

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

Acknowledgements iii

Resumo v

Abstract vii

List of Acronyms xiii

List of Figures xvii

List of Tables xix

1 Introduction 1

1.1 Contextualization . . . 1

1.2 Problem Statement . . . 3

2 Related Work 5

2.1 Subjective IQA . . . 5

2.1.1 Assessment Protocols . . . 5

2.1.2 Crowdsourcing and Remote Studies . . . 6

2.1.3 Benchmark Datasets . . . 6

2.2 Objective IQA . . . 8

2.2.1 Computational Domains . . . 8

2.2.2 Spectral Representations . . . 9

2.2.3 Full-Reference IQA . . . 9

2.2.4 No-Reference IQA . . . 21

2.2.5 Fusion-Based IQA Models . . . 24

2.3 Image distortion types . . . 27

2.3.1 Lossy compression . . . 28

2.3.2 Blur . . . 28
2.3.3 Additive noise . . . 28
2.3.4 Resampling and resolution changes . . . 29
2.3.5 Photometric and color distortions . . . 29
2.3.6 Transmission and packet-level artifacts . . . 29
2.3.7 Authentic mixtures . . . 29
2.3.8 Steganography . . . 29

3 Methodology **33**
3.1 Dataset . . . 33
3.2 Collection of Subjective Scores . . . 35
3.3 Debiasing of Subjective Scores . . . 39
3.4 Correlation of FR-IQA Metrics with Human Perception . . . 40
3.5 Fusion of FR-IQA Metrics for Pseudo-MOS Estimation . . . 42
3.6 Training a No-Reference IQA Model from Pseudo-MOS . . . 43
3.7 Results . . . 43
3.7.1 Full-Reference Fusion Metric Performance . . . 43
3.7.2 No-Reference Regression Model Performance . . . 44
3.8 Conclusion and Future Work . . . 46

A Subjective Screening Procedures **59**
A.1 Confidence Intervals . . . 59
A.2 Observer Screening . . . 59
A.2.1 Outlier Detection . . . 60
A.2.2 Rejection Rule . . . 60
A.3 Summary . . . 60

List of Acronyms

IQA	Image Quality Assessment
ITU-R BT.500-15	International Telecommunication Union - Radiocommunication Sector Recommendation Broadcasting service (television) 500-15
FR	Full-Reference
NR	No-Reference
RR	Reduced-Reference
FIQA	Facial Image Quality Assessment
ICAO	International Civil Aviation Organization
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission
MRTDs	Machine-Readable Travel Documents
MOS	Mean Opinion Score
DMOS	Difference Mean Opinion Score
SS	Single Stimulus
DS	Double Stimulus
FRLL	Face Research Lab London
HVS	Human Visual System
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error

PSNR	Peak Signal-to-Noise Ratio
PSNR-B	Blocking Effect Aware Peak Signal-to-Noise Ratio
BEF	Blocking Effect Factor
SNR	Signal-to-Noise Ratio
SAM	Spectral Angle Mapper
ERGAS	Error Relative Global Dimensionless Attribute Similarity
SSIM	Structural Similarity Index Metric
NIQE	Natural Image Quality Evaluator
NSS	Natural Scene Statistics
MSCN	Mean Subtracted Contrast Normalization
GGD	Generalized Gaussian Distribution
AGGD	Asymmetric Generalized Gaussian Distribution
MVG	Multivariate Gaussia
PIQE	Perception-based Image Quality Evaluator
LPF	Low-Pass Filter
FaceQnet	Facial Quality Network
SER-FIQ	Stochastic Embedding Robustness for Face mage Quality
MagFace	Magnitude-Aware Face
MMF	Multi-Method Fusion
PLCC	Pearson Linear Correlation Coefficient
SRCC	Spearman Rank Order Correlation Coefficient
SVR	Support Vector Regression
RBF	Radial Basis Function
LSB	Least Significant Bit

DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
GANs	Generative Adversarial Networks
SPD	Symmetric Positive Definite
STN	Spatial Transformer Network

List of Figures

3.1 Reference images from the MOS set, comprising the selected subjects from the FRLL dataset. 34
3.2 Steganographically distorted facial images from each method. 34
3.3 Django-based webapp created for the Single Stimulus test. 36
3.4 Conceptual and physical representations of the database schema. 38
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. . . 41
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 $k = 6$. 43
3.7 SHAP summary plot for $k = 6$, generated from the best CatBoost model. 44
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. 45

List of Tables

2.1 Summary of the 40 FR-IQA metrics used 10

2.2 Summary of the 5 NR-IQA metrics used 21

3.1 Partitioning of the steganographically distorted dataset. 34

3.2 Demographic and non-demographic profile of the observers. 36

3.3 ANOVA [131] 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. 40

3.4 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. 44

3.5 Performance of our pMOS-Face compared to standard NR-IQA baselines. 45

1 Introduction

1.1 Contextualization

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

Image Quality Assessment (IQA) is the process of quantifying image quality in a systematic and reproducible manner. It plays a central role in applications such as image acquisition [3], denoising [4], compression [5], [6], transmission [7], [8], and restoration [9]. Surveys and comparative studies [1], [10], [11], [12] highlight the diversity of approaches, ranging from signal-based models to data-driven learning techniques, reflecting the complexity of modeling human visual perception.

IQA methods are generally divided into subjective and objective categories. Subjective assessment relies on human observers, typically following standardized protocols such as International Telecommunication Union - Radiocommunication Sector Recommendation Broadcasting service (television) 500-15 (ITU-R BT.500-15) [13], and provides the most direct alignment with perceptual quality. However, it is costly, time-consuming, and not scalable. Objective methods, by contrast, employ computational models to automatically estimate quality, offering scalability and repeatability. Their effectiveness depends on how well they approximate subjective judgments, a long-standing challenge documented in IQA benchmarks and reviews [10].

Within objective IQA, three main categories are distinguished: Full-Reference (FR) [1], No-Reference (NR) [14], and Reduced-Reference (RR) [15]. FR approaches assume access to pristine reference images, RR methods use partial reference information, and NR approaches attempt to predict quality without any reference, making them the most challenging and increasingly important for real-world scenarios.

A specialized subfield, Facial Image Quality Assessment (FIQA), addresses the assessment

of face images with respect to both perceptual fidelity and biometric utility [16]. High-quality facial images are essential for accurate recognition in security, forensics, and surveillance. Unlike general IQA, FIQA must account for unique challenges such as pose, occlusion, and demographic variability [17], [18]. This reflects a broader trend in IQA research toward domain-specific assessment strategies [12].

The International Civil Aviation Organization (ICAO) [19] and the International Organization for Standardization / International Electrotechnical Commission (ISO/IEC) 19794–5 standard [20] establish guidelines for image quality in Machine-Readable Travel Documents (MRTDs). These guidelines ensure uniform image conditions (e.g., lighting, focus, and resolution) and consistency across datasets. While these regulations establish a technical baseline, they do not account for perceptual biases and demographic variability in FIQA.

FIQA systems have been shown to produce different quality scores depending on a person's age, gender, or skin tone [21], [22]. 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 [23]. This can cause face recognition systems to perform worse for some groups than others, raising serious concerns about fairness.

The ethical implications of these biases are profound. Political regulations, such as the European Convention on Human Rights (Article 14) [24], the Universal Declaration of Human Rights (Article 7) [25], the General Data Protection Regulation (Article 22) [26], and emerging AI governance frameworks, such as the European Artificial Intelligence Act (2024) [27] and proposals in the Artificial Intelligence Civil Rights Act (2024) of the U.S. Congress [28], 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 [23], [29]. 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 [30]. 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.

This issue is especially relevant for MRTDs, where enhanced security is critical. Incorporating steganographic imagery, such as hidden tamper-evident or authentication data, helps protect against increasingly sophisticated threats to identity documents. The necessity for such measures is underscored by growing criminal activity involving IDs. Reports from the European Border and Coast Guard Agency consistently highlight document fraud as one of the most serious threats at EU external borders [31]. Similarly, Europol identifies fraudulent IDs as a key enabler of organized crime and migrant smuggling [32]. INTERPOL's SLTD database currently contains more than 138 million records of lost or stolen travel documents, illustrating the global scale of the problem [33]. In the United States alone, identity fraud losses reached an estimated $43 billion in 2023 [34].

Steganographically embedding cryptographic signals in passport portraits can therefore serve as a complementary line of defense alongside chip-based authentication. This approach directly addresses known threats such as morphing attacks [35], photo substitution [19], and stolen-document misuse [33], while aligning with ICAO's emphasis on secure and standardized travel document images [19].

1.2 Problem Statement

Despite substantial advances in FIQA, existing methods are not designed to capture subtle but security-critical degradations, particularly those introduced by steganographic embedding. Steganography alters pixel statistics in ways that are imperceptible to human observers yet capable of undermining biometric feature extraction. Conventional FIQA approaches, whether based on handcrafted features **henniger2020biosig**, supervised learning from quality annotations [36], [37], [38], or ranking-based formulations [39], are typically trained on distortion categories represented in standard datasets (e.g., blur, noise, compression). Consequently, they fail to generalize to covert perturbations that do not visibly degrade an image but can critically impair recognition accuracy.

A second limitation lies in fairness. Empirical studies show that FIQA models often inherit demographic biases from their training data, leading to systematically lower quality scores for specific subgroups [21]. In high-stakes applications such as MRTDs, governed by ICAO and ISO/IEC standards [19], [20], these disparities raise profound ethical and legal concerns. At the same time, evidence from border security agencies highlights that document fraud, morphing attacks, and illicit ID use remain prevalent, underscoring the necessity of robust and equitable quality control mechanisms.

This motivates the development of no-reference FIQA methods explicitly sensitive to both perceptual fidelity and biometric utility in the presence of steganographic distortions. To address this gap, we propose a framework that fuses multiple full-reference metrics into a proxy ground truth aligned with subjective quality judgments. This fused metric supervises the learning of a NR-FIQA model tailored to steganographically modified facial images. By embedding considerations of task relevance, perceptual validity, and demographic fairness, our approach aims to provide a principled bridge between human perception, biometric system requirements, and the operational security demands of MRTDs.

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.500-15 [13], these methodologies rely on controlled laboratory conditions, calibrated displays, and predefined rating procedures to ensure reproducibility and validity.

2.1.1 Assessment Protocols

Protocols are typically divided into two categories: quality assessment, which captures the overall perceived quality of an image, and impairment assessment, which measures degradation relative to a reference. Two methodological paradigms are most common: the Single Stimulus (SS) method, where each image is rated independently, and the Double Stimulus (DS) method, where reference and test images are evaluated comparatively. SS protocols dominate modern datasets due to their simplicity, efficiency, and reduced observer fatigue.

Rating scales can be numerical or categorical. Numerical scales range from 1 to 5, 1 to 11, or 0 to 100, and are used to express either quality (excellent to bad) or impairment (imperceptible to very annoying). Categorical scales rely on descriptive labels, and are commonly found in ITU protocols. The choice of scale depends on whether the target measure is Mean Opinion Score (MOS) or Difference Mean Opinion Score (DMOS), and on the desired sensitivity of the test.

Each test session consists of multiple presentations. For a given presentation defined by condition j, image or sequence k, and repetition r, observer i provides a discrete score u_{ijkr}. The MOS is computed as:

$$\mathrm{MOS}_{jkr} = \frac{1}{N} \sum_{i=1}^{N} u_{ijkr} \tag{2.1}$$

where N is the number of observers. In double-stimulus protocols, the DMOS is defined as the

difference between the MOS of the reference and that of the distorted image:

$$\mathrm{DMOS}_{jkr} = \mathrm{MOS}_{\mathrm{ref},jkr} - \mathrm{MOS}_{\mathrm{dist},jkr} \tag{2.2}$$

These values are often accompanied by confidence intervals to quantify statistical uncertainty and by post-screening to remove inconsistent observers. The specific derivations, including standard deviation, confidence intervals, and the ITU-recommended kurtosis-based rejection procedure, are detailed in Appendix A.

2.1.2 Crowdsourcing and Remote Studies

While laboratory-based studies remain the benchmark for reliability, they are costly, time-consuming, and limited in participant diversity. To overcome these constraints, recent works have turned to crowdsourcing platforms such as Amazon Mechanical Turk, Figure Eight, and Prolific [40], [41]. These frameworks enable the collection of tens of thousands of ratings across heterogeneous populations at a fraction of the cost of laboratory tests.

The advantages of crowdsourcing lie in its scalability, ecological validity, and demographic diversity, which allow researchers to capture quality judgments from a broad spectrum of viewing conditions and cultural backgrounds. This is particularly relevant for applications where image quality intersects with biometric fairness, as demographic variability plays a significant role in subjective perception.

However, uncontrolled environments introduce noise: participants use uncalibrated displays, ratings may be inconsistent, and some responses are malicious or inattentive. To mitigate these issues, robust post-screening strategies are employed, including consistency checks, gold-standard reference images, and statistical filtering [40], [42]. Despite these limitations, crowdsourced IQA datasets have proven critical for training modern deep learning-based models, as they provide both scale and diversity unavailable in controlled laboratory conditions.

2.1.3 Benchmark Datasets

Several benchmark datasets have been curated to support the development and evaluation of IQA models. Early datasets such as LIVE [43], [44], [45], CSIQ [46], TID2008 [47], and TID2013 [48] follow ITU-R BT.500-15-compliant protocols, providing high-quality MOS or DMOS annotations across a wide variety of synthetic distortion types. These databases remain the most widely used for evaluating traditional full-reference IQA metrics.

The IVC [49] dataset further contributed subjective quality scores for compression-related distortions, while SIQAD [50] targeted screen content images, extending IQA research to non-

natural imagery. KADID-10k [51] scaled this approach to 10,000 distorted images with 25 distortion types and multiple intensity levels, enabling large-scale training and fine-grained benchmarking. The Waterloo IAA database [52] emphasized robustness testing through massive-scale synthetic distortions and helped reveal model overfitting issues to specific artifact distributions.

More recent contributions shifted toward naturalistic and crowdsourced assessments. KonIQ-10k [42] introduced 10,073 images collected from the web with authentic distortions and over 1.2 million quality ratings, making it one of the largest ecologically valid IQA resources. SPAQ [53] specifically addressed smartphone photography, collecting 11,125 images and crowdsourced perceptual scores to capture the variability of mobile imaging pipelines. PIPAL [54] targeted perceptual quality in image restoration, including outputs from generative models, and has become a key benchmark for deep learning-based perceptual metrics. Similarly, BAPPS [55] provided a large-scale pairwise comparison dataset for learning perceptual similarity directly from human judgments.

Beyond general-purpose IQA, domain-specific datasets remain critical. Remote sensing benchmarks such as ROSIS [56], AVIRIS [57], and HYDICE [58] provide hyperspectral imagery where distortions affect both perceptual and functional quality, but they lack standardized MOS or DMOS annotations. These datasets illustrate the difficulty of extending IQA methodologies into application-driven contexts.

The Face Research Lab London (FRLL) [59] represents a distinct contribution to IQA benchmarking, focusing on high-quality facial ICAO compliant imagery. It contains 102 unique subjects photographed under controlled studio conditions, with consistent illumination, wearing plain white t-shirts, and neutral, frontal poses. In addition to meeting passport-style requirements, the dataset also includes self-reported age, gender, and ethnicity. Attractiveness ratings (on a 1-7 scale from "much less attractive than average" to "much more attractive than average") for the neutral front faces, provided by 2513 raters aged 17-90, are also included.

This combination of demographic attributes and attractiveness ratings enables not only research that bridges perceptual quality with subjective aesthetic judgments, but also systematic analysis of demographic bias in human evaluations of facial images. These dual annotations make the dataset particularly suitable for studies involving biometric verification and MRTDs, where both technical quality and fairness considerations are critical.

Unlike general-purpose IQA benchmarks, FRLL does not include traditional synthetic distortions. Instead, its value lies in providing a pristine reference set of demographically diverse facial images, which can be systematically altered or embedded with task-driven degradations such as steganography, compression, or adversarial perturbations. This makes FRLL an essen-

tial testbed for application-driven IQA, where the objective extends beyond generic perceptual fidelity to domain-specific measures of utility and fairness.

Together, these databases trace the evolution of IQA benchmarking from controlled laboratory studies with synthetic distortions to large-scale, crowdsourced, and application-specific datasets. This progression not only enabled the development of modern no-reference and deep learning-based IQA models but also revealed persistent challenges in generalization to real-world, domain-critical scenarios.

2.2 Objective IQA

Objective IQA refers to the automatic estimation of visual quality using computational models. Unlike subjective approaches that rely on human ratings, objective IQA enables consistent and scalable evaluations, making it essential for applications such as compression, transmission, restoration, and enhancement [60], [61], [62]. Over the past decades, numerous methods have been proposed, shaped by different assumptions about the Human Visual System (HVS) and the statistical properties of natural images. Surveys and comparative studies [63], [64], [65], [66] summarize this evolution, showing a progression from simple error-based fidelity measures to advanced perceptual and learning-based frameworks. Although no metric fully replicates human judgment, objective IQA provides a reproducible and low-cost alternative that supports large-scale benchmarking and real-time quality monitoring.

A principled way to classify IQA methods is through two dimensions: the computational domain in which comparisons are performed and the spectral representation adopted.

2.2.1 Computational Domains

The computational domain defines the signal space where similarity is quantified.

- Spatial domain: Early metrics such as MSE, RMSE, and PSNR operate directly on pixel intensities by measuring absolute differences or signal-to-noise ratios [60]. Despite their simplicity, they fail to account for perceptual masking and structural consistency.
- Frequency domain: Transform-based methods use the discrete cosine transform (DCT), discrete Fourier transform (DFT), or wavelets to capture energy distributions across spatial frequencies. Examples include IFC [67] and VSNR [68], which exploit properties of the HVS such as contrast sensitivity.
- Gradient and saliency domains: These methods emphasize structural edges and perceptually important regions. The gradient magnitude similarity deviation (GMSD) [69] mea-

sures gradient consistency, while saliency-based models [70] weight distortions according to visual attention maps.

- Information-theoretic domain: Some metrics interpret quality as the amount of shared information between reference and distorted images. VIF [71] and related approaches build on mutual information, quantifying how much perceptual information can be extracted reliably from a degraded signal.

- Deep feature domain: More recent approaches leverage convolutional neural networks or vision transformers to learn perceptual embeddings. Metrics such as LPIPS [72] and DISTS [66] correlate strongly with human judgments by comparing features extracted from pre-trained networks.

2.2.2 Spectral Representations

The spectral representation specifies the channel configuration used for quality computation.

- Luminance-based: Many classical approaches, such as SSIM [64], focus solely on the luminance channel (Y). This reflects the dominant role of brightness perception in the HVS and simplifies computation.

- Color-based: Extensions to RGB or alternative color spaces, including YCbCr and CIELab, enable the incorporation of chromatic information. Examples include FSIM and its color extension FSIMc [73], or PQM [74], which attempt to capture hue and saturation degradations alongside luminance.

- Multispectral and hyperspectral: In domains such as remote sensing, quality measures generalize to dozens or even hundreds of bands. Metrics like SAM [75], ERGAS [76], and D_λ [49] ensure spectral consistency across wavelengths and are critical for preserving material properties.

This dual perspective highlights the diversity of IQA methodologies and provides a framework for their categorization. Building on it, Table 2.1 presents the 40 FR metrics and Table 2.2 summarizes the 4 NR metrics considered in this work. Each metric is characterized by its reference type, year of introduction, category domain, and tested databases, showing how the field has advanced from simple pixel-wise error measures to perceptually and semantically informed quality predictors.

2.2.3 Full-Reference IQA

Table 2.1: Summary of the 40 FR-IQA metrics used

Metric	**Year**	**Category**	**Tested Databases**
MSE [64]	2004	Error-based	LIVE, CSIQ, TID2008, TID2013, KADID-10k
RMSE [64]	2004	Error-based	LIVE, CSIQ, TID2008, TID2013, KADID-10k
MAE [64]	2004	Error-based	LIVE, CSIQ, TID2008, TID2013, KADID-10k
PSNR [64]	2004	Error-based	LIVE, CSIQ, TID2008, TID2013, KADID-10k
PSNR-B [77]	2011	Error-based	LIVE, CSIQ, TID2013
SNR [64]	2004	Error-based	LIVE, CSIQ, TID2008
SAM [75]	1992	Spectral-based	AVIRIS, HYDICE, ROSIS, TID2013
ERGAS [76]	2000	Spectral-based	AVIRIS, ROSIS, TID2013, KADID-10k
D_λ [76]	2000	Spectral-based	AVIRIS, ROSIS, TID2013, KADID-10k
SSIM [64]	2004	Structural Similarity	LIVE, CSIQ, TID2008, TID2013
DSSIM [64]	2004	Structural Similarity	LIVE, CSIQ, TID2008
MS-SSIM [78]	2003	Structural Similarity	LIVE, CSIQ, TID2013, KADID-10k
IW-SSIM [79]	2010	Structural Similarity	LIVE, CSIQ, TID2013
L-SSIM [64]	2004	Structural Similarity	—
UQI [74]	2002	Structural Similarity	—
SCC [80]	1995	Structural Similarity	—
CW-SSIM [81]	2009	Wavelet Similarity	LIVE, CSIQ, TID2013
W-SSIM [82]	2009	Wavelet Similarity	LIVE
ESSIM [83]	2006	Edge Similarity	LIVE
G-SSIM [84]	2006	Gradient Similarity	LIVE, TID2008
R-SSIM [85]	2015	Gradient Similarity	—
GMS [69]	2014	Gradient Similarity	LIVE, CSIQ, TID2008
GMSD [69]	2014	Gradient Similarity	LIVE, CSIQ, TID2008
C-SSIM [86]	2017	Color Similarity	TID2013
FSIM [73]	2011	Feature Similarity	LIVE, CSIQ, TID2008, TID2013, KADID-10k
FSIMc [73]	2011	Feature Similarity	LIVE, CSIQ, TID2008, IVC

(continuation)

Metric	Year	Category	Tested Databases
MS-FSIM [73]	2011	Color Similarity	—
SR-SIM [87]	2013	Feature Similarity	LIVE, CSIQ, TID2008
VSNR [68]	2007	Feature Similarity	LIVE, CSIQ
IFC [67]	2005	Information	LIVE
VIF [71]	2006	Information	LIVE, CSIQ, TID2008
VIFp [71]	2006	Information	LIVE
VIFc [71]	2006	Information	—
LPIPS-Alex [72]	2018	Deep Learning	BAPPS, LIVE, CSIQ, TID2013
LPIPS-Squeeze [72]	2018	Deep Learning	BAPPS, LIVE, CSIQ, TID2013
LPIPS-VGG [72]	2018	Deep Learning	BAPPS, LIVE, CSIQ, TID2013
DISTS [66]	2020	Deep Learning	LIVE, CSIQ, TID2013, KADID-10k
WaDIQaM [88]	2017	Deep Learning	LIVE, TID2013
PSM [89]	2022	Deep Learning	BAPPS
SPIQ [90]	2022	Deep Learning	KADID-10k, KonIQ-10k

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 [64], [71]. 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].

Error-Based Metrics

Error-based metrics quantify distortion by computing direct numerical differences between the reference and distorted images [64]. These measures are mathematically simple and computationally efficient, but they lack perceptual modeling and therefore correlate poorly with human opinion scores. Despite this limitation, they are still treated as FR-IQA methods and remain

widely used as baselines.

It is important to note that although several error-based metrics are standard mathematical tools in signal processing, their use in IQA involves a shift in interpretation. In this context, they are repurposed to estimate perceptual image degradation based solely on numerical pixel differences. Despite lacking any perceptual modeling, these metrics are still treated as FR-IQA methods and evaluated empirically against subjective human judgments.

Notation. Let $X = \{x_i\}_{i=1}^N$ and $Y = \{y_i\}_{i=1}^N$ be two images with N pixels each, where x_i and y_i are corresponding pixel intensities. The pixel-wise error is defined as $e_i = x_i - y_i$.

The most basic error-based metric is the Mean Squared Error (MSE) [64], which quantifies the average squared difference between corresponding pixels in the reference and distorted images:

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

The Root Mean Squared Error (RMSE) [64] is a direct extension of the mean squared error, obtained by taking its square root:

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

By restoring the metric's scale to that of the original image intensities, RMSE offers more interpretable results than MSE, especially when comparing distortion magnitudes across images.

The Mean Absolute Error (MAE) [64] measures the average absolute difference between the reference and distorted image pixels:

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

MAE penalizes all errors linearly, making it less sensitive to large deviations and outliers. This results in a more stable error measure in scenarios with isolated distortions or non-Gaussian noise.

The Peak Signal-to-Noise Ratio (PSNR) [64] expresses the fidelity between a reference and a distorted image using a logarithmic decibel (dB) scale. It is defined as:

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

where MAX is the maximum possible pixel value (typically 255 for 8-bit images). PSNR remains popular due to its simplicity, ease of interpretation, and long-standing usage in compression evaluation.

The Blocking Effect Aware Peak Signal-to-Noise Ratio (PSNR-B) [77], [93] was introduced to address one of the main weaknesses of PSNR, its insensitivity to compression artifacts such

as blocking. PSNR-B extends the original formula by adding a Blocking Effect Factor (BEF) to the denominator:

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

The BEF penalizes the presence of visible discontinuities along block boundaries, which commonly result from transform-based compression schemes.

The Signal-to-Noise Ratio (SNR) [64] quantifies the proportion of signal power relative to noise power between a reference image and its distorted counterpart. It is computed as:

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

This metric provides a direct measure of how much of the original signal is preserved in the presence of distortion.

Spectral-based Metrics

Spectral-based metrics evaluate the consistency of spectral information in multi-band or hyperspectral images [75], [76]. Each pixel is treated as a spectral vector, and quality is assessed by comparing structural relationships in the spectral domain. These methods are widely used in remote sensing and image fusion, where spectral integrity is critical.

Notation. Let $x, y \in \mathbb{R}^B$ denote the spectral vectors of the reference and distorted images at a given pixel, where B is the number of spectral bands. The Euclidean norm is denoted $\| \cdot \|$, the inner product by $\langle \cdot, \cdot \rangle$, and μ_i the mean of band i. Spatial resolutions of the high- and low-resolution images are h and l, respectively. Pairwise spectral distances between bands i and j are written as $d(r_i, r_j)$, where r_i is the i-th band.

The Spectral Angle Mapper (SAM) [75] computes the angular difference between the spectral vectors of the reference and distorted images:

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

The Error Relative Global Dimensionless Attribute Similarity (ERGAS) [76], [94] measures the average relative root mean squared error across spectral bands:

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

The D_λ metric [76] evaluates spectral distortion by averaging inter-band distances:

$$D_\lambda = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{\substack{j=1 \\ j \neq i}}^{B} d(r_i, r_j). \tag{2.11}$$

Similarity-based Metrics

Similarity-based metrics assess perceptual image quality by measuring structural or feature-based resemblance between the reference and distorted images. Rather than relying on pixel-wise errors, they capture patterns of spatial organization, texture, luminance, or phase alignment. These methods often align more closely with the HVS than purely error-based metrics [64], [69].

Notation. Let x and y denote local patches from the reference and distorted images. Their means are μ_x and μ_y, variances σ_x^2 and σ_y^2, and covariance σ_{xy}. Constants $C_1, C_2 > 0$ are used to avoid instabilities in divisions.

The Structural Similarity Index Metric (SSIM) [64] is a perceptual metric designed to evaluate image quality by comparing structural information between a reference and a distorted image. Unlike traditional error-based metrics, SSIM models luminance, contrast, and structure in a local window. Given two image patches x and y, SSIM is computed as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{2.12}$$

SSIM values range from 0 to 1, where 1 indicates perfect similarity. Multiscale SSIM (MS-SSIM) extends SSIM by computing similarity at multiple image scales. It is defined as:

The Dissimilarity SSIM (DSSIM) [64] is a reformulation of SSIM that expresses dissimilarity rather than similarity. It is derived directly from the SSIM value by:

$$\text{DSSIM}(x, y) = \frac{1 - \text{SSIM}(x, y)}{2} \tag{2.13}$$

This transformation enables DSSIM to behave like a distance function, where 0 represents identical images and higher values indicate greater perceptual difference. Despite its simplicity, DSSIM is sometimes preferred in optimization and comparison tasks where dissimilarity is more intuitive.

The Multi-Scale SSIM (MS-SSIM) [78] extends SSIM by computing the structural similarity at multiple scales, capturing both fine and coarse details. This is achieved by iteratively downsampling the input images and evaluating luminance, contrast, and structure components at each level. The final score is a weighted geometric mean of the similarity scores across scales:

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \tag{2.14}$$

where l, c, and s represent luminance, contrast, and structure components at scale j, and α_j, β_j, γ_j are weighting factors. MS-SSIM has been shown to better correlate with human perception across varying resolutions and distortion types.

The Information Content Weighted SSIM (IW-SSIM) [79] modifies MS-SSIM by incorporating local information content as a weighting mechanism. The intuition is that regions of an image

with higher information density (e.g., textures or edges) are more perceptually important and should contribute more to the quality score. IW-SSIM applies a multi-scale SSIM computation, weighted by the local mutual information between reference and distorted images:

$$\text{IW-SSIM}(x, y) = \sum_i w_i \cdot \text{SSIM}_i(x, y) \tag{2.15}$$

where w_i is the information content weight at region i, and SSIM_i is the local SSIM score. This approach improves correlation with human judgments, particularly for images with non-uniform content distribution.

The Complex Wavelet SSIM (CW-SSIM) [81] adapts the SSIM formulation to the complex wavelet domain, which is more robust to small geometric distortions such as translations and rotations. Instead of comparing pixel intensities directly, CW-SSIM compares the phase and magnitude of complex wavelet coefficients between reference and distorted images:

$$\text{CW-SSIM}(x, y) = \frac{2\sum_i |c_{x,i}||c_{y,i}| + K}{\sum_i |c_{x,i}|^2 + \sum_i |c_{y,i}|^2 + K} \cdot \frac{2\left|\sum_i c_{x,i} c_{y,i}^*\right| + K}{\sum_i |c_{x,i} c_{y,i}^*| + K} \tag{2.16}$$

where $c_{x,i}$ and $c_{y,i}$ are complex wavelet coefficients at position i, * denotes complex conjugation, and K is a small stabilizing constant. By focusing on phase alignment, CW-SSIM offers enhanced robustness in applications involving misregistration or alignment noise.

The Gradient-based SSIM (G-SSIM) [84] extends SSIM by incorporating image gradient information into the structural comparison. The motivation is that gradients reflect edge strength and direction, which are more sensitive to visual structure than raw intensity values. G-SSIM modifies the structure term in the SSIM formula to include gradient magnitudes:

$$\text{G-SSIM}(x, y) = l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(\nabla x, \nabla y)^\gamma \tag{2.17}$$

where $l(x, y)$ and $c(x, y)$ are the luminance and contrast components as in SSIM, and $s(\nabla x, \nabla y)$ is the structure similarity between the gradients of the two images. The exponents α, β, and γ control the relative importance of each component. G-SSIM offers better sensitivity to blurring and small structural degradations.

The Region-based SSIM (R-SSIM) [85] focuses on evaluating image quality within a central region of interest. This approach assumes that viewers primarily attend to central areas of an image, and thus, distortions in these regions are more perceptually significant:

$$\text{R-SSIM}(x, y) = \text{SSIM}(x|_{\Omega_c}, y|_{\Omega_c}) \tag{2.18}$$

where $\Omega_c \subset \Omega$ be the central region of the image domain Ω. This is often used in screen-centered applications, like the assessment of faces.

The Local SSIM (L-SSIM) [64] extends SSIM by applying it over multiple small patches across the image and averaging the results. This local analysis allows for finer sensitivity to spatially localized distortions that may be overlooked by global quality metrics. Given a window ω_k centered at pixel k, L-SSIM is computed as:

$$\text{L-SSIM}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \text{SSIM}(x|_{\omega_k}, y|_{\omega_k}) \tag{2.19}$$

where K is the total number of windows across the image.

The Color Structural Similarity Index (C-SSIM) [86] extends the traditional SSIM formulation by incorporating chromatic information alongside luminance, contrast, and structure. It proposes a four-component model that quantifies image distortion through a combination of local comparison functions, including a color fidelity term derived from the S-CIELAB color space [95].

$$\text{C-SSIM}(x, y) = l(x, y) \cdot C(x, y) \cdot S(x, y) \cdot \text{Cr}(x, y) \tag{2.20}$$

$$\text{Cr}(x, y) = 1 - \frac{1}{k} \Delta E(x, y) \tag{2.21}$$

where $l(x, y)$ is the luminance similarity, $C(x, y)$ is the contrast similarity, $S(x, y)$ is the structure similarity, and $\text{Cr}(x, y)$ is the color fidelity term. The term $\Delta E(x, y)$ is the color difference between images x and y in the S-CIELAB color space, and k is a scaling constant.

The Weighted Structural Similarity (W-SSIM) [82], [96] is a region-pooled extension of SSIM that emphasizes perceptually important areas of an image. Instead of computing a uniform average of local SSIM values, W-SSIM assigns a larger weight to a region of interest (ROI), such as the image center, and a smaller weight to the surrounding regions. The final score is given by:

$$\text{W-SSIM} = w \cdot \frac{\sum_x M(x)\, S(x)}{\sum_x M(x)} + (1 - w) \cdot \frac{\sum_x (1 - M(x))\, S(x)}{\sum_x (1 - M(x))} \tag{2.22}$$

where $S(x)$ denotes the local SSIM value at pixel x, $M(x)$ is a binary mask equal to 1 in the ROI and 0 elsewhere, and w is the weight assigned to the ROI. In the simple case where the ROI is defined as the central region and $w = 0.7$, this formulation reduces to the implementation described, aligning with earlier region-weighted SSIM variants proposed in the literature.

The Feature Similarity Index (FSIM) [73] assesses image quality by leveraging perceptually meaningful features. It uses phase congruency (PC) as the primary feature, which reflects the significance of local structure independent of luminance and contrast, and gradient magnitude (GM) as a secondary feature to weight salient regions. The local similarity map $S_{\text{FSIM}}(x)$ is defined as:

$$S_{\text{FSIM}}(x) = \frac{S_{\text{PC}}(x) \cdot S_{\text{GM}}(x)}{\max(PC_{\text{ref}}(x), PC_{\text{dist}}(x)) + \epsilon} \tag{2.23}$$

where $S_{\text{PC}}(x)$ and $S_{\text{GM}}(x)$ denote the PC and GM similarity between reference and distorted images at pixel x, and ϵ is a small constant to avoid division by zero. The final FSIM score is computed as the weighted average of $S_{\text{FSIM}}(x)$ over the image, using the PC map as a weighting function.

The FSIM with color (FSIMc) [73] is an extension of FSIM that integrates color information to better evaluate the perceptual quality of color images. In addition to the phase congruency (PC) and gradient magnitude (GM) features used in FSIM, FSIMc includes chrominance similarity components derived from the I and Q channels of the YIQ color space [97]. The final FSIMc score is given by:

$$\text{FSIM}_c = \frac{\sum_{x\in\Omega} S_{\text{FSIM}}(x) \cdot S_I(x) \cdot S_Q(x) \cdot PC_{\max}(x)}{\sum_{x\in\Omega} PC_{\max}(x)} \tag{2.24}$$

where $S_I(x)$ and $S_Q(x)$ are the similarity measures for the I and Q chrominance components, and $PC_{\max}(x)$ is the maximum phase congruency between the reference and distorted images at pixel x. This formulation enhances FSIM's sensitivity to color distortions while preserving its effectiveness in structural fidelity.

The Spectral Residual-based Similarity (SR-SIM) [87] leverages visual saliency derived from the spectral residual of the Fourier transform. The spectral residual highlights regions that are more likely to attract human attention, and SR-SIM incorporates this information to weight structural similarity computations. The final score is defined as:

$$\text{SR-SIM} = \frac{\sum_{x\in\Omega} S(x) \cdot \text{Sal}_{\min}(x)}{\sum_{x\in\Omega} \text{Sal}_{\min}(x)}, \tag{2.25}$$

where $S(x)$ denotes the local structural similarity at pixel x, and $\text{Sal}_{\min}(x)$ is the minimum saliency value between the reference and distorted images at that pixel. By emphasizing salient regions, SR-SIM improves correlation with human perceptual quality judgments while maintaining low computational complexity.

The Edge-based Structural Similarity (ESSIM) [83] improves SSIM by emphasizing structural fidelity in edge regions, which are perceptually more sensitive to distortions. ESSIM first extracts edge maps from the reference and distorted images using edge detectors such as Sobel or Canny, and then computes local SSIM values restricted to these edge pixels. The final score is defined as:

$$\text{ESSIM} = \frac{1}{N} \sum_{x\in\Omega_{\text{edge}}} S(x), \tag{2.26}$$

where $S(x)$ is the local SSIM value at pixel x, Ω_{edge} is the set of edge pixels, and N is the number of pixels in this set. By focusing on edges, ESSIM improves sensitivity to blur and compression artifacts that strongly affect edge integrity.

The Universal Image Quality Index (UQI) [74] models image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. UQI is defined as:

$$\text{UQI} = \frac{4\sigma_{xy}\mu_x\mu_y}{\left(\sigma_x^2 + \sigma_y^2\right)\left(\mu_x^2 + \mu_y^2\right)}, \tag{2.27}$$

where μ_x and μ_y are the mean intensities of the reference and distorted images, σ_x^2 and σ_y^2 are their variances, and σ_{xy} is the covariance between them. The index ranges from -1 to 1, with higher values indicating greater similarity. By integrating correlation, luminance, and contrast terms into a single formula, UQI set the foundation for later structural metrics such as SSIM.

The Spatial Correlation Coefficient (SCC) [80] measures the linear correlation between the reference and distorted images. It is defined as:

$$\text{SCC} = \frac{\sum_{x\in\Omega}(I_r(x) - \mu_r)(I_d(x) - \mu_d)}{\sqrt{\sum_{x\in\Omega}(I_r(x) - \mu_r)^2} \cdot \sqrt{\sum_{x\in\Omega}(I_d(x) - \mu_d)^2}}, \tag{2.28}$$

where $I_r(x)$ and $I_d(x)$ are the pixel intensities of the reference and distorted images at location x, μ_r and μ_d are their mean intensities, and Ω is the image domain. SCC ranges from -1 to 1, with higher values indicating stronger linear correlation and hence greater perceptual similarity. While simple and computationally efficient, SCC is limited because it only captures linear relationships and does not account for structural or perceptual distortions.

Information-theoretic Metrics

Information-theoretic approaches to IQA are grounded in Shannon's theory of communication [98], [99]. They model images as stochastic signals transmitted through a distortion channel, and quality is assessed by measuring how much mutual information between the reference and distorted image is preserved in the presence of perceptual noise. These methods provide a principled view of perceptual fidelity but are typically more computationally demanding than error- or similarity-based metrics.

Notation. Let C_k and D_k denote the wavelet-domain coefficients of the reference and distorted images in subband k, respectively. Natural scene statistics are modeled with Gaussian Scale Mixtures (GSM), with s_k the scale factor at subband k. When modeling the HVS, additive perceptual noise is introduced, denoted by n_k. Mutual information between random variables X and Y conditioned on s_k is written as $I(X;Y \mid s_k)$.

The Information Fidelity Criterion (IFC) [67] quantifies the mutual information shared between the reference and distorted coefficients under the GSM model:

$$\text{IFC} = \sum_k I(C_k; D_k \mid s_k)\,. \tag{2.29}$$

Higher IFC values indicate greater preservation of the statistical information present in the reference image.

The Visual Information Fidelity (VIF) index [71] extends this framework by incorporating perceptual noise to account for the HVS. It is defined as the ratio between the mutual information preserved after distortion and the information available in the reference:

$$\text{VIF} = \frac{\sum_k I(C_k; D_k \mid s_k)}{\sum_k I(C_k; C_k + n_k \mid s_k)}. \tag{2.30}$$

A value of 1 denotes perfect fidelity, while lower values correspond to stronger degradation.

VIFp [71] is a simplified pixel-domain version that avoids multiscale wavelet decomposition. Although less precise, it significantly reduces computational complexity and is useful in real-time applications.

VIFc [71] extends the metric to color images by transforming them into the YUV color space. VIF is computed independently on each channel, and the final score is given by:

$$\text{VIFc} = \tfrac{1}{3}\left(\text{VIF}_Y + \text{VIF}_U + \text{VIF}_V\right). \tag{2.31}$$

This formulation incorporates chromatic information, providing a more complete measure of perceptual fidelity in color imagery.

Deep Learning-Based Metrics

Deep learning-based approaches to IQA exploit feature representations from convolutional neural networks pretrained on large-scale vision tasks. These feature spaces capture high-level perceptual attributes such as texture, structure, and semantics, which align more closely with human visual perception than pixel-wise comparisons [72], [100], [101]. Instead of relying on handcrafted similarity measures, these methods compute distances between deep features extracted from reference and distorted images. Such approaches have shown strong correlation with human opinion scores, particularly for complex and perceptually salient distortions.

Notation. Let $\phi_l(x)$ denote the activation (feature map) of image x at layer l of a pretrained CNN with spatial size $H_l \times W_l$ and C_l channels. Features are channel-wise normalized, denoted by $\hat{\phi}_l(x)$. Channel weights are written as w_l, and similarity functions are applied element-wise over spatial positions (h, w).

Learned Perceptual Image Patch Similarity (LPIPS) [72] measures perceptual similarity by comparing deep features extracted from pretrained networks such as AlexNet, SqueezeNet, or VGG. The metric is defined as:

$$\mathrm{LPIPS}(x, y) = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h,w} \left\| \hat{\phi}_l(x)_{h,w} - \hat{\phi}_l(y)_{h,w} \right\|_2^2. \tag{2.32}$$

Different backbones yield three common variants: LPIPS-Alex, LPIPS-Squeeze, and LPIPS-VGG.

Deep Image Structure and Texture Similarity (DISTS) [66] combines structure and texture similarity in the deep feature space. For each layer l, structure similarity S_l is computed from normalized feature statistics, and texture similarity T_l is obtained from feature covariances. The final score is a convex combination:

$$\mathrm{DISTS}(x, y) = \sum_l \alpha_l S_l(x, y) + \beta_l T_l(x, y), \tag{2.33}$$

where α_l and β_l are learned weights. This design captures both structural fidelity and fine-grained textural consistency.

WaDIQaM [88] employs deep convolutional networks to learn a patch-based representation of image quality. In the full-reference setting, reference and distorted patches are passed through a siamese network [102], and local quality estimates are aggregated using a learned weighting mechanism:

$$\mathrm{WaDIQaM}(x, y) = \sum_p w_p \, q_p(x, y), \tag{2.34}$$

where $q_p(x, y)$ is the predicted quality of patch p and w_p is its learned weight. By learning both local quality prediction and patch importance, WaDIQaM achieves state-of-the-art correlation with human opinion scores in a data-driven manner.

PSM [89] builds on LPIPS to make full-reference perceptual similarity robust to small, imperceptible misalignments. It replaces shift-sensitive elements in the feature extractor with anti-aliased strided convolutions, anti-aliased pooling, reduced strides, and appropriate padding, then aggregates layerwise feature distances like LPIPS. Let $F_l(x)$ be shift-tolerant features at layer l and $\hat{F}_l$ their channel-normalized maps; the score is

$$\mathrm{PSM}(x, y) = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h,w} |\hat{F}_l(x)_{h,w} - \hat{F}_l(y)_{h,w}|_2^2, \tag{2.35}$$

where w_l are learned weights. This preserves perceptual fidelity while reducing spurious changes under 1-2 pixel shifts.

SPIQ [90] is an approach that first learns a quality-aware visual representation via self-supervised pretraining, then fits a lightweight regressor on top with limited MOS labels. Two

augmented views of the same image are encouraged to produce consistent features during pretraining; at inference a regressor maps features to a scalar quality:

$$\mathrm{SPIQ}(x) = g\big(f(x)\big), \tag{2.36}$$

where f is the self-supervised backbone and g is a trained regression head. This decouples representation learning from scarce labels and transfers well across distortions.

2.2.4 No-Reference IQA

NR-IQA estimates perceptual quality directly from a single image without requiring a pristine reference. Three modeling strategies are commonly distinguished. First, classical opinion-unaware methods rely on natural scene statistics or perceptually motivated features and operate without training on opinion scores, providing simple and interpretable baselines [103], [104]. Second, task-driven deep approaches exploit the robustness of embeddings from pretrained networks to derive quality scores without supervision, thereby aligning quality with representational stability rather than explicit perceptual labels [38]. Third, supervised regressors adapt recognition backbones to annotated labels, explicitly learning to map features to continuous quality scores [36], [37].

A complementary distinction concerns domain scope. Generic NR-IQA methods aim to capture perceptual fidelity across natural images, while face-specific approaches (FIQA) target biometric utility, where quality is linked to recognition performance rather than aesthetics. The latter group has gained traction with the emergence of embedding-based models that quantify image suitability for reliable face recognition. Considering both modeling strategies and domain scope enables a nuanced comparison between general-purpose perceptual estimators and specialized, task-aligned FIQA signals.

Table 2.2: Summary of the 5 NR-IQA metrics used

Metric	**Year**	**Category**	**Tested Databases**
NIQE [103]	2012	Classical (opinion-unaware)	LIVE, CSIQ, TID2013, KonIQ-10k
PIQE [104]	2015	Classical (perception-based)	LIVE, CSIQ, TID2013, KADID-10k
FaceQnet [36]	2019	Face-specific (supervised regression)	VGGFace2, BioSecure, FRGC
SER-FIQ [38]	2020	Face-specific (unsupervised)	LFW, VGGFace2, FRGC
MagFace [37]	2021	Face-specific (supervised)	MS-Celeb-1M, IJB-C, LFW

Classical NR-IQA

Classical NR-IQA methods operate without any training. These approaches rely on statistical regularities of natural images and perception-motivated cues to produce a scalar quality score that is easy to compute and interpret. They are attractive baselines because they generalize across content types and avoid dataset-specific supervision, but their assumptions can be violated in narrowly structured domains such as tightly cropped faces.

Natural Image Quality Evaluator (NIQE) [103] 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:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + \epsilon}, \tag{2.37}$$

from which it extracts statistical features associated with Generalized Gaussian Distribution (GGD) and Asymmetric Generalized Gaussian Distribution (AGGD). A Multivariate Gaussia (MVG) model with mean μ_r and covariance Σ_r is estimated from 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:

$$\text{NIQE}(\mathbf{f}_t) = \sqrt{(\mathbf{f}_t - \mu_r)^T \Sigma_r^{-1} (\mathbf{f}_t - \mu_r)}. \tag{2.38}$$

Lower NIQE values indicate statistics closer to the pristine manifold (better quality); higher values indicate stronger deviation.

Perception-based Image Quality Evaluator (PIQE) [104] 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:

$$S_i = \frac{1}{|B_i|} \sum_{(x,y) \in B_i} \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}, \tag{2.39}$$

and noise, estimated by subtracting a Low-Pass Filter (LPF) version of the block:

$$N_i = \text{std}\left(B_i - \text{LPF}(B_i)\right). \tag{2.40}$$

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

$$\text{PIQE} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} w_i \cdot (\alpha S_i + \beta N_i + \gamma C_i), \tag{2.41}$$

where w_i are confidence weights and α, β, γ are scaling factors for sharpness, noise, and contrast, respectively. The reported score is typically scaled to $[0, 100]$, with larger values reflecting worse perceived 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 [16], [17]. 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 [105]. Recent models leverage deep face embeddings extracted from recognition networks [38], [106]. 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 [21]. 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.

Facial Quality Network (FaceQnet) [36] formulates FIQA as regression trained on proxy targets derived from recognition performance rather than MOS. For each subject, an ICAO-compliant reference image $\mathbf{x}_{\text{ref}}$ is selected. A fixed face matcher $F(\cdot,\cdot)$ computes a similarity score $s(\mathbf{x}) = F(\mathbf{x}, \mathbf{x}_{\text{ref}})$ between a candidate image $\mathbf{x}$ and its reference. Scores are normalized to $[0, 1]$ to form targets $\tilde{s}(\mathbf{x})$. A regression network r_θ takes either the image or a recognition embedding as input and predicts $\hat{q} = r_\theta(\mathbf{x})$. Training minimizes

$$\mathcal{L}_{\text{FaceQnet}} = \frac{1}{N} \sum_{n=1}^{N} \left(r_\theta(\mathbf{x}_n) - \tilde{s}(\mathbf{x}_n) \right)^2 . \tag{2.42}$$

At inference, r_θ outputs a quality score $q_{\text{FQ}} \in [0, 1]$. Because targets come from a matcher, FaceQnet quality is explicitly tied to verification utility, but remains dependent on the chosen backbone and matcher used to generate the proxy labels.

Stochastic Embedding Robustness for Face mage 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:

$$\text{sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}, \quad \forall i, j \in \{1, \ldots, T\}, i < j. \tag{2.43}$$

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

$$\text{SER-FIQ} = \frac{1}{\binom{T}{2}} \sum_{i<j} \text{sim}(\mathbf{e}_i, \mathbf{e}_j), \tag{2.44}$$

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.

Magnitude-Aware Face (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:

$$\mathcal{L}_{\text{MagFace}} = -\log \frac{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))}}{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))} + \sum_{j \neq y} e^{s \cdot \cos(\theta_j)}}, \tag{2.45}$$

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.

2.2.5 Fusion-Based IQA Models

Individual FR metrics capture different aspects of perceptual degradation. Error-based metrics respond to amplitude distortions, structure-based metrics track local similarity, and learning-based metrics correlate with semantic distortions. This complementarity motivates fusion: combining heterogeneous scores to approximate MOS more faithfully than any single metric.

Early supervised fusion combined normalized metric scores linearly. [65] introduced the Multi-Method Fusion (MMF) framework, where a regression model learns weights that map a vector of FR-IQA scores to MOS. Non-linear learners extend this principle by modeling interactions between metrics and mitigating collinearity. In a biometric setting, [107] demonstrated that tree ensembles trained on ISO-compliant facial images improve correlation with utility-focused targets. Subsequent studies confirmed that carefully designed fusion outperforms standalone metrics in Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SRCC) against human judgments [21].

When large-scale MOS is scarce, weak supervision turns to proxy signals and relative constraints: one line of work fuses multiple FR-IQA outputs into pseudo-MOS via averaging

or reliability-weighted fusion [108]; another infers labels for unlabeled samples through semi-supervised and positive unlabeled objectives within FR-IQA [109]; a complementary path replaces absolute scores with pairwise or listwise orderings induced by controlled degradations [39]; and, building on these proxy signals, cascaded predictors trained on pseudo-MOS can bootstrap quality estimation when subjective ground truth is limited [110].

Regression Models

Fusion-based IQA learns a mapping from heterogeneous IQA scores to a single perceptual estimate. Let each image have a score vector $\mathbf{x} \in \mathbb{R}^p$ collecting p harmonized metrics and a target $y \in \mathbb{R}$ given by MOS. We first align polarity so that larger scores mean better quality, then optionally standardize each metric. Fusion then solves empirical risk minimization

$$\min_{\theta} \ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \ f(\mathbf{x}_i; \theta)) \quad \text{with} \quad f : \mathbb{R}^p \to \mathbb{R} \tag{2.46}$$

using squared or robust losses. To handle redundancy we may select the top k metrics or use a low-dimensional projection estimated from the score matrix. The learned fusion f combines complementary cues to approximate human judgment more faithfully than any single metric.

There are four useful categories for fusion on tabular metric vectors. Linear models assume an approximately additive mapping and control complexity via shrinkage or sparsity [111], [112], [113], [114]. Kernel methods capture smooth non-linear interactions through implicit feature maps with margin based objectives [115]. Bagging and gradient boosting ensembles model interactions and are robust to heterogeneous scales [116], [117], [118], [119]. Modern boosted-tree frameworks optimize second order objectives with explicit regularization and efficient growth strategies [120], [121], [122].

Linear Regression [111] fits an affine predictor by minimizing squared residuals:

$$\min_{\mathbf{w}, b} \ \sum_{i=1}^{n} \Big(y_i - \mathbf{w}^\top \mathbf{x}_i - b \Big)^2 . \tag{2.47}$$

The normal equations yield a closed form when $\mathbf{X}$ is full rank. Coefficients provide direct attribution to each metric, but estimates can be unstable under collinearity.

Ridge Regression [112] stabilizes linear fusion with L_2 shrinkage

$$\min_{\mathbf{w}, b} \ \sum_{i=1}^{n} \Big(y_i - \mathbf{w}^\top \mathbf{x}_i - b \Big)^2 + \lambda \|\mathbf{w}\|_2^2, \quad \lambda > 0. \tag{2.48}$$

The solution $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$ trades small bias for reduced variance and handles correlated metrics well.

Bayesian Ridge [113] is a probabilistic ridge with Gaussian prior and noise

$$\mathbf{w} \mid \alpha \sim \mathcal{N}(\mathbf{0},\, \alpha^{-1}\mathbf{I}), \quad y_i \mid \mathbf{x}_i, \mathbf{w}, \beta \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b,\, \beta^{-1}), \tag{2.49}$$

$$\mathbf{S} = \left(\alpha \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X}\right)^{-1}, \quad \mathbf{m} = \beta\, \mathbf{S}\, \mathbf{X}^\top \mathbf{y}. \tag{2.50}$$

Posterior $p(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{m}, \mathbf{S})$ gives predictive mean $\hat{y}_* = \mathbf{m}^\top \mathbf{x}_* + b$ and variance $\beta^{-1} + \mathbf{x}_*^\top \mathbf{S} \mathbf{x}_*$, useful to quantify uncertainty in fused quality.

ElasticNet [114] promotes sparse yet stable linear fusion with mixed penalties

$$\min_{\mathbf{w},\, b} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^\top \mathbf{x}_i - b\right)^2 + \alpha \left(\rho \|\mathbf{w}\|_1 + \tfrac{1-\rho}{2} \|\mathbf{w}\|_2^2\right), \tag{2.51}$$

with $\alpha > 0$ and $\rho \in [0, 1]$. It can drop redundant metrics while shrinking the rest to mitigate collinearity.

Support Vector Regression (SVR) [115] fits $f(\mathbf{x})$ with an ϵ-insensitive loss and capacity control

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}^*} \ \tfrac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{2.52}$$

$$\text{s.t. } y_i - f(\mathbf{x}_i) \le \epsilon + \xi_i, \quad f(\mathbf{x}_i) - y_i \le \epsilon + \xi_i^*, \quad \xi_i, \xi_i^* \ge 0, \tag{2.53}$$

with kernelized predictor $f(\mathbf{x}) = \sum_j (\alpha_j - \alpha_j^*) K(\mathbf{x}, \mathbf{x}_j) + b$. Using the Radial Basis Function (RBF) kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$ captures smooth non-linear interactions between metrics.

Random Forest [116] employs bagging regression trees to reduce variance and requires minimal preprocessing. Nodes choose splits that maximize variance reduction

$$\Delta = \mathrm{Var}(S) - \frac{|S_L|}{|S|} \mathrm{Var}(S_L) - \frac{|S_R|}{|S|} \mathrm{Var}(S_R), \tag{2.54}$$

and the fusion prediction averages T trees

$$\hat{y}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}). \tag{2.55}$$

Extremely randomized trees [117] sample split thresholds at random for each feature before selecting the best split, which increases diversity and often lowers variance. Aggregation is identical to Random Forest, with similar variance reduction analysis.

Gradient Boosting [118] employs additive modeling to fit weak learners to pseudo residuals. For squared error, residuals are $r_{im} = y_i - f_{m-1}(\mathbf{x}_i)$. Each stage fits a tree b_m to $\{(\mathbf{x}_i, r_{im})\}$ and updates

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu\, b_m(\mathbf{x}), \tag{2.56}$$

with learning rate $\nu \in (0, 1]$. This captures non-linear metric interactions important in perceptual fusion.

Histogram Gradient Boosting [119] is a histogram variant of gradient boosting that bins each feature into B buckets and accumulates gradient and Hessian statistics per bin to evaluate splits efficiently. It approximates standard boosting while improving speed and memory on many-metric settings.

XGBoost [120] optimizes a second-order regularized objective with per-leaf closed-form weights:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \qquad \Omega(f) = \gamma T + \tfrac{\lambda}{2} \sum_{j=1}^{L} w_j^2. \tag{2.57}$$

With gradients $g_i = \partial_{\hat{y}} \ell$ and Hessians $h_i = \partial_{\hat{y}}^2 \ell$, the optimal leaf weight and split gain are

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \qquad \text{Gain} = \tfrac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma, \tag{2.58}$$

where $G_\cdot = \sum g_i$ and $H_\cdot = \sum h_i$. This yields accurate and regularized fusion on small to medium labeled sets.

LightGBM [121] builds trees leaf-wise (best-first) under depth or leaf constraints and uses gradient-based one-side sampling and exclusive feature bundling, selecting each split by maximizing a second-order regularized gain:

$$\text{Gain} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} - \gamma, \tag{2.59}$$

where $G_\cdot = \sum g_i$ and $H_\cdot = \sum h_i$ are the sums of first and second derivatives over the node and γ penalizes splits.

CatBoost [122] performs ordered boosting with symmetric trees to mitigate prediction shift by computing gradients on permutation prefixes and updating leaves with second-order estimates while optimizing a standard regression loss:

$$f_t(\mathbf{x}) = f_{t-1}(\mathbf{x}) + \nu \, b_t(\mathbf{x}), \tag{2.60}$$

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \tag{2.61}$$

where $G_j = \sum g_i$ and $H_j = \sum h_i$ are (ordered) gradient and Hessian sums over leaf j, ν is the learning rate, and λ is the leaf regularization.

2.3 Image distortion types

Image distortions in IQA are typically organized into operational families that mirror failures across capture, processing, compression, transmission, and display, with taxonomies grounded

in vision science and formal subjective testing protocols [2], [10], [12], [13]. Core classes include lossy compression artifacts from transform quantization and entropy coding, blur from optical defocus or motion, stochastic noise from sensor and electronics, geometric and resampling effects such as aliasing and interpolation blur, and photometric or color shifts induced by exposure, tone mapping, and white-balance errors; real acquisition pipelines often produce authentic mixtures of these mechanisms, especially in mobile, webcam, and surveillance scenarios [3], [14], [15].

2.3.1 Lossy compression

Lossy compression artifacts arise from transform quantization and entropy coding to reduce bitrate; JPEG is dominant for images and induces blocking, ringing, and loss of high frequency detail, while modern codecs vary the artifact profile but share similar roots in quantization and in loop filtering [5], [6]. These artifacts are ubiquitous in social media uploads, cloud messaging, and browser pipelines where repeated encode decode cycles may accumulate damage; the effect is perceived as block boundaries, staircase edges, haloing near strong transitions, and texture washing out fine facial features [1], [10].

2.3.2 Blur

Blur covers optical defocus from finite depth of field and focusing errors, Gaussian like point spread from lens or sensor integration, and motion blur caused by camera or subject movement; all smear edges, reduce acutance, and degrade recognizability of small structures such as eyes or mouth corners [12], [60]. Real world drivers include handheld shooting, long exposures in low light, rolling shutter readout, and surveillance cameras with slow shutters or compression integrated stabilization [3], [45].

2.3.3 Additive noise

Additive noise includes sensor read noise, signal dependent shot noise, thermal noise, and coarse impulse like contamination from faulty pixels or transmission induced bit flips; it manifests as grain, color speckle, or salt and pepper dots that mask texture and create false edges [4], [60]. Noise dominates in dim scenes, small pixel sensors, and high ISO capture and interacts nonlinearly with denoising and sharpening in the camera pipeline [3]. Lab datasets simulate Gaussian, Poisson, and impulse processes across standard deviations or rates [41], [45], [48].

2.3.4 Resampling and resolution changes

Resampling and resolution changes introduce aliasing, moiré, interpolation blur, and haloing when images are downscaled or upscaled without ideal prefiltering; repeated resampling in social platforms or display adaptation pipelines compounds these effects [52], [60]. Edge directed sharpening and unsharp masking can further create overshoot undershoot ripples, perceived as halos around high contrast structures [10].

2.3.5 Photometric and color distortions

Photometric and color distortions cover exposure errors, gamma mis calibration, global and local contrast changes, white balance shifts, channel noise imbalance, and saturation clipping; these arise from auto exposure and auto white balance failures, tone mapping, display pipeline mismatches, and ad hoc enhancement filters [3], [97]. Perceptually, these alter scene realism, skin tone fidelity, and local visibility of detail; benchmark datasets include luminance and contrast change variants, color shifts, and mean intensity shifts to probe metric sensitivity [41], [48].

2.3.6 Transmission and packet-level artifacts

Transmission and packet level artifacts occur when compressed bitstreams traverse unreliable channels; packet loss, bit errors, or jitter map to block corruption, freezing, or tearing in images and video, especially over wireless links or congested networks [7], [8]. Although less common in still images than in streaming video, partial downloads and progressive JPEG decoding can leave visible blocky regions or coarse approximations that degrade perceived quality [5], [43].

2.3.7 Authentic mixtures

Authentic mixtures combine several mechanisms in the wild: for example, a low light smartphone selfie uploaded to a social network may exhibit high ISO noise, heavy denoising and sharpening, aggressive JPEG compression, slight motion blur, and color cast; in surveillance, long exposure blur, compression, and downsampling co occur [40], [42], [53]. These combinations motivate evaluating IQA models beyond single distortions and under realistic capture conditions, using subjective methodologies aligned with standards such as ITU R BT.500 [13] and comprehensive performance studies of metric behavior across datasets and artifact families [1], [2].

2.3.8 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. Classi-

cal 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 **steganography2018adaptive**, 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 [123]. Adaptive methods further refine this by analyzing local characteristics of the image, dynamically choosing to embed regions to minimize perceptual distortion [124].

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.

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) [125] to encode and decode information, we can obtain various results depending on the method used.

StegaStamp

StegaStamp [126] 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.

Code Face

Code Face [127] 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. Code Face proves effective for embedding messages in high-resolution facial portraits, making it suitable for privacy-preserving biometric IDs.

RiemStega

RiemStega [128] 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 [129] 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 [59] (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 to support both supervised learning and independent evaluation. A small, demographically diverse subset of 15 identities, shown in Fig. 3.1 was annotated with MOS through subjective tests. This set was split into a training portion (12 identities, Fig. 3.1a) used to fit the fusion model, and a disjoint test portion (3 identities, Fig. 3.1b) held out for final evaluation. The remaining 87 identities formed the unlabeled set, to which pseudo-MOS were later assigned. Together, the annotated and pseudo-labeled subsets were used to train the NR-IQA model. A summary of these partitions, including the number of subjects and images per subset, is presented in Table 3.1.

The steganographic distortions are visible in Fig. 3.2, which shows examples of distorted facial images from each method. The distortions are applied to the original images, and the resulting 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.

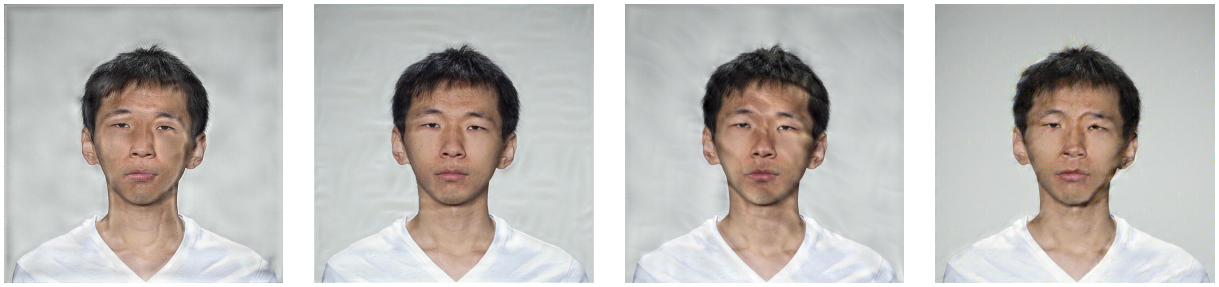
Table 3.1: Partitioning of the steganographically distorted dataset.

Subset	**Subjects**	**Images**	**Purpose**
MOS train set	12	432	Train FR fusion model
MOS test set	3	108	Final evaluation (disjoint)
Pseudo-MOS set	87	3,132	Generate pseudo-MOS labels
NR train set (total)	99	3,564	Train NR model (MOS train + pseudo-MOS)
Total	**102**	**3,672**	Full dataset (all stego levels)



(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.



(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 **ITU-R-BT500** recommendation and adopted the Single Stimulus (SS) method. The test was implemented using a custom Django web application seen in Fig. 3.3. Prior to the test session, participants signed an informed consent form and filled out a registration form providing demographic and environmental information such as age, gender, education, country of origin and ethnicity, and others. After this, observers were shown 10 randomly selected sample images. These examples were designed to help them understand the variation in quality across the dataset.

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 3.2 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 **ITU-R-BT500**, 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. 3.4a 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. 3.4b, 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 partici-



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

Table 3.2: 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

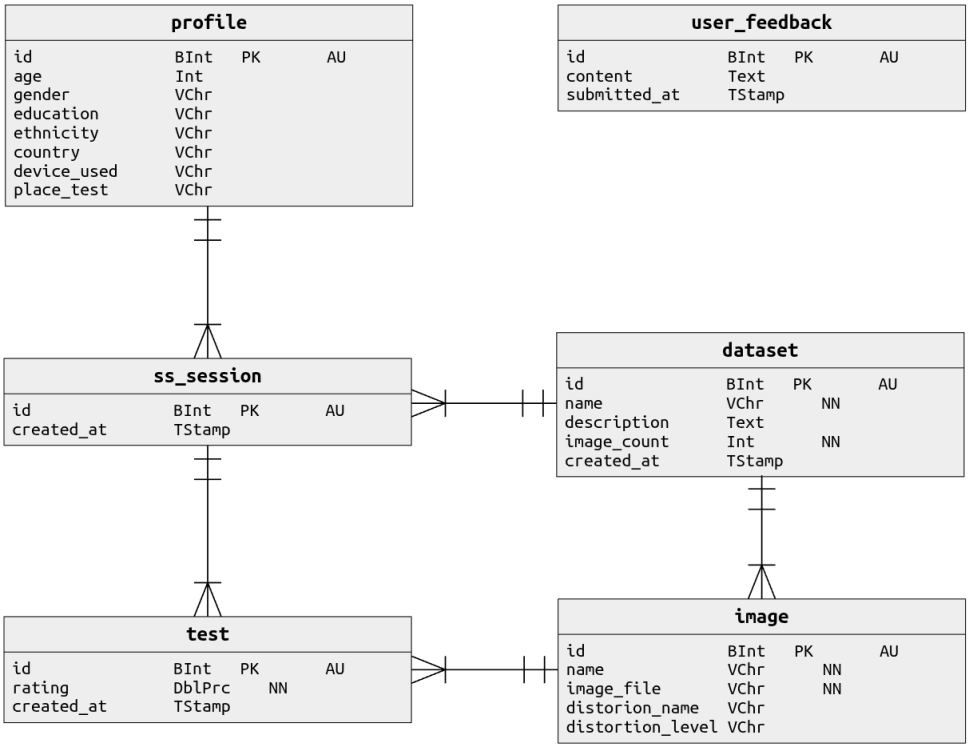
(e) Country

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

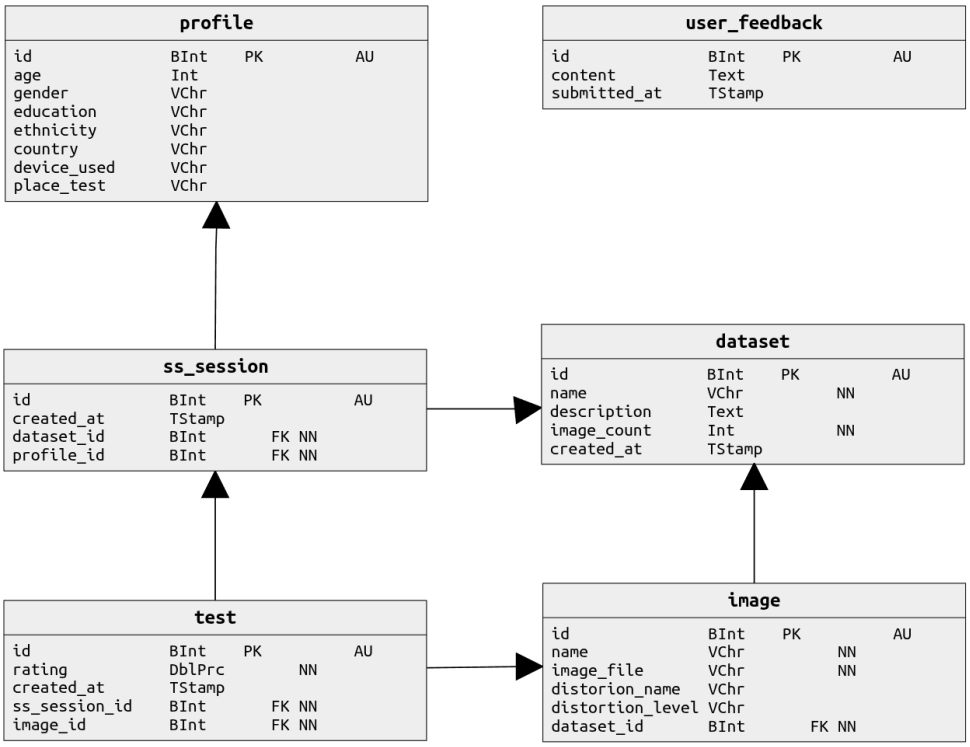
(f) Device Used

Device	Count
Laptop	94
Phone	65
Desktop	42
Other	1

pants. 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 descriptor such as MOS and confidence intervals per image and per distortion level. The relational model also 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 [130], 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 3.3, 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 3.3: ANOVA [131] 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. 3.5. 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 [63].

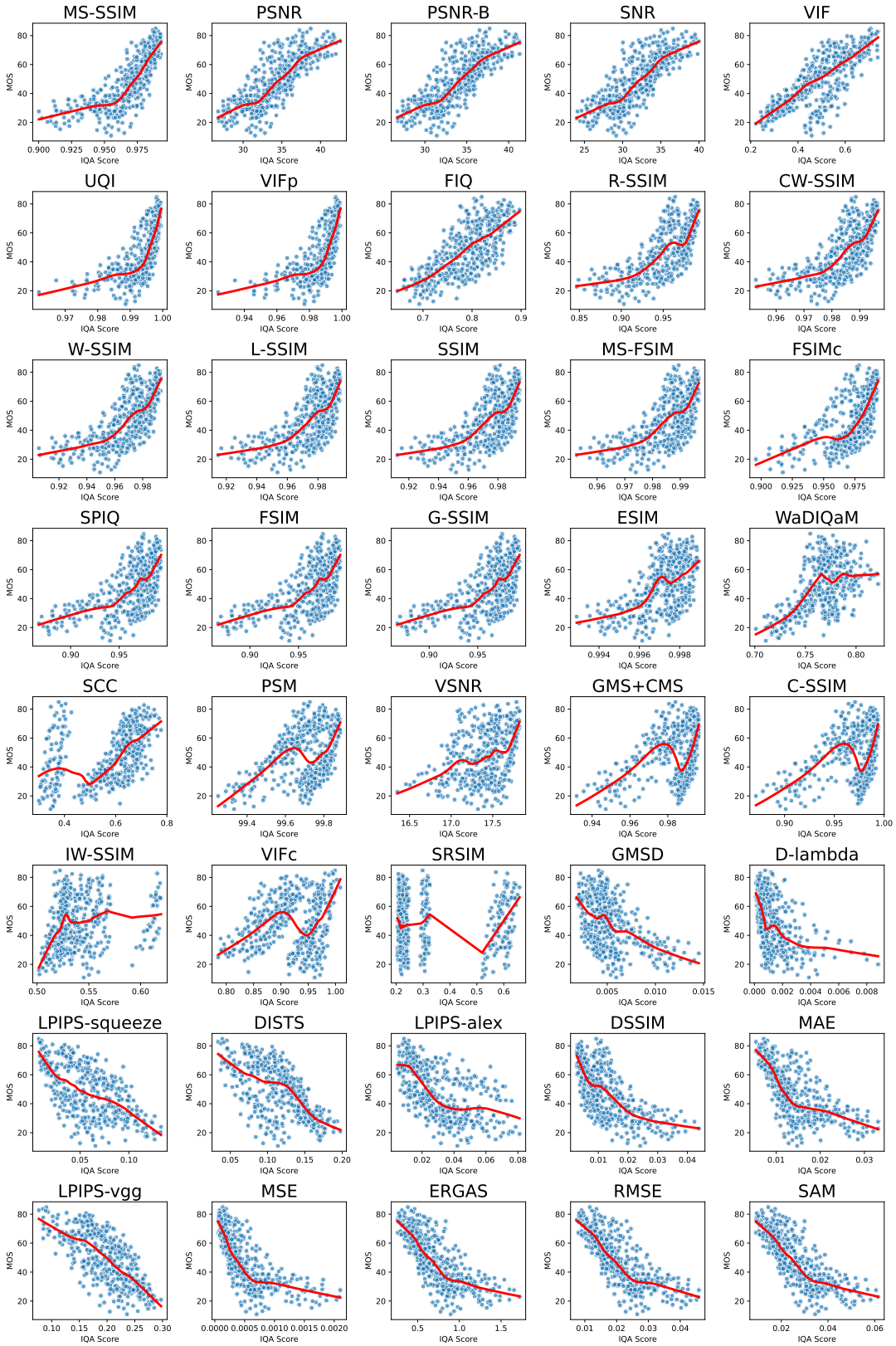


Figure 3.5: Scatter plots illustrating the relationship between MOS and 40 individual full-reference IQA metrics. The red line represents a smoothed local trend curve.

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

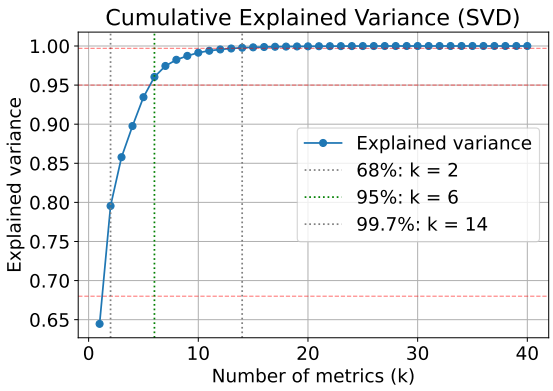
To identify which metrics align best with human perception, we compute both the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-Order Correlation Coefficient (SRCC) [132] which respectively quantify the linearity and monotonicity of the relationship between metric scores and MOS. To determine the appropriate number of metrics to retain for fusion, we applied Singular Value Decomposition (SVD) and the Picard criterion [133]. 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. 3.6a, indicating an optimal dimensionality of $k = 6$.

To validate this truncation point, we examined the Picard plot in Fig. 3.6b, 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.

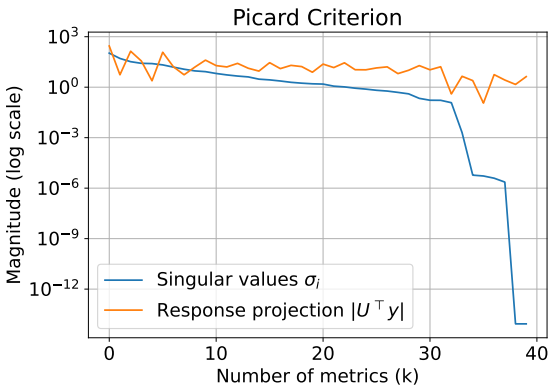
After selecting the top $k = 6$ FR-IQA metrics, we trained a diverse set of supervised regressors to map these features to subjective MOS scores. The models included both linear and non-linear types:

- Linear models: Linear Regression [111], Ridge Regression [112], Bayesian Ridge [113], ElasticNet [114].
- Kernel-based: Support Vector Regression [115] (SVR).
- Ensembles: Random Forest [116], Extra Trees [117], Gradient Boosting [118], HistGradientBoosting [119].
- Boosted Trees: XGBoost[120], LightGBM [121], CatBoost [122].

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 [134] pretrained on ImageNet [135], with its final classification layer replaced by a lightweight multi-layer perceptron (MLP) regressor [113]. 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 [136] (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. 3.7, MS-SSIM [78] (Multiscale SSIM) and UQI **wang2004ssim** (Universal Quality Index) had the largest positive impact on predictions, while PSNR-B **ma2011psnr**, VIF**sheikh2006image**, (PSNR-Blocking), PSNR [137] and SNR **gonzalez2002digital** (Signal-to-Noise Ratio) contributed with moderate

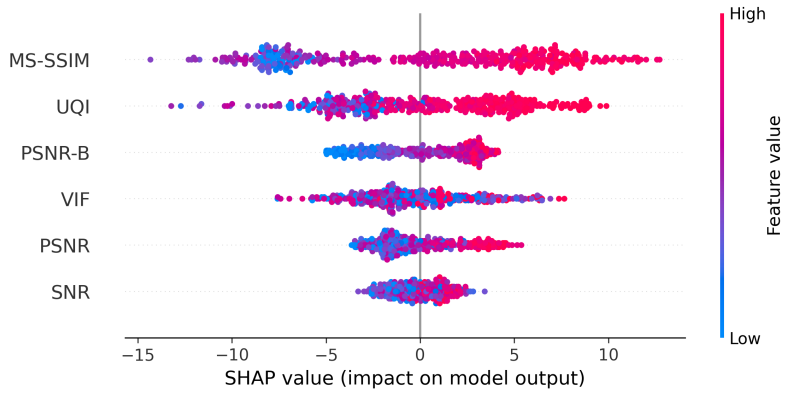


Figure 3.7: SHAP summary plot for $k = 6$, generated from the best CatBoost model.

Table 3.4: 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

influence. The color encoding further highlights how high and low feature values affect the model output.

When comparing the fusion metric with the individual FR metrics, Table 3.4 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.

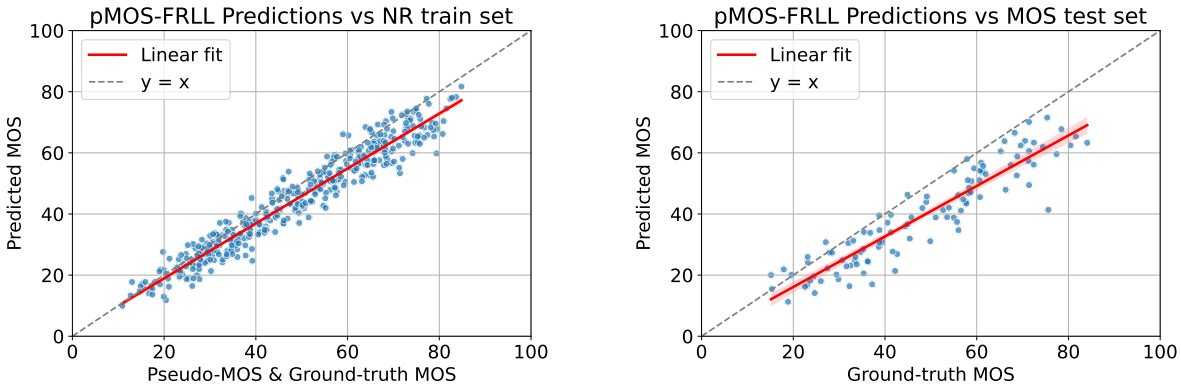
3.7.2 No-Reference Regression Model Performance

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. 3.8 shows two evaluation scenarios:

Table 3.5: 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

in Fig. 3.8(a), the model's predictions are compared against the pseudo and ground-truth MOS used during training, yielding a PLCC of 0.9285, SRCC of 0.9345, MSE of 120.2, and MAE of 9.158. Fig. 3.8(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 [138] and PIQE **piqe2016**, as well as task-specific approaches like SER-FIQ [38] and MagFace [37]. This highlights the robustness of our weakly supervised strategy for NR-IQA on steganographically degraded facial images. Quantitative results are reported in Table 3.5.

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] S. Athar and Z. Wang, "A comprehensive performance evaluation of image quality assessment algorithms," *IEEE Access*, vol. 7, pp. 140 030–140 070, 2019, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2943319.

[2] C. Ma, Z. Shi, Z. Lu, S. Xie, F. Chao, and Y. Sui, *A survey on image quality assessment: Insights, analysis, and future outlook*, 2025. arXiv: 2502.08540 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2502.08540.

[3] V. Mishra, S. Kumar, and N. Shukla, "Image acquisition and techniques to perform image acquisition," *SAMRIDDHI : A Journal of Physical Sciences, Engineering and Technology*, vol. 9, Jul. 2017. DOI: 10.18090/samriddhi.v9i01.8333.

[4] M. Elad, B. Kawar, and G. Vaksman, *Image denoising: The deep learning revolution and beyond – a survey paper –*, 2023. arXiv: 2301.03362 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2301.03362.

[5] M. Rabbani and P. W. Jones, *Digital image compression techniques*. SPIE press, 1991, vol. 7.

[6] J. R. Devadason et al., "A comprehensive survey of image compression methods: From prediction models to advanced techniques," *Multimedia Tools and Applications*, pp. 1–39, 2025.

[7] M. Chandra, D. Agarwal, and A. Bansal, "Image transmission through wireless channel: A review," in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 2016, pp. 1–4. DOI: 10.1109/ICPEICES.2016.7853121.

[8] R. Jayanthi and K. J. Singh, "Image encryption techniques for data transmission in networks: A survey," *International Journal of Advanced Intelligence Paradigms*, vol. 12, no. 1-2, pp. 178–191, 2019.

[9] N. Singhal, A. Kadam, P. Kumar, H. Singh, A. Thakur, and P. Pranay, "Study of recent image restoration techniques: A comprehensive survey," *Jordanian Journal of Computers and Information Technology*, p. 1, Jan. 2025. DOI: 10.5455/jjcit.71-1735034495.

[10] Z. Wang and A. C. Bovik, "Modern image quality assessment," 2006.

[11] K.-H. Thung and P. Raveendran, "A survey of image quality measures," in *2009 International Conference for Technical Postgraduates (TECHPOS)*, 2009, pp. 1–4. DOI: 10.1109/TECHPOS.2009.5412098.

[12] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Science China Information Sciences*, vol. 63, no. 11, p. 211 301, 2020.

[13] international telecommunication union, "Methodologies for the subjective assessment of the quality of television pictures," itu, geneva, switzerland, recommendation itu-r bt.500-15, May 2023.

[14] V. Kamble and K. Bhurchandi, "No-reference image quality assessment algorithms: A survey," *Optik*, vol. 126, no. 11-12, pp. 1090–1097, 2015.

[15] S. Dost, F. Saud, M. Shabbir, M. G. Khan, M. Shahid, and B. Lovstrom, "Reduced reference image and video quality assessments: Review of methods," *EURASIP Journal on Image and Video Processing*, vol. 2022, no. 1, p. 1, 2022.

[16] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–49, Jan. 2022, ISSN: 1557-7341. DOI: 10.1145/3507901. [Online]. Available: http://dx.doi.org/10.1145/3507901.

[17] M. El-Abed, C. Charrier, and C. Rosenberger, "Quality assessment of image-based biometric information," *EURASIP Journal on Image and Video Processing*, vol. 2015, Dec. 2015. DOI: 10.1186/s13640-015-0055-8.

[18] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

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

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

[21] 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.

[22] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, *Face quality estimation and its correlation to demographic and non-demographic bias in face recognition*, 2020. arXiv: 2004.01019 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2004.01019.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] U. Congress, *S.5152 - artificial intelligence civil rights act of 2024*, Accessed: [date of access], 2024. [Online]. Available: https://www.congress.gov/bill/118th-congress/senate-bill/5152.

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

[30] T. Morkel, J. H. Eloff, and M. S. Olivier, "An overview of image steganography.," in *Issa*, vol. 1, 2005, pp. 1–11.

[31] frontex, "Risk analysis for 2023," frontex, Tech. Rep., 2023.

[32] europol, "European union serious and organised crime threat assessment," europol, Tech. Rep., 2022.

[33] interpol, "Stolen and lost travel documents database statistics," interpol, Tech. Rep., 2023.

[34] javelin strategy and research, "Identity fraud study 2023," Tech. Rep., 2023.

[35] frontex, "Technical report on morphing attacks and countermeasures," frontex, Tech. Rep., 2019.

[36] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality assessment for face recognition based on deep learning," in *2019 International Conference on Biometrics (ICB)*, IEEE, 2019, pp. 1–8.

[37] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 225–14 234.

[38] P. Terhorst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5651–5660.

[39] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1040–1049.

[40] 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, 2015.

[41] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2019, pp. 1–3.

[42] V. Hosu, H. 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.

[43] *Laboratory for Image and Video Engineering - The University of Texas at Austin*, https://live.ece.utexas.edu/research/quality/. Accessed: Aug. 6, 2025.

[44] H. Sheikh, "Live image quality assessment database release 2," *http://live. ece. utexas. edu/research/quality*, 2005.

[45] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[46] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of electronic imaging*, vol. 19, no. 1, pp. 011 006–011 006, 2010.

[47] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of modern radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

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

[49] N. Carnec, P. Le Callet, and D. Barba, "Subjective quality assessment irccyn/ivc database," *IRCCyN/IVC*, 2005, `https://ivc.univ-nantes.fr/en/databases`.

[50] F. Yang, W. Lin, X. Zhang, C.-C. J. Kuo, and S. Yao, "Siqad: A database for screen content image quality assessment," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4098–4110, 2015.

[51] H. Lin, L. Jin, J. Wang, W. Lu, S. Li, and K. Zeng, "Kadid-10k: A large-scale artificially distorted iqa database," *Electronic Imaging*, vol. 2019, no. 11, pp. 41–1, 2019, KADID-10k.

[52] K. Ma, Z. Duanmu, Z. Wu, and Z. Wang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, 2016.

[53] Y. Fang, H. Zhu, Z. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3677–3686.

[54] H. Zheng et al., "Pipal: A large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision*, Springer, 2021, pp. 633–651.

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[56] *Rosis: Reflective optics system imaging spectrometer*, `https://earth.esa.int/eogateway/instruments/rosis`, European Space Agency, 2000.

[57] *Airborne visible/infrared imaging spectrometer (aviris)*, https://aviris.jpl.nasa.gov/, NASA Jet Propulsion Laboratory, 1987.

[58] "Hyperspectral digital imagery collection experiment (hydice)," US Naval Research Laboratory, Tech. Rep., 1995, https://apps.dtic.mil/sti/citations/ADA362372.

[59] L. DeBruine and B. Jones, "Face Research Lab London Set," May 2017. DOI: 10.6084/m9.figshare.5047666.v5. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666.

[60] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.

[61] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[62] kai lin, vlad hosu, and dietrich saupe, "Surpassing human perception in image quality assessment," in *ieee international conference on image processing (icip)*, 2020, pp. 916–920. DOI: 10.1109/icip40778.2020.9190936.

[63] 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.

[64] 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.

[65] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 22, Dec. 2012. DOI: 10.1109/TIP.2012.2236343.

[66] 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*, vol. 44, no. 5, pp. 2567–2581, 2022. DOI: 10.1109/TPAMI.2020.3045810.

[67] H. R. Sheikh, A. C. Bovik, and G. De 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, 2005.

[68] 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, 2007.

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

[70] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *2007 IEEE Conference on computer vision and pattern recognition*, Ieee, 2007, pp. 1–8.

[71] 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.

[72] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, *The unreasonable effectiveness of deep features as a perceptual metric*, 2018. arXiv: 1801.03924 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1801.03924.

[73] 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. DOI: 10.1109/TIP.2011.2109730.

[74] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002. DOI: 10.1109/97.995823.

[75] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," in *Summaries of the 4th Annual JPL Airborne Earth Science Workshop*, JPL Publication 92-41, vol. 1, Pasadena, CA, USA: Jet Propulsion Laboratory, 1992, pp. 147–149.

[76] L. Wald, "Quality of high-resolution synthesized images: Is there a simple criterion?" *International Journal of Remote Sensing*, vol. 21, no. 11, pp. 2123–2130, 2000.

[77] 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.

[78] E. P. S. Zhou Wang and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, 2003. DOI: 10.1109/ACSSC.2003.1292216.

[79] 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.

[80] A. Eskicioglu and P. Fisher, "Image quality measures and their performance," *IEEE Transactions on Communications*, vol. 43, no. 12, pp. 2959–2965, 1995. DOI: 10.1109/26.477498.

[81] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009. DOI: 10.1109/TIP.2009.2025923.

[82] c. li chang and a. c. bovik alan c., "Three-component weighted structural similarity index," in *proceedings of spie visual communications and image processing*, vol. 7242, 2009, 72420z. DOI: 10.1117/12.805827.

[83] G.-H. Chen, C.-L. Yang, L.-M. Po, and S.-L. Xie, "Edge-based structural similarity for image quality assessment," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 2, 2006, pp. II–II. DOI: 10.1109/ICASSP.2006.1660497.

[84] G.-h. Chen, C.-l. Yang, and S.-l. Xie, "Gradient-based structural similarity for image quality assessment," in *2006 International Conference on Image Processing*, 2006, pp. 2929–2932. DOI: 10.1109/ICIP.2006.313132.

[85] F. Yan and M. Cao, "An improved method of ssim based on visual regions of interest," in *2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 2015, pp. 1–4. DOI: 10.1109/ICSPCC.2015.7338911.

[86] M. A. Hassan and M. S. Bashraheel, "Color-based structural similarity image quality assessment," in *2017 8th International Conference on Information Technology (ICIT)*, 2017, pp. 691–696. DOI: 10.1109/ICITECH.2017.8079929.

[87] L. Zhang and H. Li, "Sr-sim: A fast and high performance iqa index based on spectral residual," pp. 1473–1476, 2012. DOI: 10.1109/ICIP.2012.6467149.

[88] 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, 2017.

[89] A. Ghildyal and F. Liu, "Shift-tolerant perceptual similarity metric," in *European Conference on Computer Vision*, Springer, 2022, pp. 91–107.

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

[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] 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.

[94] Y. Zhang, "Understanding image fusion quality metrics," in *Proceedings of the 21st ISPRS Congress: Silk Road for Information from Imagery*, ISPRS, Beijing, China, 2008, pp. 90–95.

[95] X. Zhang, D. Silverstein, J. Farrell, and B. Wandell, "Color image quality metric s-cielab and its application on halftone texture visibility," in *Proceedings IEEE COMPCON 97. Digest of Papers*, 1997, pp. 44–48. DOI: 10.1109/CMPCON.1997.584669.

[96] c. li chang and a. c. bovik alan c., "Content-partitioned structural similarity index for image quality assessment," *signal processing: image communication*, vol. 25, no. 7, pp. 517–526, 2010. DOI: 10.1016/j.image.2010.03.004.

[97] national television system committee, "Engineering aspects of television transmission," *proceedings of the ire*, vol. 41, no. 6, pp. 809–819, 1953. DOI: 10.1109/jrproc.1953.274365.

[98] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[99] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[100] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[101] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, Springer, 2016, pp. 694–711.

[102] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[103] 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, 2012.

[104] N. Venkatanath, D. Praneeth, S. C. Sumohana, S. M. Swarup, et al., "Blind image quality evaluation using perception based features," in *2015 twenty first national conference on communications (NCC)*, IEEE, 2015, pp. 1–6.

[105] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002: Evaluation report," 2003.

[106] L. Best-Rowden and A. K. Jain, "Longitudinal study of automatic face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 148–162, 2017.

[107] O. Henniger, B. Fu, and C. Chen, "On the assessment of face image quality based on handcrafted features," in *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2020, pp. 1–5.

[108] L. Lai, J. Chu, and L. Leng, "No-reference image quality assessment based on quality awareness feature and multi-task training," *Journal of Multimedia Information System*, vol. 9, no. 2, pp. 75–86, 2022. DOI: 10.33851/JMIS.2022.9.2.75. [Online]. Available: https://doi.org/10.33851/JMIS.2022.9.2.75.

[109] Y. Cao, Z. Wan, D. Ren, Z. Yan, and W. Zuo, "Incorporating semi-supervised and positive-unlabeled learning for boosting full reference image quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5851–5861.

[110] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, "End-to-end blind image quality prediction with cascaded deep neural network," *IEEE Transactions on image processing*, vol. 29, pp. 7414–7426, 2020.

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

[112] D. W. Marquardt and R. D. Snee, "Ridge regression in practice," *The American Statistician*, vol. 29, no. 1, pp. 3–20, 1975.

[113] C. M. Bishop and M. E. Tipping, "Bayesian regression and," *Advances in Learning Theory: Methods, Models, and Applications*, vol. 190, p. 267, 2003.

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

[115] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

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

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

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

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

[120] T. Chen et al., "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[121] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[122] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, *Catboost: Unbiased boosting with categorical features*, 2019. arXiv: `1706.09516 [cs.LG]`. [Online]. Available: `https://arxiv.org/abs/1706.09516`.

[123] H. Kaur and J. Rani, "A survey on different techniques of steganography," in *MATEC web of conferences*, EDP Sciences, vol. 57, 2016, p. 02 003.

[124] S. Baluja, "Hiding images in plain sight: Deep steganography," *Advances in neural information processing systems*, vol. 30, 2017.

[125] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[126] M. Tancik, B. Mildenhall, and R. Ng, "Stegastamp: Invisible hyperlinks in physical photographs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2117–2126.

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

[128] 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 *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2025, pp. 7572–7581.

[129] F. Shadmand, I. Medvedev, L. Schirmer, J. Marcos, and N. Gonçalves, "Stampone: Addressing frequency balance in printer-proof steganography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4367–4376.

[130] 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.

[131] 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.

[132] 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.

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

[134] 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.

[135] 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.

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

[137] 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.

[138] 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.

Appendix A

Subjective Screening Procedures

This appendix reproduces the statistical screening and post-processing steps recommended by ITU-R BT.500 for subjective image quality assessment (IQA). These procedures were referenced in Chapter 2 but are presented here in full detail for completeness.

A.1 Confidence Intervals

To quantify uncertainty in scores, the sample standard deviation is computed:

$$S_{jkr} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(u_{ijkr} - \bar{u}_{jkr})^2}. \tag{A.1}$$

The 95% confidence interval is then given by:

$$\mathrm{CI}_{95} = 2 \cdot \frac{S_{jkr}}{\sqrt{N}}. \tag{A.2}$$

This interval assumes approximate normality and indicates the range within which the true population mean lies with 95% probability.

A.2 Observer Screening

After score aggregation, observers must be screened to eliminate inconsistent or unreliable participants. ITU-R BT.500 prescribes a kurtosis-based post-screening procedure, especially when the number of subjects is small ($N < 20$) or when non-expert assessors are involved.

For each presentation (j, k, r), the following statistics are computed:

- Mean score $\bar{u}_{jkr}$
- Standard deviation S_{jkr}

- Kurtosis coefficient $\beta_{2,jkr}$, given by:

$$\beta_{2,jkr} = \frac{m_4}{(m_2)^2}, \quad m_x = \frac{1}{N}\sum_{i=1}^{N}\left(u_{ijkr} - \bar{u}_{jkr}\right)^x . \tag{A.3}$$

If $2 \leq \beta_{2,jkr} \leq 4$, the distribution is considered normal; otherwise, it is treated as non-normal.

For each observer i, two counters are defined:

- P_i: number of times the observer's score is an outlier above the acceptable range
- Q_i: number of times the observer's score is an outlier below the acceptable range

A.2.1 Outlier Detection

If the distribution is normal:

- If $u_{ijkr} \geq \bar{u}_{jkr} + 2S_{jkr}$, then $P_i = P_i + 1$
- If $u_{ijkr} \leq \bar{u}_{jkr} - 2S_{jkr}$, then $Q_i = Q_i + 1$

If the distribution is non-normal:

- If $u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20}\, S_{jkr}$, then $P_i = P_i + 1$
- If $u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20}\, S_{jkr}$, then $Q_i = Q_i + 1$

A.2.2 Rejection Rule

After processing all presentations, observer i is rejected if both conditions hold:

- $\frac{P_i+Q_i}{L} > 0.05$
- $\left|\frac{P_i-Q_i}{P_i+Q_i}\right| < 0.3$

where L is the total number of presentations.

A.3 Summary

The post-screening protocol removes observers whose responses deviate excessively from the group distribution. This procedure improves the reliability of subjective scores and is widely used in IQA studies, particularly when the pool of assessors is small or the distortions under study are subtle.