR-IQA models often diverge from these judgments, especially in complex scenarios involving subtle degradations or diverse populations. To bridge this gap, new strategies are needed to align NR-IQA with subjective perception, either by integrating perceptual priors, modeling bias sources, or learning from pseudo-MOS annotations that reflect human preferences [51], [52].

To address the limitations of existing approaches, we propose a NR-IQA framework tailored to facial images affected by subtle or task-relevant distortions, such as those introduced by steganography. Our method relies on a fusion strategy that combines multiple FR metrics, trained using supervised regression against a curated set of subjective scores. This fused metric serves as a proxy ground truth to guide the learning of a NR-IQA model, enabling it to approximate perceptual quality without requiring access to pristine reference images. The framework is evaluated across a steganographically augmented dataset of facial images, incorporating multiple embedding levels and distortion types. By integrating perceptual fidelity, recognition utility, and robustness to imperceptible alterations, the proposed approach seeks to bridge the gap between unbiased human perception and automatic assessment.

2 Related Work

2.1 Subjective IQA

Subjective image quality assessment is grounded in human visual perception and remains the gold standard for evaluating visual fidelity. Standardized by ITU-R BT.500 [49] and ITU-T P.910 [53], traditional methodologies rely on controlled laboratory conditions, calibrated displays, and predefined rating procedures to ensure reproducibility and validity.

2.1.1 Assessment Protocols

Assessment protocols are typically divided into two categories: quality assessment, which estimates the overall perceived quality of an image, and impairment assessment, which evaluates the severity of degradation relative to an undistorted reference. In image evaluation, two main methodological paradigms are used: Single Stimulus (SS), where one image is shown at a time and rated for quality; and Double Stimulus (DS), where both the reference and distorted images are shown together to assess perceptual difference. Rating scales can be numerical (e.g., 1 to 11, or 1 to 100) or categorical, with labels for quality (e.g., bad, poor, fair, good, excellent) or impairment (e.g., very annoying, annoying, slightly annoying, perceptible but not annoying, imperceptible). Each method differs in cognitive demand, sensitivity to bias, and statistical power, and the appropriate choice depends on the specific goals of the assessment.

After subjective scores are collected, typically from at least thirty observers per image, the data undergo rigorous statistical screening to ensure reliability and consistency. The first step involves identifying and discarding outlier observers whose scoring behavior significantly deviates from the population. This is usually performed by computing the Pearson correlation coefficient between an observer's scores and the preliminary mean scores across all images. Observers whose correlation falls below a predefined threshold (e.g., $r < 0.75$) are flagged as unreliable and removed. Following outlier removal, the scores are averaged to compute the final Mean Opinion Score (MOS) or Difference Mean Opinion Score (DMOS), depending on whether quality or impairment was assessed.

The MOS is computed as the arithmetic mean of all valid scores for a given image:

$$\text{MOS} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{2.1}$$

where S_i is the score given by the i-th subject, and N is the number of valid observers.

For double-stimulus tests, the DMOS reflects the perceptual degradation with respect to a reference:

$$\text{DMOS} = \text{MOS}_{\text{reference}} - \text{MOS}_{\text{distorted}} \tag{2.2}$$

Confidence intervals for MOS/DMOS are also computed, usually assuming a normal distribution. Statistical tests such as t-tests or ANOVA may be applied to assess differences between image groups or distortion types. This rigorous post-processing pipeline ensures that the final subjective scores are both perceptually meaningful and statistically valid.

2.1.2 Crowdsourcing

Beyond laboratory-controlled protocols, crowdsourcing has emerged as a scalable alternative for subjective IQA. Platforms such as Amazon Mechanical Turk (AMT) and Prolific allow researchers to gather perceptual scores from a geographically and demographically diverse participant pool. To mitigate the loss of environmental control, studies often integrate reliability checks, such as repeated image pairs or gold-standard trials [50]. While the quality of crowdsourced labels may vary, prior work has shown that when properly filtered, crowdsourced MOS can achieve correlation levels comparable to those from lab-based assessments [52]. This approach enables the creation of large-scale IQA datasets, significantly expanding the empirical foundation for training and benchmarking objective metrics.

2.1.3 Benchmark Datasets

Several benchmark datasets have been curated to support the development and evaluation of IQA models. The LIVE [54], TID2013 [55], and CSIQ [56] datasets follow ITU-R BT.500-compliant procedures, offering high-quality MOS/DMOS annotations across a range of distortion types. More recent datasets, such as PIPAL [52] and KonIQ-10k [50], leverage crowdsourced assessments to achieve broader coverage of real-world content and perceptual variance. These datasets not only facilitate benchmarking but also allow for training data-driven IQA models in both full and no-reference scenarios. Nonetheless, domain-specific datasets, particularly those for biometric, medical, or forensic applications, remain scarce, limiting the generalizability of learned metrics in those contexts.

2.2 Objective IQA

Objective IQA refers to the automatic estimation of visual quality using computational models. Unlike subjective methods that rely on human ratings, objective IQA provides consistent, scalable evaluations suitable for real-world tasks such as compression, transmission, restoration, and enhancement [54], [57]. A wide range of objective metrics has been proposed, reflecting different assumptions about the human visual system and the nature of image distortions. Surveys such as Wang et al. [58], Liu et al. [59], and Ding et al. [60] review dozens of existing metrics, highlighting both classical signal-based models and recent learning-based approaches. While no metric perfectly matches human perception, objective methods offer a low-cost, repeatable alternative for benchmarking, model optimization, and large-scale quality monitoring.

These metrics can be further characterized by the domain in which they operate and the spectral representation they use. The domain refers to the signal space employed during computation. Classical metrics operate in the spatial domain, comparing pixel intensities directly, as seen in PSNR and SSIM [61]. Others use the frequency domain, employing transforms such as the discrete cosine transform (DCT) or wavelet decompositions to capture structural differences, as in VIF [4] and IFC [62]. A third group targets the spectral domain, often used in remote sensing, where images are compared across multiple wavelength bands using metrics such as ERGAS [63] and SAM [64]. Additional domains include gradient-based methods that emphasize edge strength or direction, such as GMSD [65], and deep feature domains, which assess perceptual similarity using activations from convolutional neural networks trained on natural images, as in LPIPS [66] and DISTS [60].

The spectral representation, or color spectrum, denotes the channel configuration used by the metric. Many classical approaches focus solely on the luminance channel (Y), which aligns closely with human perception [4], [67]. Others compute scores over grayscale or full RGB images, as seen in FSIM [68] and its color extension FSIMc. Some metrics operate in alternative color spaces such as YCbCr or CIELab [69], while hyperspectral or multiband metrics process dozens of spectral channels, particularly in remote sensing applications [70], [71]. Table **??** summarizes the 40 IQA metrics considered in this work, detailing their reference type, computational domain, publication year, number of tested databases, and spectral characteristics.

Table 2.1: Summary of the 40 IQA metrics used

Metric	Type	Domain	Color Spectrum	Year	Tested DBs
MSE [72],	Error-based	Spatial	Grayscale	1977	Many
MAE [72],	Error-based	Spatial	Grayscale	1977	Many

(continuation)

Metric	Type	Domain	Year	Tested Databases	Spectrum	
RMSE [72]	Error-based	Spatial	Grayscale	1977	Many	
PSNR [72]	Error-based	Spatial	Grayscale / RGB	1977	Many	
PSNR-B	Error-based	Spatial	Grayscale / RGB	2008	Few	
SNR	Error-based	Spatial	Grayscale / RGB	1977	Many	
ERGAS	Error-based	Remote sensing	Multiband	2000	Few	
D_λ	Error-based	Remote sensing	Multiband	2004	Few	
SAM	Error-based	Remote sensing	Multiband	1993	Few	
SCC	Error-based	Spatial	Grayscale / RGB	1904	Many	
C-SSIM [73]	FR	spatial	2017	2	RGB	
CW-SSIM [74]	FR	complex wavelet	2009	1	grayscale	
DISTS [75]	FR	deep features	2020	4	RGB	
DSSIM [61]	FR	spatial	2004	—	grayscale	
ERGAS [63]	FR	spectral	2000	1	multi-band	
ESIM (ESSIM) [76]	FR	spatial (edges)	2013	2	grayscale	
FIQ (IFC) [77]	FR	wavelet	2005	1	Y (luma)	
FSIM [78]	FR	spatial	2011	3	grayscale	
FSIMc [78]	FR	spatial	2011	3	RGB	
G-SSIM [79]	FR	spatial (gradients)	2012	2	grayscale	
GMS + CMS [78]	FR	spatial	2011	3	RGB	
GMSD [65]	FR	spatial (gradients)	2014	3	grayscale	
IW-SSIM [80]	FR	spatial	2011	4	grayscale	
L-SSIM [61]	FR	spatial	2004	—	grayscale	
LPIPS-Alex [81]	FR	deep features	2018	4	RGB	
LPIPS-Squeeze [81]	FR	deep features	2018	4	RGB	
LPIPS-VGG [81]	FR	deep features	2018	4	RGB	
MAE [72]	FR	spatial	N/A	—	grayscale	
PSNR [72]	FR	spatial	N/A	—	grayscale	
PSNR-B [82]	FR	spatial	2011	2	grayscale	
SNR [72]	FR	spatial	N/A	—	grayscale	
SAM [64]	FR	spectral	1993	1	multi-band	
SCC [70]	FR	spectral	1998	1	multi-band	
R-SSIM [68]	FR	spatial	2012	3	RGB	

(continuation)

Metric	Type	Domain	Year	Tested Databases	Spectrum	
RMSE [72]		FR	spatial	N/A	—	grays
PSM [69]		FR	spatial	2002	1	grays
MS-FSIM [78]		FR	spatial	2011	3	grays
MS-SSIM [83]		FR	spatial	2003	1	grays
MSE [72]		FR	spatial	N/A	—	grays
SPIQ [84]		NR	deep features	2022	3	RGB
SR-SIM [85]		FR	spatial	2012	3	RGB
SSIM [61]		FR	spatial	2004	1	grays
UQI [86]		FR	spatial	2002	1	grays
VIF [87]		FR	wavelet	2006	1	Y (lu
VIFc [87]		FR	wavelet	2006	1	Y/Cb
VIFp [88]		FR	pixel-domain	2005	1	Y (lu
VSNR [67]		FR	wavelet	2007	1	Y (lu
WaDIQaM [89]		FR	deep features	2018	3	RGB
W-SSIM [90]		FR	spatial	2011	1	grays
D_λ (QNR) [71]		FR	spectral/spatial	2008	1	multi

2.2.1 Full-Reference IQA

FR-IQA models estimate the perceptual quality of a distorted image by comparing it to an undistorted reference. These methods assume complete access to the original image and are commonly used in contexts where ground-truth data is available, such as image compression, transmission, enhancement, and restoration [54], [58]. By directly quantifying deviations from the reference, FR-IQA provides a consistent and interpretable benchmark for evaluating distortion. However, their applicability is limited to scenarios where high-quality reference images exist, making them unsuitable for many real-world applications [91], [92].

Fidelity-Based Metrics

Fidelity-based metrics quantify distortion by measuring direct numerical differences between the reference and the distorted image. These models operate in the spatial domain and are widely used due to their simplicity, computational efficiency, and clear interpretability [57], [93]. The most common examples include the mean squared error (MSE), root mean squared error (RMSE), and peak signal-to-noise ratio (PSNR). Although these metrics are sensitive to pixel-level changes, they often fail to capture perceptually relevant distortions, especially when the structure or semantic content of the image is preserved [58], [67]. Extensions such as PSNR-B address some of these limitations by incorporating blocking artifact penalties [94], while other domain-specific metrics, such as ERGAS [95] and SAM [96], are used in remote sensing to assess multiband and hyperspectral image fidelity.

The Mean Squared Error (MSE) measures the average squared pixel-wise difference:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} x_i - y_i{}^2 \tag{2.3}$$

The Root Mean Squared Error (RMSE) is the square root of MSE:

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{2.4}$$

The Mean Absolute Error (MAE) computes the average of absolute differences:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |x_i - y_i| \tag{2.5}$$

The Peak Signal-to-Noise Ratio (PSNR) expresses the logarithmic ratio between the maximum possible signal value and the MSE between a reference and a distorted image:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX^2}{\text{MSE}}\right) \tag{2.6}$$

where MAX is the maximum possible pixel value (typically 255 for 8-bit images). Although widely used for its simplicity and interpretability, PSNR correlates poorly with human perception, especially in the presence of structured distortions.

To address this limitation, PSNR-B extends the original formulation by incorporating a blocking effect factor (BEF), which penalizes compression artifacts such as those introduced by block-based codecs:

$$\text{PSNR-B} = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE} + \text{BEF}} \right) \tag{2.7}$$

The Signal-to-Noise Ratio (SNR) compares the signal power to the noise power:

$$\text{SNR} = 10 \cdot \log_{10} \left(\frac{\sum x_i^2}{\sum (x_i - y_i)^2} \right) \tag{2.8}$$

The Spectral Angle Mapper (SAM), often used in hyperspectral imagery, computes the angle between image vectors:

$$\text{SAM} = \cos^{-1} \left(\frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \right) \tag{2.9}$$

The ERGAS (Erreur Relative Globale Adimensionnelle de Synthèse) metric expresses relative global error and is frequently used in image fusion and remote sensing:

$$\text{ERGAS} = 100 \cdot \frac{h}{l} \cdot \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\frac{\text{RMSE}_i}{\mu_i} \right)^2} \tag{2.10}$$

where h and l are the spatial resolutions of the high and low resolution images respectively, RMSE_i is the RMSE of band i, and μ_i is the mean of the reference band.

These metrics are particularly useful when high pixel fidelity is essential, but they often fail to align with human perception, especially in cases where structural or semantic information is preserved.

Perceptual Similarity Metrics

Perceptual similarity metrics aim to model the characteristics of the human visual system (HVS), focusing not on raw signal differences but on how distortions affect perceived quality. These models evaluate structural, luminance, and contrast relationships within local regions of the image and are typically more aligned with subjective scores than purely fidelity-based metrics [58], [67]. The Structural Similarity Index (SSIM) is a foundational method in this category, comparing local windows between the reference and distorted images. Multiscale variants such as MS-SSIM [83] improve robustness across spatial resolutions. Other models, such as FSIM [78],

incorporate phase congruency and gradient information to better reflect visual saliency, while GMSD [65] measures image quality by analyzing the standard deviation of gradient magnitude similarity. Additional extensions include SR-SIM [85], which prioritizes salient regions, and W-SSIM [90], which introduces perceptual weighting across scales. These methods offer improved correlation with human opinion scores, particularly in the presence of structural or perceptual distortions.

The Structural Similarity Index (SSIM) compares local image patches between the reference and distorted images, measuring similarity in luminance, contrast, and structure. It is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.11}$$

where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the local variances, σ_{xy} is the local covariance, and C_1, C_2 are stabilizing constants.

Multiscale SSIM (MS-SSIM) extends SSIM by computing similarity at multiple image scales. It is defined as:

$$\text{MS-SSIM}(x, y) = \prod_{j=1}^{M} [l_j(x, y)]^{\alpha_j} \cdot [c_j(x, y)]^{\beta_j} \cdot [s_j(x, y)]^{\gamma_j} \tag{2.12}$$

where l_j, c_j, and s_j are the luminance, contrast, and structure comparisons at scale j, and α_j, β_j, γ_j are weights.

The Feature Similarity Index (FSIM) combines phase congruency (PC) and gradient magnitude (GM) to evaluate perceptual quality:

$$\text{FSIM}(x, y) = \frac{\sum_{i\in\Omega} T(i) \cdot PC_m(i)}{\sum_{i\in\Omega} PC_m(i)} \tag{2.13}$$

where $T(i)$ is a similarity function combining gradient and phase congruency similarity at pixel i, and $PC_m(i)$ is the maximum phase congruency across both images.

The Gradient Magnitude Similarity Deviation (GMSD) is based on the standard deviation of pixel-wise gradient similarity maps:

$$\text{GMS}(i) = \frac{2G_x(i)G_y(i) + C}{G_x(i)^2 + G_y(i)^2 + C} \tag{2.14}$$

$$\text{GMSD}(x, y) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\text{GMS}(i) - \overline{\text{GMS}}\right)^2} \tag{2.15}$$

where $G_x(i)$ and $G_y(i)$ are gradient magnitudes of the reference and distorted images, and C is a small constant.

The Spectral Residual Similarity Index (SR-SIM) evaluates similarity by combining spectral residual saliency and gradient magnitude:

$$\text{SR-SIM}(x, y) = \frac{\sum_{i \in \Omega} S_r(i) \cdot G_m(i) \cdot Q(i)}{\sum_{i \in \Omega} S_r(i) \cdot G_m(i)} \tag{2.16}$$

where $S_r(i)$ is the spectral residual saliency map, $G_m(i)$ is the gradient magnitude, and $Q(i)$ is the local structural similarity.

The Information-Weighted Structural Similarity Index (IW-SSIM) assigns greater importance to image regions that carry more structural information. The final score is a weighted average of SSIM values computed at each local window:

$$\text{IW-SSIM}(x, y) = \frac{\sum_{i=1}^{N} w_i \cdot \text{SSIM}_i(x, y)}{\sum_{i=1}^{N} w_i} \tag{2.17}$$

where w_i is the information weight for window i, often computed using local entropy or local variance.

The Wavelet-based SSIM (W-SSIM) extends SSIM into the complex wavelet domain. Instead of operating directly on image intensities, it evaluates phase consistency across wavelet subbands. The general form follows a similar structure to SSIM, applied to wavelet coefficients:

$$\text{W-SSIM}(x, y) = \prod_{s} \prod_{o} \text{SSIM}(x_{s,o}, y_{s,o}) \tag{2.18}$$

where s indexes the scale and o the orientation in the wavelet decomposition, and $x_{s,o}$ and $y_{s,o}$ are the corresponding coefficients.

The Edge Strength Similarity Metric (ESSIM or ESIM) modifies SSIM by replacing the structure component with a comparison of edge magnitudes. The edge strength is typically extracted via a Sobel or Prewitt operator. The structure term σ_{xy} is replaced by a directional edge similarity $E(x, y)$, such that:

$$\text{ESSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2E(x, y) + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(E(x, x) + E(y, y) + C_2)} \tag{2.19}$$

where $E(x, y)$ denotes the dot product of edge maps from the two images.

While perceptual similarity metrics approximate human visual perception by analyzing structural and saliency-based cues, they remain largely deterministic and handcrafted.

Information-theoretic Metrics

Information-theoretic approaches offer a fundamentally different perspective, treating image quality assessment as a problem of quantifying the amount of perceptually relevant information

preserved in the distorted image. These models are rooted in natural scene statistics and signal fidelity theory, where both the reference and distorted images are interpreted as stochastic signals [62]. The Information Fidelity Criterion (IFC) and the Visual Information Fidelity (VIF) index model the image source and distortion process probabilistically, estimating mutual information between the reference and distorted signals as mediated by models of the human visual system [87]. This framework enables a content-aware assessment of quality that accounts for the statistical dependencies and perceptual importance of image structures.

The IFC models image patches in the wavelet domain using Gaussian scale mixture (GSM) models. It estimates the mutual information between the reference image x and the distorted image y conditioned on the distortion model and natural scene statistics:

$$\mathrm{IFC}(x, y) = \sum_{i \in \Omega} I(C_{x,i}; C_{y,i} \mid z_i) \tag{2.20}$$

where $C_{x,i}$ and $C_{y,i}$ are wavelet coefficients of the reference and distorted images in patch i, and z_i is a local variance parameter estimated under a GSM model. The total information is accumulated over all image patches Ω.

The VIF index extends this by incorporating a model of the human visual system. It computes the ratio of information that can be extracted by a hypothetical observer from the distorted image versus the reference:

$$\mathrm{VIF}(x, y) = \frac{\sum_{i=1}^{N} I(E_i; F_i)}{\sum_{i=1}^{N} I(E_i; G_i)} \tag{2.21}$$

Here, E_i represents the coefficients of the reference image, F_i those of the distorted image after HVS filtering, and G_i those of the reference after the same filtering. The numerator measures the information preserved in the distorted image, and the denominator the total information in the reference. Mutual information is computed under assumed Gaussian models of natural scenes and additive noise for distortion and perceptual masking.

Several variants of the VIF metric have been proposed to adapt the model to different domains and signal characteristics. The pixel-domain variant, VIFp [88], simplifies the original VIF model by operating directly on pixel intensities rather than wavelet coefficients. This avoids the need for multi-resolution decomposition and is computationally more efficient, making it suitable for real-time or resource-constrained applications. VIFp retains the same information fidelity framework, estimating mutual information between corresponding local patches.

Another extension is VIFc, which extends the VIF model to handle color images [87]. Instead of converting the image to grayscale, VIFc applies the information fidelity model to each color channel, often in the YCbCr color space. A weighted fusion of channel-wise VIF scores is then

used to produce a final quality estimate:

$$\text{VIFc}(x, y) = \sum_{k \in \{Y, Cb, Cr\}} w_k \cdot \text{VIF}_k(x, y) \tag{2.22}$$

where w_k are empirically determined weights that reflect the perceptual importance of each component. Typically, the luminance channel (Y) receives the highest weight due to its dominant role in human perception.

These variants preserve the conceptual integrity of the original VIF model while offering improved flexibility for different input modalities and computational budgets.

Deep Learning-based Metrics

Learning-based full-reference IQA models leverage deep neural networks to capture perceptual similarity in a data-driven manner. Unlike traditional models that rely on handcrafted features or analytic formulations of human perception, these methods operate in learned feature spaces derived from large-scale image datasets. The central assumption is that perceptual similarity can be approximated by comparing intermediate activations of pretrained convolutional neural networks (CNNs) [81]. The Learned Perceptual Image Patch Similarity (LPIPS) metric exemplifies this approach by computing a weighted L_2 distance between feature maps extracted from networks such as VGG or AlexNet. Subsequent models, such as DISTS [60], refine this idea by balancing structural and texture similarity through adaptive weighting schemes. Other approaches, like WaDIQaM [89], train task-specific regressors over deep features extracted from both the reference and distorted images. These models achieve high correlation with human opinion scores, particularly in cases involving complex or perceptually subtle distortions, but often require careful normalization and calibration of feature space distances.

The LPIPS metric computes the distance between deep feature maps of a reference image x and a distorted image y as follows:

$$\text{LPIPS}(x, y) = \sum_{l} \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \| w_l \odot (\hat{f}_l^x(h, w) - \hat{f}_l^y(h, w)) \|_2^2 \tag{2.23}$$

where $\hat{f}_l^x$ and $\hat{f}_l^y$ are the normalized feature maps at layer l for images x and y, respectively, with spatial dimensions $H_l \times W_l$. The weights w_l are learned scalars for each channel, and $\odot$ denotes element-wise multiplication.

The DISTS metric combines feature similarity and texture similarity between corresponding feature maps:

$$\text{DISTS}(x, y) = \sum_{l} \alpha_l \cdot S_l(x, y) + \beta_l \cdot T_l(x, y) \tag{2.24}$$

where $S_l(x, y)$ is the structure similarity component computed as the cosine similarity between deep features at layer l, and $T_l(x, y)$ is the texture similarity component, often based on the mean and variance of feature activations. The coefficients α_l and β_l are learned to balance structure and texture contributions for each layer l.

WaDIQaM-FR learns a patch-based quality prediction model using paired deep features from the reference image x and the distorted image y. Given a patch pair (x_i, y_i), deep features are extracted using a shared-weight CNN encoder, and the resulting representations are concatenated and passed through a regression network to predict a local quality score q_i. The final quality score is the weighted average over all patches:

$$Q(x, y) = \sum_{i=1}^{N} w_i \cdot q_i \tag{2.25}$$

where $q_i = f_\theta(x_i, y_i)$ is the predicted quality for patch i, w_i is a learned spatial weight indicating the perceptual relevance of patch i, and N is the total number of patches. The function f_θ denotes the trained regression model with parameters θ.

During training, the network minimizes the error between predicted quality scores and human-annotated MOS or DMOS labels using an appropriate loss function (e.g., MSE or Huber loss). Unlike LPIPS or DISTS, WaDIQaM does not rely on pretrained networks, allowing it to adapt feature representations to the task of quality prediction.

In summary, full-reference IQA models span a spectrum from simple error-based metrics to perceptually and statistically grounded approaches. Fidelity-based methods offer computational efficiency but lack alignment with human perception. Perceptual similarity models incorporate structural and saliency cues to improve correlation with subjective quality. Information-theoretic metrics formalize quality as mutual information preservation under natural scene statistics. Learning-based models leverage deep features or end-to-end training to approximate perceptual similarity in high-dimensional spaces. Each class reflects a trade-off between interpretability, perceptual fidelity, and computational complexity, making their suitability highly context-dependent.

2.2.2 No-Reference IQA

Traditional NR-IQA

Among traditional NR-IQA techniques, PIQE [97] (Perception-based Image Quality Evaluator) and NIQE [98] (Natural Image Quality Evaluator) are widely used due to their computational simplicity. PIQE begins by dividing the input image into non-overlapping 16×16 blocks, denoted B_i, and identifying distorted blocks based on edge content and local variance thresholds. For each block, it computes perceptual features using the gradient magnitude:

$$S_i = \frac{1}{|B_i|} \sum_{(x,y) \in B_i} \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}, \tag{2.26}$$

and noise, estimated by subtracting a low-pass filtered (LPF) version of the block:

$$N_i = \text{std}\left(B_i - \text{LPF}(B_i)\right). \tag{2.27}$$

The global PIQE score is computed via a weighted aggregation over the distorted blocks, $\mathcal{D}$:

$$\text{PIQE} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} w_i \cdot (\alpha S_i + \beta N_i + \gamma C_i), \tag{2.28}$$

where w_i are confidence weights and α, β, γ are scaling factors for sharpness, noise, and contrast, respectively. The final score is scaled between 0 (best quality) and 100 (worst quality).

In contrast, NIQE is grounded in natural scene statistics (NSS), using a statistical deviation model to quantify perceptual degradation. First, it applies Mean Subtracted Contrast Normalization (MSCN) to locally standardize luminance:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + \epsilon}, \tag{2.29}$$

from which it extracts statistical features associated with Generalized Gaussian Distributions (GGD) and Asymmetric GGD (AGGD). A multivariate Gaussian (MVG) model with mean μ_r and covariance Σ_r is estimated to a set of pristine images. During inference, the Mahalanobis distance between the test image features $\mathbf{f}_t$ and the natural image model is used to compute the NIQE score:

$$\text{NIQE}(\mathbf{f}_t) = \sqrt{(\mathbf{f}_t - \mu_r)^T \Sigma_r^{-1} (\mathbf{f}_t - \mu_r)}. \tag{2.30}$$

This formulation enables NIQE to operate in a completely opinion-unaware and unsupervised fashion, relying solely on deviations from natural image statistics to infer perceptual quality.

Facial Image Quality Assessment

FIQA refers to the estimation of biometric utility from face images, typically in the context of face recognition, verification, or detection pipelines. Unlike generic NR-IQA, FIQA does

not aim to measure aesthetic or perceptual quality, but rather the suitability of a face image for recognition purposes [99], [100]. This quality is influenced by a range of factors, including resolution, pose, occlusion, blur, compression, and illumination.

Traditional FIQA methods relied on handcrafted features such as sharpness, symmetry, or inter-eye distance [101]. Recent models leverage deep face embeddings extracted from recognition networks [46], [102]. These models typically apply a regression or ranking loss to map features to quality scores correlated with recognition accuracy.

FIQA is an inherently task-dependent quality problem. Unlike general-purpose IQA, it must account for the characteristics of the recognition model and dataset. Furthermore, FIQA methods often correlate poorly with perceptual quality metrics, as an image may appear visually high-quality but remain unsuitable for matching due to pose or occlusion. This motivates the need to explore perceptually grounded IQA models tailored to face data, and to investigate how traditional metrics can be adapted, fused, or compared against biometric-specific quality indicators.

In contrast to traditional statistical NR-IQA methods, recent approaches such as SER-FIQ [46] (Stochastic Embedding Robustness for Face Image Quality) and MagFace [45] adopt deep learning-based strategies tailored specifically for facial image quality. SER-FIQ builds on the intuition that high-quality facial images produce consistent embeddings under stochastic dropout perturbations, while low-quality images result in more variable and unstable representations. Given a face image, a pre-trained face recognition model (e.g., ArcFace) with dropout enabled generates T stochastic embeddings $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T\} \in \mathbb{R}^d$. The pairwise cosine similarities between these embeddings are then computed:

$$\text{sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}, \quad \forall i, j \in \{1, \ldots, T\}, i < j. \tag{2.31}$$

The final SER-FIQ score is derived from the mean or standard deviation of the similarity matrix, often defined as:

$$\text{SER-FIQ} = \frac{1}{\binom{T}{2}} \sum_{i<j} \text{sim}(\mathbf{e}_i, \mathbf{e}_j), \tag{2.32}$$

where higher scores indicate more stable (and thus higher-quality) facial representations. This method is fully unsupervised and exploits intrinsic model uncertainty as an indicator for quality.

MagFace, on the other hand, is a supervised approach that explicitly learns a quality-aware embedding space. It introduces a margin-adaptive loss that couples facial feature magnitude with image quality. For an embedding $\mathbf{e} \in \mathbb{R}^d$, its L_2-norm $\|\mathbf{e}\|$ is used as a measure for quality. The classification loss is modified as:

$$\mathcal{L}_{\text{MagFace}} = -\log \frac{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))}}{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))} + \sum_{j \neq y} e^{s \cdot \cos(\theta_j)}}, \tag{2.33}$$

where θ_y is the angle between $\mathbf{e}$ and the weight vector of the correct class, s is a scaling factor, and $m(\|\mathbf{e}\|)$ is a dynamic margin that increases with quality. The final MagFace quality score is proportional to the magnitude $\|\mathbf{e}\|$, constrained during training to lie within a fixed interval $[l, u]$ for regularization. Unlike SER-FIQ, MagFace learns to embed quality-awareness directly into the feature space through supervision, enabling both verification and quality prediction in a unified framework.

2.2.3 Fusion-Based IQA Models

To address the perceptual limitations of individual image quality metrics, several fusion-based approaches have been proposed. Liu et al. [59] introduced a multi-method fusion (MMF) framework that linearly combines full-reference IQA scores via supervised regression to better approximate MOS. Henniger et al. [43] developed a Random Forest model trained on handcrafted features derived from ISO-compliant facial images, improving performance in biometric settings. These works demonstrated that integrating complementary quality cues can significantly enhance perceptual alignment, outperforming standalone metrics in correlation with human judgment [103].

In parallel, the scarcity of large-scale human-labeled datasets has motivated weakly supervised learning strategies. Chen et al. [51] generated pseudo-MOS labels by averaging outputs of multiple FR-IQA metrics. RankIQA [47] applied relative ranking supervision on synthetically degraded images to learn ordinal quality representations. Wu et al. [104] used cascaded convolutional regressors trained on pseudo-MOS labels, demonstrating that pseudo-supervision can bootstrap learning when subjective ground truth is limited.

These developments support the feasibility of learning perceptual quality through fusion, either via model ensembles or by generating training supervision. Our approach builds on this principle: we construct a regression-based fusion model that aggregates multiple FR-IQA scores to predict perceptual quality, using human-labeled MOS as supervision. This model is then applied to a larger unlabeled dataset to generate pseudo-MOS scores, which in turn are used to train a no-reference regressor. This hybrid full-to-no-reference framework enables scalable, perceptually grounded quality estimation even in the absence of pristine reference images.

2.3 Steganography

Steganography has evolved significantly with the rise of digital media and advanced compression algorithms, leading to a variety of embedding methods and detection strategies. Classical steganographic methods can be broadly categorized into spatial, frequency, and adaptive

domain techniques. Spatial domain methods directly manipulate pixel values, most notably through Least Significant Bit (LSB) substitution [105], to encode binary data within an image. Frequency domain methods operate on transformed representations of the image, using tools such as the Discrete Cosine Transform (DCT) or Discrete Wavelet Transform (DWT), embedding data in coefficients less sensitive to compression [106]. Adaptive methods further refine this by analyzing local characteristics of the image, dynamically choosing embedding regions to minimize perceptual distortion [107].

However current methods cover a wide range of applications, there has been a growing interest in printer-proof steganography, which survives the printing and scanning process. This is particularly relevant in contexts where physical copies of images are distributed, such as MRTDs. The challenge lies in ensuring that the embedded information remains detectable and robust against distortions introduced by printing and scanning processes.

2.3.1 Printer-proof Steganography

Recent advances employ deep learning to design robust, end-to-end steganographic pipelines that simulate print-scan degradations during training. We highlight four such methods that represent the current state of the art in printer-proof steganography for natural and facial images. The printer-proof steganography methods used were based on Generative Adversarial Networks (GANs) [108] to encode and decode information, we can obtain various results depending on the method used.

StegaStamp

StegaStamp [42] introduced the first deep learning-based pipeline for robust steganography under real-world distortions. It jointly trains an encoder-decoder architecture with simulated perturbations, such as blur, color shift, projective warps, and JPEG compression, to mimic the print-scan process. A random bitstring is embedded into the image, and the decoder learns to retrieve it despite these degradations. The training loss combines a cross-entropy bit loss, an L_2 reconstruction loss, and a perceptual LPIPS loss to preserve visual quality. StegaStamp showed that it is possible to reliably recover short messages (e.g., 56 bits) from physically printed and scanned images, setting a benchmark for printer-proof robustness.

CodeFace

CodeFace [41] extends StegaStamp with a focus on facial images used in ID documents. It introduces facial-specific modules such as a Spatial Transformer Network (STN) for geometric

alignment and a perceptual identity-preserving loss based on FaceNet embeddings. This ensures that the stego face remains visually and biometrically consistent with the original. The network is trained adversarially, with a discriminator encouraging naturalness and a decoder optimizing message recovery under print-scan distortions. CodeFace proves effective for embedding messages in high-resolution facial portraits, making it suitable for privacy-preserving biometric IDs.

RiemStega

RiemStega [109] innovates by introducing a Riemannian manifold-based loss. Instead of comparing pixel values, it represents images through symmetric positive definite (SPD) covariance matrices and minimizes the affine-invariant Riemannian distance between the cover and stego image descriptors. This statistical approach preserves global structural and texture patterns more robustly under degradation. The method maintains high visual fidelity while increasing message recoverability post-printing, and demonstrates strong performance on both generic and facial image datasets.

StampOne

StampOne [110] incorporates frequency-domain awareness by applying DWT and gradient maps to both image and message during encoding. The model balances frequency bands via a learned attention mechanism and includes frequency-domain discriminators in its loss functions. This results in encoded images with spectral properties closely matching those of the original, enhancing resilience to printer-related distortions such as color shifts and sensor noise. StampOne outperforms prior methods in both visual similarity and decoding accuracy, especially for high-frequency facial features.

3 Methodology

3.1 Dataset

Our dataset is derived from the publically available Face Research Lab London [111] (FRLL) set, comprising 102, 1350 × 1350, full coloured images of frontal, ICAO-compliant adult faces. Self-reported age, gender and ethnicity are included. Attractiveness ratings (on a 1–7 scale from "much less attractiveness than average" to "much more attractive than average") for the faces from 2513 people (ages 17–90) are included as well. The FRLL dataset was selected for its high-fidelity ICAO-compliant images, controlled acquisition conditions, and extensive demographic annotations, which are crucial for analyzing perceptual quality and potential demographic biases in subjective ratings. Each image was encoded using four printer-proof steganographic methods, each applied at nine different intensity levels, $x \mid x = 0.6 + 0.1 \times k, k \in \{0, \ldots, 8\}$, yielding a total of 3,672 distorted images.

The dataset was partitioned into four subsets, as follows:

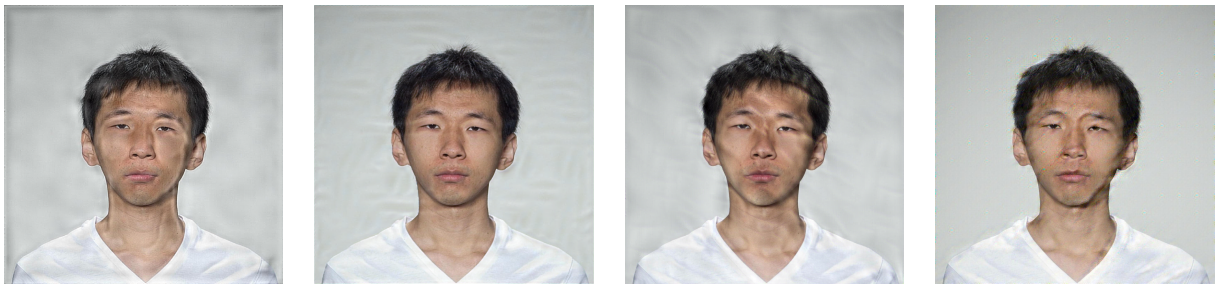
- MOS set (15 identities, 540 images): a core set of demographically diverse subjects, shown in Fig. **??**, with subjective MOS annotations. It is split into:
  - MOS train set (12 identities): used to train the FR fusion model, regressing FR-IQA metrics to human MOS.
  - MOS test set (3 identities): 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 steganographic distortions are visible in Fig. **??**, which shows examples of distorted facial images from each method. The distortions are applied to the original images, and the re-

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

Figure 3.1: Reference images from the MOS set, comprising the selected subjects from the FRLL dataset.

sulting stego images are used for both subjective evaluation and training of the NR-IQA model. The selected steganographic methods represent diverse classes of distortion processes (geometric, color-based, local texture perturbation) commonly encountered in practical steganographic applications. The chosen intensity levels ensure a broad perceptual range from barely noticeable to clearly visible distortions.



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

Figure 3.2: Steganographically distorted facial images from each method.

3.2 Collection of Subjective Scores

We followed the ITU-R BT.500–15 [49] recommendation and adopted the Single Stimulus (SS) method. The test was implemented using a custom Django web application seen in Fig. **??**. 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. Table **??** summarizes the demographic profile of the participant pool. 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 [49], resulting in the exclusion of one participant.

The database was implemented using the Django web framework, which provides an object-relational mapping (ORM) layer that directly maps Python model classes to database tables. The ORM also enforces data integrity constraints and simplifies querying for statistical analysis. The database backend used was PostgreSQL.

The database schema was designed to ensure structured storage and traceability of all subjective quality evaluation data. Fig. **??** presents the conceptual data model, which defines the main entities and their relationships. The profile entity stores participant metadata, including age, gender, education, ethnicity, country of origin, and device used. The ss_session entity models individual test sessions and is linked to both the participant profile and the dataset being evaluated. The image entity encodes each image's filename, associated distortion type (distortion_name), and distortion level (distortion_level). The test entity stores individual subjective ratings, including the precise timestamp of submission, linked to both the session and the image presented. An auxiliary user_feedback entity captures optional, anonymous, observer comments and critiques.

The physical implementation of this schema, shown in Fig. **??**, is realized as a relational database with explicit foreign key constraints to ensure referential integrity. The dataset_id and profile_id fields in ss_session formally enforce the linkage to specific datasets and participants. The test table connects each subjective rating to its session (ss_session_id) and image (image_id), enabling consistent aggregation of ratings into a statistical descriptors such as MOS and confidence intervals per image and per distortion level. The relational model also

Figure 3.3: Django-based webapp created for the Single Stimulus test.

Table 3.1: Demographic and non-demographic profile of the observers.

(a) Age group

Group	Count
18–25	85
26–40	54
41–60	52
Above 60	9
Under 18	2

(b) Gender

Gender	Count
Male	143
Female	59

(c) Ethnicity

Group	Count
White	177
Latino	8
Black	6
Mixed / Multiple	3
Others	7

(d) Education

Level	Count
Bachelor's	92
Master's	63
Doctorate	29
High School	14
Other	4

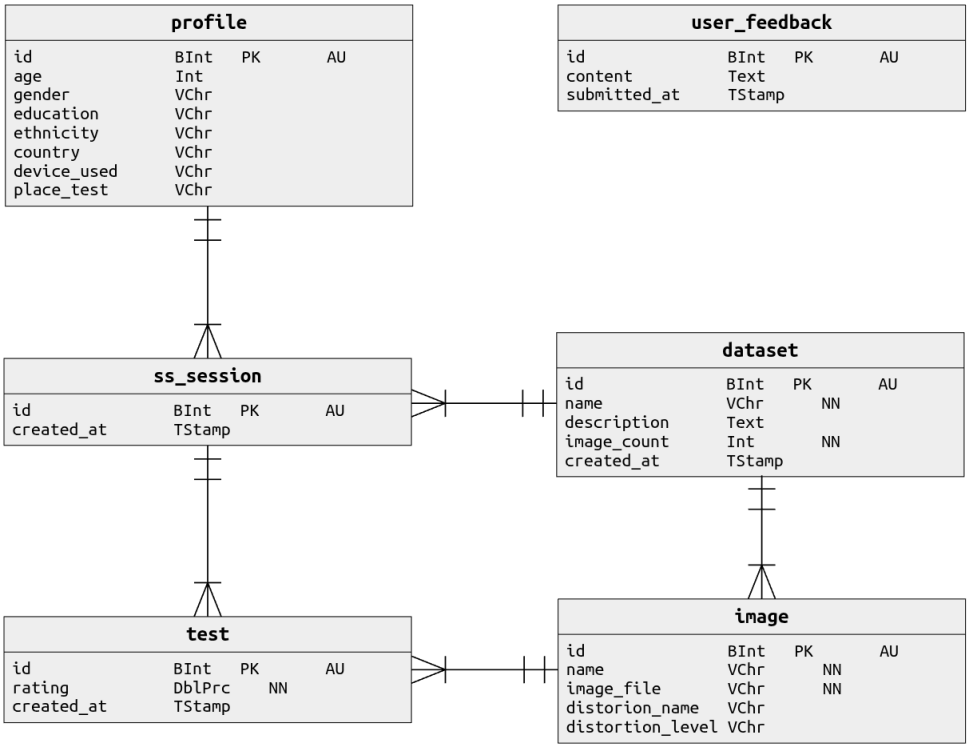
(e) Country

Country	Count
Portugal	170
Brazil	16
Russia	3
India	2
Ghana	2
Others	9

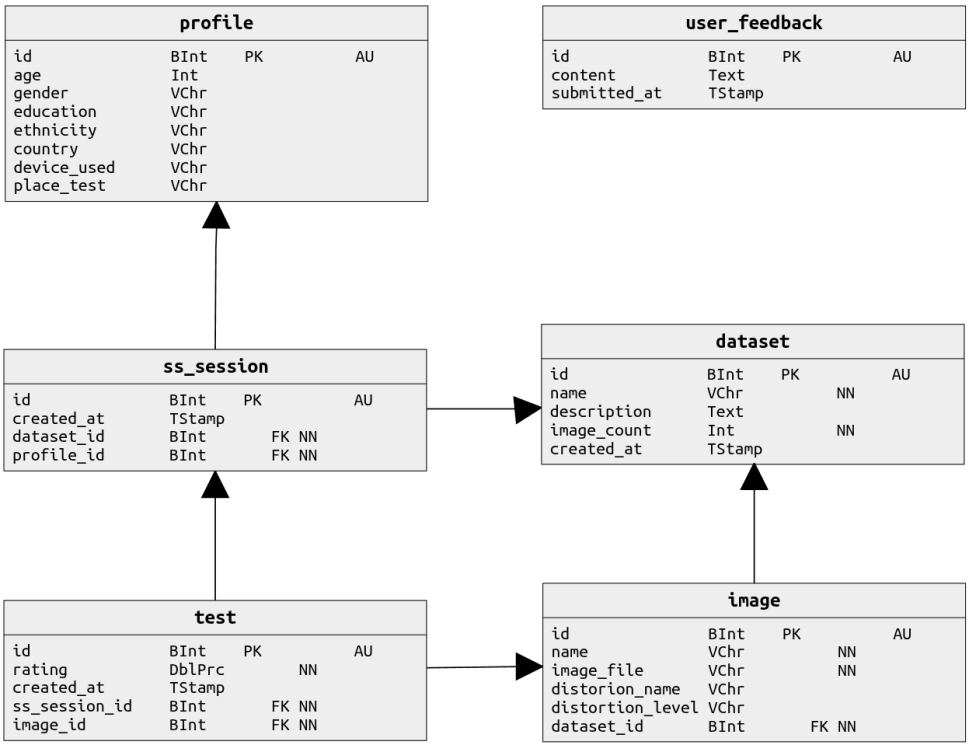
(f) Device Used

Device	Count
Laptop	94
Phone	65
Desktop	42
Other	1

facilitates efficient querying for downstream statistical analysis and supports reproducibility of the experimental protocol in line with ITU-R BT.500-15 recommendations.



(a) Conceptual database schema showing entity relationships and high-level structure.



(b) Physical database schema illustrating actual tables, columns, and implementation details.

Figure 3.4: Conceptual and physical representations of the database schema.

3.3 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 [112], 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 **??**, 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.

Table 3.2: ANOVA [113] 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

3.4 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. **??**. Several metrics 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 [114].

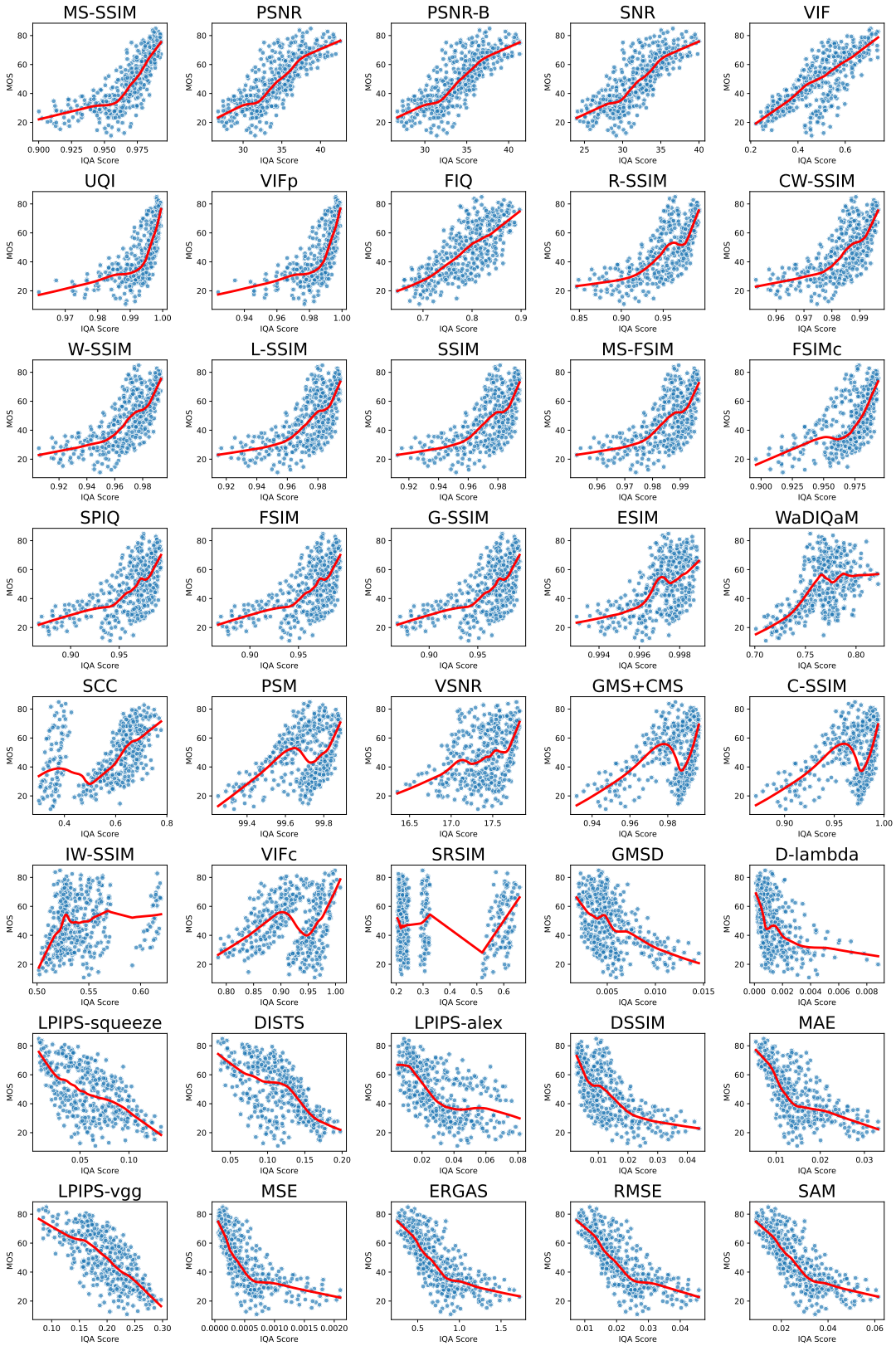


Figure 3.5: Scatter plots illustrating the relationship between MOS and 40 individual full-reference IQA metrics. The red line represents a smoothed local trend curve.

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

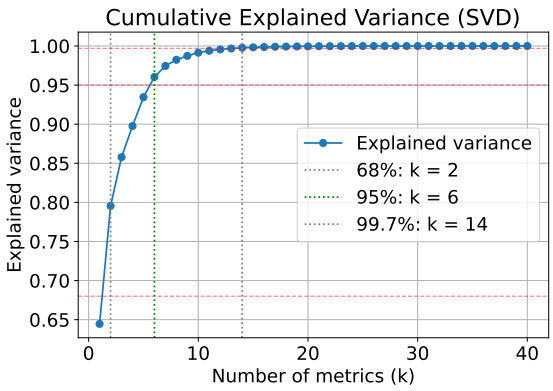
To identify which metrics align best with human perception, we compute both the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC) [115] which respectively quantify the linearity and monotonicity of the relationship between metric scores and MOS. To determine the appropriate number of metrics to retain for fusion, we applied Singular Value Decomposition (SVD) and the Picard criterion [116]. 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. **??**a, indicating an optimal dimensionality of $k = 6$.

To validate this truncation point, we examined the Picard plot in Fig. **??**b, 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.

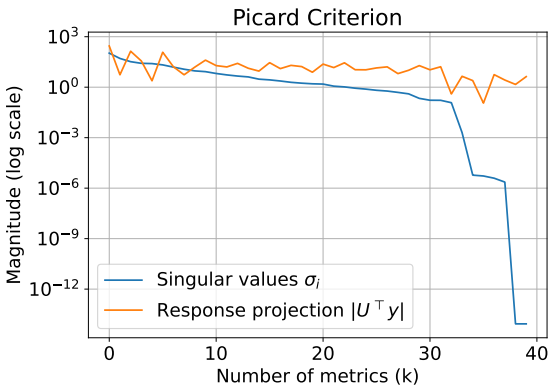
After selecting the top $k = 6$ FR-IQA metrics, we trained a diverse set of supervised regressors to map these features to subjective MOS scores. The models included both linear and non-linear types:

- Linear models: Linear Regression [117], Ridge Regression [118], Bayesian Ridge [119], ElasticNet [120].
- Kernel-based: Support Vector Regression [121] (SVR).
- Ensembles: Random Forest [122], Extra Trees [123], Gradient Boosting [124], HistGradientBoosting [125].
- Boosted Trees: XGBoost[126], LightGBM [127], CatBoost [128].

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



(a) Cumulative variance explained by SVD components.



(b) Magnitude of singular values and projected response.

Figure 3.6: 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 $k = 6$.

3.6 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 [129] pretrained on ImageNet [130], with its final classification layer replaced by a lightweight multi-layer perceptron (MLP) regressor [119]. 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.

3.7 Results

3.7.1 Full-Reference Fusion Metric Performance

To interpret feature contributions, we used SHAP [131] (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. **??**, MS-SSIM [132] (Multiscale SSIM) and UQI [133] (Universal Quality Index) had the largest positive impact on predictions, while PSNR-B [56], VIF[54], (PSNR-Blocking), PSNR [55] and SNR [57] (Signal-to-Noise Ratio) contributed with moderate influence. The color encoding further highlights how high and low

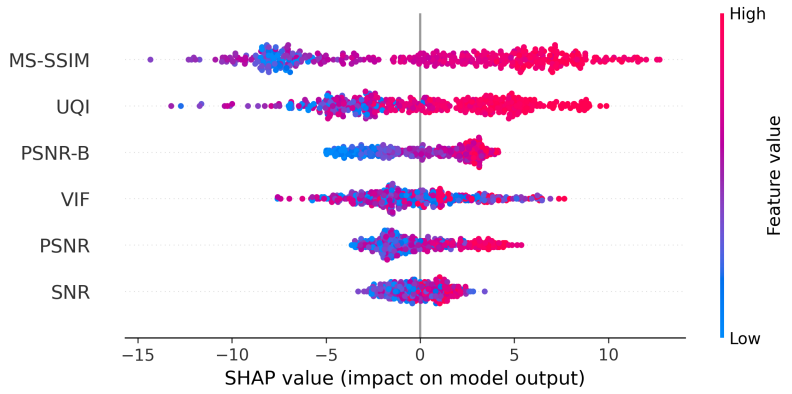


Figure 3.7: SHAP summary plot for $k = 6$, generated from the best CatBoost model.

Table 3.3: 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

feature values affect the model output.

When comparing the fusion metric with the individual FR metrics, Table **??** 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.

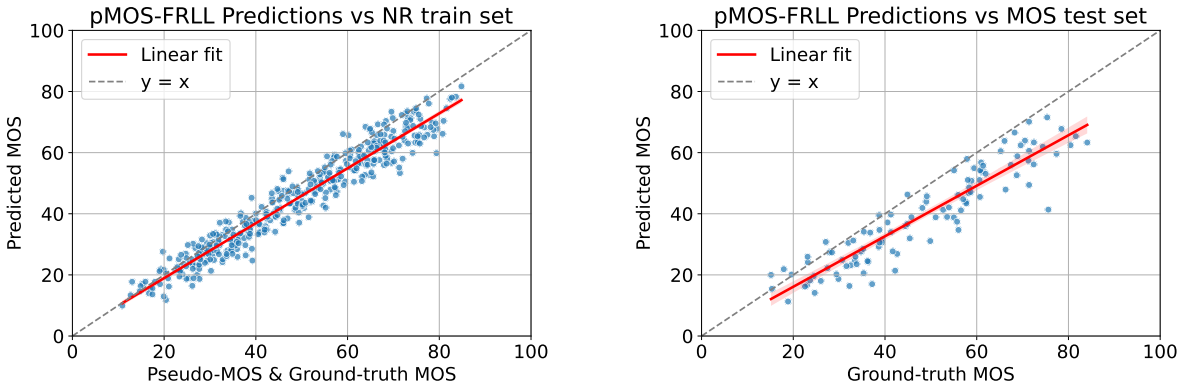
3.7.2 No-Reference Regression Model Performance

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. **??** shows two evaluation scenarios: in Fig. **??**(a), the model's predictions are compared against the pseudo and ground-truth MOS

Table 3.4: 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

used during training, yielding a PLCC of 0.9285, SRCC of 0.9345, MSE of 120.2, and MAE of 9.158. Fig. **??**(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.

Figure 3.8: 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 [98] and PIQE [97], as well as task-specific approaches like SER-FIQ [46] and MagFace [45]. This highlights the robustness of our weakly supervised strategy for NR-IQA on steganographically degraded facial images. Quantitative results are reported in Table **??**.

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.

3.8 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.

Bibliography

[1] Z. Wang and A. C. Bovik, *Modern image quality assessment* (Synthesis Lectures on Image, Video, and Multimedia Processing 1). San Rafael, CA, USA: Morgan & Claypool, 2006, vol. 2, pp. 1–156.

[2] K.-H. Thung and P. Raveendran, "A survey of image quality measures," *IEEE Transactions on Image Processing*, pp. 1–15, 2021.

[3] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140 030–140 043, 2019. [Online]. Available: `https://doi.org/10.1109/ACCESS.2019.2943319`.

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

[5] R. Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR*, pp. 586–595, 2018.

[6] A. Ciancio, A. L. N. T. Costa, E. A. B. Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[7] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 1, Rochester, NY, USA, Sep. 2002, I:477–I:480.

[8] U. Engelke, H.-J. Zepernick, and T. M. Kusuma, "Subjective quality assessment for wireless image communication: The wireless imaging quality database," in *Proc. 5th Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, Scottsdale, AZ, USA, Jan. 2010, pp. 1–5.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[10] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[11] M. Pedersen and J. Y. Hardeberg, "Survey of full-reference image quality metrics," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 1, pp. 1–80, 2024. [Online]. Available: `https://doi.org/10.1561/0600000037`.

[12] I. T. Union, *Recommendation itu-r bt.500-15: Methodologies for the subjective assessment of the quality of television images*, Available: https://www.itu.int/rec/R-REC-BT.500/en, 2023.

[13] A. Zaric et al., "Image quality assessment - comparison of objective measures with results of subjective test," pp. 113–118, 2010.

[14] *Subjective video quality assessment methods for multimedia applications*, International Telecommunication Union, Apr. 2008.

[15] N. Ponomarenko et al., "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[16] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[17] *Amazon mechanical turk*, Accessed: Aug. 20, 2019. [Online]. Available: https://www.mturk.com/.

[18] H. R. Sheikh, A. C. Bovik, and G. d. Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[19] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[20] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," in *Proc. SPIE Electron. Imaging*, vol. 5666, San Jose, CA, USA, Mar. 2005, pp. 149–159.

[21] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 202–211, Jun. 2009.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[23] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[24] J. Hernandez-Ortega et al., "Faceqnet: Quality assessment for face recognition based on deep learning," *Image and Vision Computing*, pp. 1–15, 2020.

[25] F. Boutros et al., "Iqface: A face image quality assessment model based on deep learning," *IEEE Biometrics Compendium*, pp. 123–135, 2021.

[26] S. Li et al., "Biofacenet: Optimized quality assessment for secure biometrics," *Journal of Biometrics Research*, pp. 1–16, 2021.

[27] W. Xu et al., "Secureqnet: Integrating quality and security for identity documents," *IEEE Transactions on Biometrics*, pp. 345–360, 2020.

[28] X. Luo et al., "Deepiq: A learning-based metric for facial image quality," *Pattern Recognition Letters*, pp. 12–20, 2018.

[29] B. Jo, I. K. Park, and S. Hong, "Perceptual metric for face image quality with pixel-level interpretability," *Neurocomputing*, vol. 614, p. 128 780, 2025. [Online]. Available: `https://doi.org/10.1016/j.neucom.2024.128780`.

[30] B. Jo et al., "Perceptual metric for face image quality with pixel-level interpretability," *Neurocomputing*, vol. 614, p. 128 780, 2025.

[31] 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.

[32] *Doc 9303 machine readable travel documents, part 3: Specifications common to all mrtds*, International Civil Aviation Organization, Sep. 2015.

[33] *Information technology – biometric sample quality – part 5: Face image data*, International Organization for Standardization, Apr. 2010.

[34] C. of Europe, *European convention on human rights, article 14: Prohibition of discrimination*, Council of Europe Treaty Series No. 5, 1950. [Online]. Available: `https://www.echr.coe.int/Documents/Convention_ENG.pdf`.

[35] U. Nations, *Universal declaration of human rights, article 7: Equality before the law*, General Assembly Resolution 217 A, 1948. [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[36] E. Union, *General data protection regulation (eu) 2016/679, article 22: Automated individual decision-making, including profiling*, Official Journal of the European Union, L 119, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[37] E. Union, *Artificial intelligence act (ai act) - regulation (eu) 2024/1689*, Official Journal of the European Union, L 1689, 12 July 2024, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[38] T. W. House, *Blueprint for an ai bill of rights: Making automated systems work for the american people*, Office of Science and Technology Policy, United States, 2022. [Online]. Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

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

[40] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. 2010, ISBN: 9780521190190. DOI: 10.1017/CBO9781139192903.

[41] F. Shadmand, I. Medvedev, and N. Goncalves, "Codeface: A deep learning printer-proof steganography for face portraits," *IEEE Access*, vol. 9, pp. 167 282–167 291, 2021.

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

[43] O. Henniger et al., "On the assessment of face image quality based on handcrafted features," in *Proceedings of BIOSIG*, 2020.

[44] K. Hernandez-Ortega et al., "Faceqnet: Quality assessment for face recognition based on deep learning," in *ICB*, 2019.

[45] Q. Meng et al., "Magface: A universal representation for face recognition and quality assessment," in *CVPR*, 2021.

[46] P. Terhörst et al., "Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *CVPR*, 2020.

[47] S. Liu, J. Lin, Y. Fang, Y. Wang, and W.-S. Lin, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1040–1049.

[48] Ž. Babnik and V. Štruc, *Assessing bias in face image quality assessment*, 2022. arXiv: 2211.15265 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2211.15265.

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

[50] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[51] 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.

[52] L. Jin et al., "Pipal: A large-scale image quality assessment dataset for perceptual image restoration," in *ECCV*, 2020.

[53] G. IOO, "Recommendation itu-t p. 10 vocabulary for performance and quality of service. itu-t, 07," 2006.

[54] 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.

[55] J.-C. Yoo and C. Ahn, "Image matching using peak signal-to-noise ratio-based occlusion detection," *IET image processing*, vol. 6, no. 5, pp. 483–495, 2012.

[56] 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. DOI: 10.1109/JSTSP.2012.2204259.

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

[58] 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.

[59] 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.

[60] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

[62] H. R. Sheikh and A. C. Bovik, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2117–2128, 2005.

[63] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, 2000.

[64] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, et al., "The spectral image processing system (SIPS) – Interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2, pp. 145–163, 1993.

[65] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[66] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[67] D. M. Chandler and S. S. Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[68] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[69] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 1, 2002, pp. 477–480.

[70] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[71] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 5, pp. 593–602, 2008.

[72] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd ed.)* Upper Saddle River, NJ: Prentice Hall, 2008.

[73] M. A. Hassan and M. S. Bashraheel, "Color-based Structural Similarity Image Quality Assessment," in *Proc. 8th Int. Conf. on Information Technology (ICIT)*, 2017, pp. 691–696.

[74] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2385–2401, 2009.

[75] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image Quality Assessment: Unifying Structure and Texture Similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, 2022.

[76] X. Zhang, X. Mou, W. Wang, and G. Shi, "Edge strength similarity for image quality assessment," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 319–322, 2013.

[77] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, 2005.

[78] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[79] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, 2012.

[80] Q. Li and Z. Wang, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, 2011.

[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.

[82] K. Yim and A. C. Bovik, "Quality assessment of deblocked images," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 88–98, 2011.

[83] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Systems & Computers*, 2003, pp. 1398–1402.

[84] P. Chen, L. Li, Q. Wu, and J. Wu, "SPIQ: A self-supervised pre-trained model for image quality assessment," *IEEE Signal Process. Lett.*, vol. 29, pp. 513–517, 2022.

[85] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2012, pp. 1473–1476.

[86] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[87] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[88] H. R. Sheikh and A. C. Bovik, *Pixel-domain visual information fidelity (VIF) implementation*, Online: http://live.ece.utexas.edu/research/Quality/VIF.htm, 2005.

[89] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, 2018.

[90] U. Engelke, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, 2011.

[91] A. Mittal, A. K. Moorthy, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[92] V. Hosu, J. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

[93] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[94] J. Yim and A. C. Bovik, "Quality metric for image compression based on blocking effect," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 88–98, 2011.

[95] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The arsis concept and its implementation," *Photogrammetric Engineering and Remote Sensing*, vol. 66, no. 1, pp. 49–61, 2000.

[96] F. A. Kruse, J. W. Boardman, and J. F. Huntington, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," in *Summaries of the 4th Annual JPL Airborne Geoscience Workshop*, JPL, vol. 1, 1993, pp. 15–16.

[97] A. Venkatanath, R. Praneeth, C. S. Babu, M. A. Gopalakrishna, A. S. Nair, and A. Prabhakar, "Blind image quality evaluation using perception based features," *National Conference on Communications (NCC)*, pp. 1–6, 2015. DOI: 10.1109/NCC.2015.7084923.

[98] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013. DOI: 10.1109/LSP.2012.2227726.

[99] N. Damer, F. Boutros, M. Fang, F. Kirchbuchner, and A. Kuijper, "Local fusion for face image quality assessment," in *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–10.

[100] L. Best-Rowden and A. K. Jain, "Automatic face image quality prediction," *arXiv preprint arXiv:1806.07394*, 2018.

[101] P. Grother, P. J. Phillips, and R. Micheals, "Face recognition vendor test 2002 performance report," National Institute of Standards and Technology, Tech. Rep. NISTIR 6965, 2003.

[102] P. Terhörst, B. Steffen, N. Damer, and A. Kuijper, "The relationship between face image quality and recognition performance: A longitudinal study," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1626–1639, 2022.

[103] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: Too bias, or not too bias?," 2020. arXiv: 2002.06483 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2002.06483.

[104] C. Wu, K. Ma, Z. Duanmu, S. Wang, and W. Zeng, "Cascaded perceptual quality assessment using pseudo-reference learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 9654–9666, 2020.

[105] S. Baluja, "Hiding images in plain sight: Deep steganography," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/838e8afb1ca34354ac209f53d90c3a43-Paper.pdf.

[106] A. Fkirin, G. Attiya, and A. El-Sayed, "Steganography literature survey, classification and comparative study," *Communications on Applied Electronics*, vol. 5, pp. 13–22, Sep. 2016. DOI: 10.5120/cae2016652384.

[107] B. Demirci and K. Tutuncu, "Adaptive lsb steganography based on chaos theory and random distortion," *Advances in Electrical and Computer Engineering*, vol. 18, pp. 15–22, Jan. 2018.

[108] 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. arXiv: 1706.08500 [cs.LG].

[109] 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)*, Tucson, USA: IEEE, 2025.

[110] F. Shadmand, I. Medvedev, L. Schirmer Silva, J. Sena Marcos, and N. Gonçalves, "Stampone: Addressing frequency balance in printer-proof steganography," in *CVPR Workshops*, 2024, pp. 4367–4376. DOI: 10.1109/CVPRW63382.2024.00440.

[111] L. DeBruine and B. Jones, "Face research lab london set," *Open Science Framework*, 2017. DOI: 10.17605/OSF.IO/FPZ3U.

[112] A. Clapés, M. Oliu, M. Farrús, and S. Escalera, "From apparent to real age: Gender, age, ethnicity and error perception," in *CVPR Workshops*, 2018, pp. 0-0.

[113] A. Ross, V. L. Willson, A. Ross, and V. L. Willson, "One-way anova," *Basic and advanced statistical tests: Writing results sections and creating tables and figures*, pp. 21–24, 2017.

[114] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140 030–140 070, 2019. DOI: 10.1109/ACCESS.2019.2943319.

[115] 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. DOI: 10.1016/j.image.2017.05.001.

[116] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*. SIAM, 1998.

[117] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd. Springer, 2009.

[118] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[119] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[120] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[121] D. Basak, S. Pal, and D. Patranabis, "Support vector regression," *Neural Information Processing - Letters and Reviews*, vol. 11, Nov. 2007.

[122] L. Breiman, "Random forests," English, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, ISSN: 0885-6125. DOI: 10.1023/A:1010933404324. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324.

[123] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[124] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[125] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[126] 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*, ser. KDD '16, ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785.

[127] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *NeurIPS*, vol. 30, 2017.

[128] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," in *NeurIPS*, vol. 31, 2018.

[129] 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)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[130] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.

[131] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[132] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," 2003.

[133] 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.

Appendix A

Sample Appendix