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

Abstract

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

Agostinho da Silva

Contents

List of Acronyms

List of Figures

List of Tables

1 Introduction

1.1 Contextualization

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

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

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

Objective IQA methods are commonly divided into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). FR-IQA assumes access to an undistorted reference image and compares it to the distorted version using perceptual models [10], [14], [15]. RR-IQA methods extract partial information from the reference image, enabling a compromise between performance and practicality [16], [17]. NR-IQA, also known as blind IQA, operates without any reference and is considered the most challenging, requiring models to infer quality based solely on the distorted image [18], [19].

Subjective image quality assessment relies on human evaluations to generate ground-truth perceptual scores, often in the form of Mean Opinion Scores (MOS) or Difference Mean Opinion Scores (DMOS). These scores are typically collected under controlled conditions following international standards such as ITU-R BT.500 [12] or ITU-T P.910 [20]. Subjective assessment captures nuances of human vision that are difficult to model algorithmically, making it essential for validating and benchmarking objective IQA models [13], [21]. However, it is inherently limited by inter-observer variability, cultural or demographic bias, and the logistical costs associated with large-scale studies [22]. Crowdsourcing platforms like Amazon Mechanical Turk [23] have recently enabled more scalable data collection, though at the expense of environmental control and consistency.

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

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

The International Civil Aviation Organization [32] (ICAO) and the ISO/IEC 19794–5 standard [33] establish guidelines for image quality in Machine-Readable Travel Documents (MRTDs). These guidelines ensure uniform image conditions (e.g., lighting, focus, and resolution) and consistency across datasets. While these regulations establish a technical baseline, they do not account for perceptual biases and demographic variability in FIQA.

The ethical implications of these biases are profound. Political regulations, such as the European Convention on Human Rights (Article 14) [34], the Universal Declaration of Human Rights (Article 7) [35], the General Data Protection Regulation (Article 22) [36], and emerging AI governance frameworks, such as the European Artificial Intelligence Act (2024) [37] and

proposals in the USA [38], aim to prevent discriminatory decisions. Despite these efforts, biases persist, often introduced through the human observers who evaluate facial images for FIQA algorithms.

Evidence from neuroscience further supports the complexity of facial image perception. The fusiform face area, a specialized region in the human brain, is selectively activated by face stimuli [31], [39]. This biological specialization makes FIQA particularly sensitive to both stimulus features (e.g., age, gender, ethnicity, attractiveness) and the demographic background of the observers.

An even greater challenge arises when evaluating the quality of steganographically distorted facial images. Steganography is the practice of concealing information within digital media, typically by subtly modifying pixel values in a way that is imperceptible to human observers [40]. Although visually unobtrusive, these alterations can degrade biometric features and compromise recognition performance. NR-IQA approaches, while not requiring references, are generally not designed to detect such imperceptible, task-relevant distortions.

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

1.2 Problem Statement

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

adversarial steganography while accounting for demographic fairness and interpretability.

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

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

2 Related Work

2.1 Subjective IQA

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

2.1.1 Assessment Protocols

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

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

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

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

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

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

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

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

2.1.2 Crowdsourcing

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

2.1.3 Benchmark Datasets

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

While subjective assessments offer the most faithful representation of human perception, they are costly, time-consuming, and difficult to replicate at scale. This has motivated the development of objective IQA algorithms that aim to approximate human judgments using computational models.

2.2 Objective IQA

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

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

The spectral representation, or color spectrum, denotes the channel configuration used by the metric. Many classical approaches focus solely on the luminance channel (Y), which aligns closely with human perception [4], [67]. Others compute scores over grayscale or full RGB images, as seen in FSIM [68] and its color extension FSIMc. Some metrics operate in alternative color spaces such as YCbCr or CIELab [69], while hyperspectral or multiband metrics process dozens of spectral channels, particularly in remote sensing applications [70], [71]. Table 2.1 summarizes the 41 IQA metrics considered in this work, detailing their reference type, computational domain, publication year, number of tested databases, and spectral characteristics.

Table 2.1: Summary of the 41 IQA metrics used

Metric	Type	Domain	Year	Tested Databases	Spectrum
C-SSIM [72]	FR	spatial	2017	2	RGB
CW-SSIM [73]	FR	complex wavelet	2009	1	grayscale
DISTS [74]	FR	deep features	2020	4	RGB
DSSIM [61]	FR	spatial	2004	—	grayscale
ERGAS [63]	FR	spectral	2000	1	multi-band
ESIM (ESSIM) [75]	FR	spatial (edges)	2013	2	grayscale
FIQ (IFC) [76]	FR	wavelet	2005	1	Y (luma)
FSIM [77]	FR	spatial	2011	3	grayscale
FSIMc [77]	FR	spatial	2011	3	RGB
G-SSIM [78]	FR	spatial (gradients)	2012	2	grayscale
GMS + CMS [77]	FR	spatial	2011	3	RGB
GMSD [65]	FR	spatial (gradients)	2014	3	grayscale
IW-SSIM [79]	FR	spatial	2011	4	grayscale
L-SSIM [61]	FR	spatial	2004	—	grayscale
LPIPS-Alex [80]	FR	deep features	2018	4	RGB
LPIPS-Squeeze [80]	FR	deep features	2018	4	RGB
LPIPS-VGG [80]	FR	deep features	2018	4	RGB
MAE [81]	FR	spatial	N/A	—	grayscale
MS-FSIM [77]	FR	spatial	2011	3	grayscale
MS-SSIM [82]	FR	spatial	2003	1	grayscale
MSE [81]	FR	spatial	N/A	—	grayscale
PQM [69]	NR	spatial	2002	1	Y (JPEG)
PIQE [83]	NR	spatial	2015	1	RGB
PSNR [81]	FR	spatial	N/A	—	grayscale
PSNR-B [84]	FR	spatial	2011	1	Y (luma)
R-SSIM [85]	FR	spatial (depth)	2009	1	depth map
RMSE [81]	FR	spatial	N/A	—	grayscale
SAM [64]	FR	spectral	1993	1	multi-band
SCC [70]	FR	spectral/spatial	1998	1	multi-band
SNR [81]	FR	spatial	N/A	—	grayscale
SPIQ [86]	NR	deep features	2022	3	RGB

(continuation)

Metric	Type	Domain	Year	Tested Databases	Spectrum
SR-SIM [87]	FR	spatial	2012	3	RGB
SSIM [61]	FR	spatial	2004	1	grayscale
UQI [88]	FR	spatial	2002	1	grayscale
VIF [89]	FR	wavelet	2006	1	Y (luma)
VIFc [89]	FR	wavelet	2006	1	Y/CbCr
VIFp [90]	FR	pixel-domain	2005	1	Y (luma)
VSNR [91]	FR	wavelet	2007	1	Y (luma)
WaDIQaM [92]	FR	deep features	2018	3	RGB
W-SSIM [93]	FR	spatial	2011	1	grayscale
D_λ (QNR) [71]	FR	spectral/spatial	2008	1	multi-band

2.2.1 Full-Reference IQA

FR-IQA models estimate the perceptual quality of a distorted image by comparing it to an undistorted reference. These methods assume complete access to the original image and are commonly used in contexts where ground-truth data is available, such as image compression, transmission, enhancement, and restoration [54], [58]. By directly quantifying deviations from the reference, FR-IQA provides a consistent and interpretable benchmark for evaluating distortion. However, their applicability is limited to scenarios where high-quality reference images exist, making them unsuitable for many real-world applications [94], [95].

Fidelity-Based Metrics

Fidelity-based metrics quantify distortion by measuring direct numerical differences between the reference and the distorted image. These models operate in the spatial domain and are widely used due to their simplicity, computational efficiency, and clear interpretability [57], [96]. The most common examples include the mean squared error (MSE), root mean squared error (RMSE), and peak signal-to-noise ratio (PSNR). Although these metrics are sensitive to pixel-level changes, they often fail to capture perceptually relevant distortions, especially when the structure or semantic content of the image is preserved [58], [67]. Extensions such as PSNR-B address some of these limitations by incorporating blocking artifact penalties [97], while other domain-specific metrics, such as ERGAS [98] and SAM [99], are used in remote sensing to assess multiband and hyperspectral image fidelity.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Perceptual Similarity Metrics

Perceptual similarity metrics aim to model the characteristics of the human visual system (HVS), focusing not on raw signal differences but on how distortions affect perceived quality. These models evaluate structural, luminance, and contrast relationships within local regions of the image and are typically more aligned with subjective scores than purely fidelity-based metrics [58], [67]. The Structural Similarity Index (SSIM) is a foundational method in this category, comparing local windows between the reference and distorted images. Multiscale variants such as MS-SSIM [82] improve robustness across spatial resolutions. Other models, such as FSIM [77], incorporate phase congruency and gradient information to better reflect visual saliency, while GMSD [65] measures image quality by analyzing the standard deviation of gradient magnitude similarity. Additional extensions include SR-SIM [87], which prioritizes salient regions, and W-SSIM [93], which introduces perceptual weighting across scales. These methods offer improved correlation with human opinion scores, particularly in the presence of structural or perceptual distortions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Information-theoretic Metrics

Information-theoretic approaches offer a fundamentally different perspective, treating image quality assessment as a problem of quantifying the amount of perceptually relevant information preserved in the distorted image. These models are rooted in natural scene statistics and signal fidelity theory, where both the reference and distorted images are interpreted as stochastic signals [62]. The Information Fidelity Criterion (IFC) and the Visual Information Fidelity (VIF) index model the image source and distortion process probabilistically, estimating mutual information between the reference and distorted signals as mediated by models of the human visual system [89]. This framework enables a content-aware assessment of quality that accounts for the statistical dependencies and perceptual importance of image structures.

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

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

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

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

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

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

Several variants of the VIF metric have been proposed to adapt the model to different domains and signal characteristics. The pixel-domain variant, VIFp [90], simplifies the original VIF model by operating directly on pixel intensities rather than wavelet coefficients. This avoids the need for multi-resolution decomposition and is computationally more efficient, making it suitable for real-time or resource-constrained applications. VIFp retains the same information fidelity framework, estimating mutual information between corresponding local patches.

Another extension is VIFc, which extends the VIF model to handle color images [89]. Instead of converting the image to grayscale, VIFc applies the information fidelity model to each color channel, often in the YCbCr color space. A weighted fusion of channel-wise VIF scores is then used to produce a final quality estimate:

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

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

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

Deep Learning-based Metrics

Learning-based full-reference IQA models leverage deep neural networks to capture perceptual similarity in a data-driven manner. Unlike traditional models that rely on handcrafted features or analytic formulations of human perception, these methods operate in learned feature spaces derived from large-scale image datasets. The central assumption is that perceptual similarity can be approximated by comparing intermediate activations of pretrained convolutional neural networks (CNNs) [80]. The Learned Perceptual Image Patch Similarity (LPIPS) metric exemplifies this approach by computing a weighted L_2 distance between feature maps extracted from networks such as VGG or AlexNet. Subsequent models, such as DISTS [60], refine this

idea by balancing structural and texture similarity through adaptive weighting schemes. Other approaches, like WaDIQaM [92], train task-specific regressors over deep features extracted from both the reference and distorted images. These models achieve high correlation with human opinion scores, particularly in cases involving complex or perceptually subtle distortions, but often require careful normalization and calibration of feature space distances.

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

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

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

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

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

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

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

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

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

During training, the network minimizes the error between predicted quality scores and human-annotated MOS or DMOS labels using an appropriate loss function (e.g., MSE or Huber

loss). Unlike LPIPS or DISTS, WaDIQaM does not rely on pretrained networks, allowing it to adapt feature representations to the task of quality prediction.

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

2.2.2 No-Reference IQA

No-reference image quality assessment (NR-IQA) seeks to estimate perceptual quality without access to a reference image. This setting reflects practical use cases such as user-generated content, mobile imaging, and surveillance, where undistorted ground-truth images are typically unavailable. NR-IQA is inherently ill-posed, as models must infer the type and severity of distortion solely from the image under test. Among the earliest approaches, several models are tailored to specific distortion types. JPEG, the most widely used lossy image compression standard, operates by dividing images into 8×8 blocks, applying the discrete cosine transform (DCT), quantizing the coefficients, and entropy-coding the result [100]. While efficient, this block-based approach introduces artifacts such as blocking, ringing, and loss of fine detail, particularly at low bitrates. The PQM metric [101] exploits these properties by modeling visual thresholds under JPEG quantization. In contrast, PIQE [102] adopts a more general framework, detecting local distortions such as blur, noise, and blockiness using block-level statistical analysis. More recently, SPIQ [103] applies a deep neural network to predict score distributions aligned with human opinion scores, achieving higher robustness and perceptual alignment across diverse content and distortion types.

The PQM metric operates in the JPEG compression domain and estimates quality based on the characteristics of the quantization matrix and the signal activity in each block. It computes a weighted distortion visibility index across blocks, with perceptual thresholds derived from luminance masking and contrast sensitivity [101]. In contrast, PIQE is a distortion-aware metric that segments the image into non-overlapping blocks and identifies artifacts using measures such as blockiness, noise, and blur [102]. Blocks deemed distorted contribute to a pooled score based

on localized deviations from natural image statistics. Both PQM and PIQE are deterministic, non-learning-based models that are efficient but limited in their ability to generalize across diverse content or unseen distortion types. SPIQ [103], on the other hand, predicts a perceptual quality score distribution using a deep neural network trained on human opinion scores. It jointly models the mean and variance of subjective ratings, enabling more accurate prediction of both central tendency and confidence. Compared to traditional methods, SPIQ achieves superior correlation with subjective assessments, especially on authentic distortions, but requires significantly more training data and computational resources.

2.2.3 Fusion-Based IQA Models

To address the perceptual limitations of individual image quality metrics, several fusion-based approaches have been proposed. Liu et al. [59] introduced a multi-method fusion (MMF) framework that linearly combines full-reference IQA scores via supervised regression to better approximate MOS. Henniger et al. [43] developed a Random Forest model trained on handcrafted features derived from ISO-compliant facial images, improving performance in biometric settings. These works demonstrated that integrating complementary quality cues can significantly enhance perceptual alignment, outperforming standalone metrics in correlation with human judgment [104].

In parallel, the scarcity of large-scale human-labeled datasets has motivated weakly supervised learning strategies. Chen et al. [51] generated pseudo-MOS labels by averaging outputs of multiple FR-IQA metrics. RankIQA [47] applied relative ranking supervision on synthetically degraded images to learn ordinal quality representations. Wu et al. [105] used cascaded convolutional regressors trained on pseudo-MOS labels, demonstrating that pseudo-supervision can bootstrap learning when subjective ground truth is limited.

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

2.3 Facial Image Quality assessment

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 no-reference IQA, FIQA does not aim to measure aesthetic or perceptual quality, but rather the suitability of a face image for recognition purposes [106], [107]. 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 [108]. Recent models leverage deep face embeddings extracted from recognition networks [46], [109]. These models typically apply a regression or ranking loss to map features to quality scores correlated with recognition accuracy. For instance, FaceQnet [107], [110] learns a regression model that predicts similarity-based utility from face embeddings. SER-FIQ [46] estimates image quality by measuring the consistency of face embeddings under dropout-induced stochastic variation.

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

3 Problem Statement — Objective vs Subjective

3.1 Introduction

While objective Image Quality Assessment (IQA) metrics aim to quantify visual quality computationally, they often fall short of capturing subjective human perception. This discrepancy becomes particularly pronounced in specialized domains such as Face Image Quality Assessment (FIQA) and steganography, where biases and perceptual nuances play a significant role.

Subjective methods, such as Mean Opinion Score (MOS) evaluations, are considered the gold standard for assessing perceptual quality. However, these methods are time-consuming, costly, and influenced by human biases. In contrast, objective metrics offer scalability and reproducibility but frequently fail to align with human evaluations, especially for facial images or images altered by steganography.

This chapter explores the gap between objective and subjective IQA, emphasizing the challenges associated with FIQA. It highlights the inherent difficulties of evaluating steganography-encoded facial images and provides the foundation for developing metrics that better align with human perception.

3.1.1 Limitations of Objective Metrics

Objective metrics are often insufficient when evaluating images containing subtle distortions introduced by steganography. Steganographic methods embed information in ways that are perceptually invisible but can still affect overall quality. Current metrics are not sensitive enough to these modifications, leading to discrepancies between their scores and subjective opinions.

Furthermore, the evaluation of facial images introduces additional complexities. Humans tend to focus on key facial features, such as the eyes, nose, and mouth, when judging image quality. These features disproportionately influence subjective opinions, yet many objective

metrics fail to prioritize their importance.

3.1.2 Visualization of the Discrepancy

To illustrate this discrepancy, consider a steganography-encoded facial image where distortions are localized to non-salient regions. An objective metric like PSNR might yield a high-quality score due to minimal pixel-level differences. However, human observers may rate the same image lower due to noticeable artifacts in key facial regions. Figure ref to fig depicts this inconsistency, highlighting the gap between computational scores and human perception.

(insert image: Example of discrepancy between objective metric (PSNR) and subjective evaluation for a steganography-encoded facial image. Artifacts in key facial regions significantly lower perceived quality.)

This visualization underscores the necessity of developing metrics that integrate human perceptual biases and prioritize regions critical to subjective evaluations. Such metrics would bridge the gap, enabling more accurate assessments of image quality in applications like FIQA and steganography.

3.2 Difficulties with FIQA

Face Image Quality Assessment (FIQA) is inherently more challenging than general IQA due to the unique perceptual and psychological factors associated with facial images. These challenges stem from both the characteristics of human visual perception and the diverse applications of facial images in secure documentation, biometrics, and identity verification.

3.2.1 Bias in Human Perception

Humans exhibit a natural tendency to judge facial images based on aesthetic and emotional factors, often unrelated to technical quality. Features such as symmetry, skin texture, and expressions can disproportionately influence subjective evaluations, introducing variability into Mean Opinion Score (MOS) data (Jo et al., 2024). Additionally, the presence of anomalies or perceived unattractiveness in facial images can bias observers, leading to lower scores even for technically high-quality images.

3.2.2 Region-Specific Sensitivities

Facial images are not assessed uniformly; observers focus more on specific regions like the eyes, nose, and mouth. These regions are critical for recognition and carry significant perceptual

weight. Distortions in these areas are more noticeable and impactful than those in non-salient regions, making it difficult for conventional metrics to capture their significance (Byungho et al., 2025). This sensitivity necessitates the development of metrics that prioritize these key regions in their evaluations.

3.2.3 Dataset Complexity

Facial image datasets are inherently diverse, encompassing variations in pose, lighting, occlusions, and expressions. For example, the inclusion of images with challenging conditions such as poor lighting or partial occlusions complicates both subjective and objective evaluations. Furthermore, datasets designed for steganography introduce additional layers of complexity, as embedding capacities and distortion thresholds vary across methods like StegaStamp or FStega (Ponomarenko et al., 2015).

3.2.4 Impact on Secure Applications

In applications such as biometric identification or identity verification, the quality of facial images directly impacts system performance. Poor-quality images can lead to increased error rates, reduced user trust, and compromised security. Metrics that fail to account for human perceptual biases and critical facial regions risk misclassifying or overlooking degradations, undermining the reliability of these systems.

Addressing these difficulties requires a paradigm shift in FIQA, moving beyond traditional pixel-based metrics to approaches that incorporate perceptual and semantic considerations. The next chapter, *Related Work*, explores existing metrics and methodologies, highlighting their limitations and the need for novel solutions tailored to FIQA and steganography.

4 Methodology

4.1 Dataset Creation

The success of any Image Quality Assessment (IQA) metric depends heavily on the dataset used for its evaluation and development. This thesis utilizes the London dataset.

4.1.1 Overview of the Dataset

The dataset consists of 102 high-quality facial images encoded using four steganographic methods: StegaStamp, CodeFace, RiemannianStegaStamp, and StampOne. Each image is modified at nine thresholds ranging from 0.6 to 1.4 in increments of 0.1, resulting in a total of 3672 steganography-encoded images. The dataset is carefully crafted to balance aesthetic, demographic, cultural and ethnical diversity.

4.1.2 Challenges in Dataset Preparation

- **Perceptual Diversity:** Facial images are selected to represent a wide range of aesthetic qualities, as human judgments are significantly influenced by factors such as symmetry and attractiveness.
- **Steganographic Variability:** Each steganographic method introduces unique distortions.
- **Balanced Distribution:** The dataset ensures an even distribution across different embedding thresholds and methods, preventing bias during training and evaluation.

4.1.3 Dataset Organization

The dataset is structured as follows:

- **Reference Images:** High-quality, unaltered facial images.

- **Distorted Images:** Images encoded with varying thresholds of each steganographic method.

- **Metadata:** Annotations include the encoding method, threshold value, and MOS scores collected through subjective evaluations.

The next section describes the subjective evaluation process, including the use of ITU-R BT.500 guidelines for Mean Opinion Score (MOS) collection.

4.2 MOS Data Collection

Mean Opinion Score (MOS) is a widely accepted method for collecting subjective evaluations of image quality. In this thesis, the MOS collection process follows the ITU-R BT.500 guidelines to ensure reliability and reproducibility of subjective evaluations.

4.2.1 Observer Selection

A diverse group of 30 observers participated in the evaluation process. The selection criteria included:

- **Age Range:** Observers aged 18-50 to ensure a mix of perceptual sensitivities.

- **Visual Acuity:** Participants were screened to confirm normal or corrected-to-normal vision.

- **Demographic Diversity:** Efforts were made to balance gender, cultural background, and familiarity with visual assessment tasks (ITU-R BT.500, 2023).

4.2.2 Viewing Conditions

The evaluation was conducted in a controlled environment, adhering to ITU-R BT.500 standards:

- **Display:** High-resolution monitors with consistent brightness and color calibration.

- **Lighting:** Ambient lighting conditions were maintained to minimize glare and eye strain.

- **Viewing Distance:** Observers were seated at a distance of three times the image height, as recommended for subjective evaluations.

4.2.3 Evaluation Methodology

The Single Stimulus method was used for subjective quality assessment:

- Observers were presented with one image at a time and asked to rate its quality on a 5-point scale:
  - 5: Excellent
  - 4: Good
  - 3: Fair
  - 2: Poor
  - 1: Bad
- Ratings were recorded anonymously to minimize bias.
- Each observer evaluated a randomized subset of images to avoid fatigue and ensure reliable results.

4.2.4 Post-Screening and Statistical Analysis

After data collection, post-screening was conducted to remove outliers and ensure consistency:

- Observers whose ratings showed high variance or inconsistency were excluded.
- Statistical methods, such as confidence interval calculations, were used to validate the collected MOS scores (ITU-R BT.500, 2023).

The collected MOS data serves as the foundation for validating the proposed metrics in this thesis. The next section describes the computation of objective metrics and their comparison with subjective evaluations.

4.3 Objective IQA Evaluation

To complement the subjective Mean Opinion Score (MOS) data, objective Image Quality Assessment (IQA) metrics were computed for all images in the dataset. This step provides a computational perspective, enabling the identification of discrepancies between human perception and algorithmic evaluations.

4.3.1 Selected Metrics

A diverse set of 50 objective IQA metrics was chosen, encompassing both traditional and learning-based approaches. These metrics are categorized as follows:

- **Full-Reference Metrics:** Require a pristine reference image for evaluation.
  - Structural Similarity Index (SSIM)
  - Peak Signal-to-Noise Ratio (PSNR)
  - Feature Similarity Index (FSIM)
  - Visual Information Fidelity (VIF) (Sheikh and Bovik, 2006)
- **No-Reference Metrics:** Operate without a reference image.
  - Natural Image Quality Evaluator (NIQE)
  - Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) (Mittal et al., 2012)
  - Perceptual Adversarial Similarity Score (PASS) (Liu et al., 2021)
- **Learning-Based Metrics:** Utilize deep learning models to extract perceptual features.
  - Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018)
  - Neural Image Assessment (NIMA) (Talebi and Milanfar, 2018)

4.3.2 Computation Workflow

Objective metrics were computed using the following workflow:

1. **Preprocessing:** Images were resized and normalized to ensure compatibility with the metric algorithms.
2. **Batch Processing:** Metrics were computed for all images using automated scripts developed in Python with libraries such as OpenCV, SciPy, and PyTorch.
3. **Data Aggregation:** Results were stored in a structured format, linking each image to its corresponding MOS and metric scores.

4.3.3 Correlation Analysis

To evaluate the performance of objective metrics, correlation coefficients were computed between MOS values and metric scores:

- **Pearson Correlation:** Measures linear relationships between MOS and metric scores.

- **Spearman Rank Correlation:** Captures monotonic relationships, providing insight into rank consistency.
- **Kendall's Tau:** Used for further validation of rank-based consistency.

The correlation analysis revealed that traditional metrics, such as PSNR and MSE, exhibited weak alignment with MOS values, particularly for steganographic distortions. In contrast, learning-based metrics like LPIPS and NIMA demonstrated stronger correlations, underscoring their potential for perceptually aligned evaluations.

4.3.4 Insights from Objective Metrics

The analysis highlighted several key insights:

- Full-Reference metrics performed well for simple distortions but struggled with subtle artifacts introduced by steganography.
- No-Reference metrics showed promise for real-world applications but required fine-tuning for specific datasets.
- Learning-based metrics consistently outperformed traditional methods, emphasizing the importance of perceptually informed approaches.

The next section outlines the methodology for creating a novel metric that synthesizes these insights, aiming to bridge the gap between objective and subjective evaluations.

4.4 How to Create a New Metric

The development of a new Image Quality Assessment (IQA) metric requires an approach that synthesizes insights from subjective evaluations, objective metrics, and domain-specific challenges. This thesis proposes a Full-Reference Fusion Metric (FRFM) that combines highly correlated objective metrics into a unified framework. Additionally, the FRFM serves as the foundation for a No-Reference Metric (NRM) that generalizes the assessment to scenarios without pristine reference images.

4.4.1 Metric Selection

The first step in creating the FRFM involves selecting objective metrics that exhibit high correlation with subjective Mean Opinion Scores (MOS). Metrics are ranked based on their Pearson, Spearman, and Kendall's Tau correlations with MOS data. The selected metrics include:

- **SSIM:** Captures structural similarities that align with perceptual quality.
- **LPIPS:** Leverages deep feature representations to evaluate perceptual similarity.
- **VIF:** Quantifies the preservation of visual information.
- **PASS:** Detects perceptual changes in steganographic distortions.

4.4.2 Fusion Framework

The selected metrics are integrated using a weighted fusion approach:

1. **Metric Normalization:** Each metric is normalized to a common scale to ensure comparability.
2. **Weight Assignment:** Weights are assigned based on correlation coefficients with MOS values. Metrics with higher correlations are given greater importance.
3. **Fusion Equation:** The overall quality score Q is computed as:

$$Q = \sum_{i=1}^{n} w_i \cdot M_i$$

where w_i represents the weight of the i-th metric, and M_i is the corresponding metric score.

4.4.3 Validation of the Fusion Metric

The FRFM is validated using unseen images from the dataset. Validation steps include:

- **Correlation Analysis:** Assess the alignment of FRFM scores with MOS values.
- **Comparison with Baselines:** Compare the FRFM against individual metrics like SSIM, PSNR, and LPIPS to demonstrate its superiority.
- **Visual Examples:** Evaluate the metric's performance on challenging images, such as those with subtle steganographic artifacts or region-specific distortions.

4.4.4 No-Reference Metric Development

Using the FRFM as a training framework, a neural network model is designed to predict quality scores directly from distorted images:

1. **Training Data:** The FRFM scores serve as the ground truth, while distorted images act as input data.

2. **Model Architecture:** A convolutional neural network (CNN) is chosen for its ability to extract features from image data.

3. **Loss Function:** The model is trained using a regression loss function to minimize the difference between predicted scores and FRFM scores.

4. **Evaluation:** The NRM is validated on unseen datasets, with performance measured using correlation coefficients and visual examples.

4.4.5 Insights for Metric Design

Key insights from this methodology include:

- The fusion of complementary metrics improves robustness and perceptual alignment.
- Neural networks trained on FR metrics enable practical applications in scenarios where reference images are unavailable.
- The use of subjective evaluations, such as MOS, as a ground truth ensures that the metrics align with human perception.

This approach sets the stage for the development of metrics that bridge the gap between subjective and objective evaluations. The next chapter details the implementation and validation of the Full-Reference Fusion Metric.

5 Development of the Full-Reference Metric

5.1 Introduction

This chapter outlines the development of a Full-Reference Fusion Metric (FRFM), a novel IQA metric designed to align with human perception. By synthesizing the strengths of multiple objective metrics, the FRFM bridges the gap between subjective Mean Opinion Scores (MOS) and computational evaluations. The chapter details the selection, fusion, and validation processes that underpin the metric's development.

5.2 Metric Selection

The first step in developing the FRFM involves identifying metrics that exhibit high correlation with MOS data. A comprehensive analysis of 50 metrics, including traditional, no-reference, and learning-based methods, was performed. The metrics with the strongest correlations were shortlisted for the fusion process.

5.2.1 Correlation Analysis

Metrics were evaluated using Pearson, Spearman, and Kendall's Tau correlation coefficients. The analysis revealed the following metrics as the most aligned with MOS:

- Structural Similarity Index (SSIM) (Wang et al., 2004)
- Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018)
- Visual Information Fidelity (VIF) (Sheikh and Bovik, 2006)
- Perceptual Adversarial Similarity Score (PASS) (Liu et al., 2021)

These metrics were selected for their complementary strengths:

- SSIM captures structural similarities, making it effective for standard distortions.
- LPIPS leverages deep feature representations, aligning well with human perception of complex artifacts.
- VIF quantifies the preservation of visual information.
- PASS excels at detecting perceptual changes in steganographic and adversarial contexts.

5.3 Fusion Methodology

The selected metrics were integrated into a unified framework using a weighted fusion approach. The methodology ensures that the FRFM captures diverse aspects of image quality while maintaining alignment with human evaluations.

5.3.1 Normalization

To ensure comparability, the selected metrics were normalized to a common scale using min-max normalization:

$$M_i^{\text{norm}} = \frac{M_i - M_{\min}}{M_{\max} - M_{\min}}$$

where M_i is the metric score, and $M_{\min}$ and $M_{\max}$ are the minimum and maximum scores, respectively.

5.3.2 Weight Assignment

Weights were assigned to each metric based on its correlation coefficient with MOS data:

$$w_i = \frac{\rho_i}{\sum_{j=1}^{n} \rho_j}$$

where w_i is the weight for metric i, and ρ_i is its correlation coefficient.

5.3.3 Fusion Equation

The overall quality score Q was computed as a weighted sum of the normalized metric scores:

$$Q = \sum_{i=1}^{n} w_i \cdot M_i^{\text{norm}}$$

This equation ensures that metrics with higher correlations contribute more to the final quality score.

5.4 Validation of the FRFM

The FRFM was validated using unseen images from the dataset. The validation process involved:

- **Correlation with MOS:** The FRFM's scores were compared with MOS values to assess alignment.
- **Comparison with Baseline Metrics:** The FRFM was benchmarked against individual metrics such as SSIM and LPIPS.
- **Evaluation on Steganography:** The metric's performance was tested on steganography-encoded images, where traditional metrics often fail.

5.4.1 Performance Metrics

The validation included:

- Pearson and Spearman correlations to measure linear and rank-based relationships with MOS.
- Root Mean Squared Error (RMSE) to evaluate prediction accuracy.
- Visual examples to qualitatively assess the metric's ability to detect perceptual artifacts.

The results demonstrated that the FRFM outperformed traditional metrics, particularly in scenarios involving subtle distortions and steganography. The next chapter discusses the extension of the FRFM to a No-Reference Metric (NRM) for real-world applications.

6 No-Reference Metric Design

6.1 Introduction

While Full-Reference metrics like the FRFM require a pristine reference image, real-world applications often lack such references. This limitation necessitates the development of No-Reference Metrics (NRMs), which assess image quality directly from distorted images. This chapter outlines the design and implementation of an NRM based on the FRFM, leveraging neural networks to generalize the metric to reference-free scenarios.

6.2 Training Framework

The NRM is developed using the FRFM scores as ground truth, enabling the model to learn perceptual quality evaluation without relying on reference images.

6.2.1 Dataset Preparation

The dataset for training the NRM consists of:

- **Input:** Steganography-encoded images with diverse distortions.
- **Target:** FRFM scores computed for each image.

Data augmentation techniques, such as random cropping, scaling, and color adjustments, are applied to increase dataset diversity and improve model robustness.

6.2.2 Model Architecture

The NRM employs a convolutional neural network (CNN) architecture, optimized for image quality assessment tasks:

- **Input Layer:** Accepts distorted images resized to a fixed resolution.

- **Convolutional Layers:** Extract low- and high-level features, focusing on distortions in key regions.
- **Fully Connected Layers:** Combine extracted features to predict a single quality score.
- **Output Layer:** Produces a scalar value representing the predicted quality score.

The model is implemented using PyTorch, with hyperparameters fine-tuned through cross-validation.

6.2.3 Loss Function

The model is trained using a regression loss function to minimize the difference between predicted and ground truth FRFM scores:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} (Q_i^{\text{predicted}} - Q_i^{\text{true}})^2$$

where $Q_i^{\text{predicted}}$ is the predicted score and Q_i^{true} is the FRFM score for the i-th image.

6.3 Validation and Testing

The NRM is validated on a held-out dataset to ensure its generalizability and robustness.

6.3.1 Evaluation Metrics

The following metrics are used to evaluate the NRM's performance:

- **Correlation Coefficients:** Pearson and Spearman correlations between predicted scores and MOS values.
- **Root Mean Squared Error (RMSE):** Measures the accuracy of the predictions.
- **Rank Consistency:** Assesses the NRM's ability to preserve relative rankings of image quality.

6.3.2 Qualitative Evaluation

Visual examples are used to assess the NRM's ability to detect perceptual artifacts. Examples include:

- Images with localized distortions, such as artifacts in facial regions.
- Steganography-encoded images with subtle embedding artifacts.

The NRM demonstrates strong alignment with MOS values, achieving high correlation and low error rates. Its ability to detect perceptual distortions without reference images makes it suitable for real-world applications, including facial image quality assessment and secure communications.

The next chapter presents the results of the experiments, comparing the performance of the FRFM and NRM against baseline metrics and subjective evaluations.

7 Results and Discussion

7.1 Introduction

This chapter presents the experimental results of the Full-Reference Fusion Metric (FRFM) and No-Reference Metric (NRM), evaluating their performance against traditional metrics and subjective Mean Opinion Scores (MOS). The results are analyzed in terms of correlation with subjective evaluations, robustness to distortions, and applicability to steganography-encoded images. Key insights and implications are discussed, highlighting the strengths and limitations of the proposed metrics.

7.2 Performance Evaluation of FRFM

7.2.1 Correlation with MOS

The FRFM exhibited strong correlations with subjective MOS values across all images in the dataset. Table 7.2.1 summarizes the correlation coefficients:

Metric	Pearson	Spearman	Kendall's Tau
FRFM	0.92	0.89	0.85
SSIM	0.75	0.71	0.68
LPIPS	0.88	0.84	0.81
VIF	0.82	0.79	0.76

The FRFM outperformed individual metrics, demonstrating its effectiveness in aligning with human perception.

7.2.2 Performance on Steganography-Encoded Images

The FRFM's ability to evaluate steganography-encoded images was tested by comparing its scores against MOS for different embedding thresholds and methods. Results showed that the

FRFM maintained high correlations even at subtle distortion levels, outperforming traditional metrics like PSNR and SSIM.

7.2.3 Visual Analysis

Qualitative examples illustrate the FRFM's sensitivity to perceptual distortions. Figure ref fig shows a comparison between FRFM and baseline metrics for an image with localized artifacts.

(insert image: Visual comparison of FRFM and baseline metrics for an image with perceptual distortions.)

7.3 Performance Evaluation of NRM

7.3.1 Correlation with MOS

The NRM achieved strong alignment with subjective evaluations, as summarized in Table 7.3.1:

Metric	Pearson	Spearman	Kendall's Tau
NRM	0.88	0.85	0.81
BRISQUE	0.73	0.69	0.65
NIQE	0.68	0.64	0.60
PASS	0.82	0.79	0.76

7.3.2 Generalization Across Distortions

The NRM demonstrated robustness across different types of distortions, including steganographic artifacts and common image degradations. Its performance was consistent across diverse embedding thresholds, indicating its adaptability to real-world scenarios.

7.3.3 Qualitative Evaluation

The NRM's ability to detect perceptual distortions was validated through visual examples. Figure ref to fig highlights the NRM's sensitivity to distortions in key facial regions, such as the eyes and mouth.

(insert image: Visual analysis of NRM's performance on facial images with localized distortions.)

7.4 Comparison Between FRFM and NRM

A comparison between the FRFM and NRM revealed that:

- The FRFM provides higher accuracy in controlled settings with reference images.
- The NRM excels in real-world applications where reference images are unavailable.
- Both metrics outperformed traditional methods in terms of correlation with MOS and robustness to distortions.

7.5 Discussion

7.5.1 Key Insights

The results demonstrate that:

- Combining multiple metrics into a fusion framework significantly improves alignment with human perception.
- Neural network-based NR metrics can generalize well to unseen distortions and datasets.
- The proposed metrics are particularly effective for steganography-encoded images, addressing limitations of traditional approaches.

7.5.2 Limitations and Future Work

Despite their strengths, the proposed metrics have limitations:

- The FRFM's reliance on reference images limits its applicability in real-world scenarios.
- The NRM requires extensive training data, which may not be available for all applications.
- Future work could explore hybrid approaches that integrate reference-based and reference-free methodologies.

8 Conclusion and Future Work

8.1 Conclusion

This thesis addressed the challenge of aligning objective Image Quality Assessment (IQA) metrics with human perception, particularly in the context of facial images and steganography-encoded distortions. The research was driven by the need for metrics that capture subtle, perceptually significant artifacts while maintaining computational feasibility.

The primary contributions of this work include:

- Development of the Full-Reference Fusion Metric (FRFM), which integrates multiple highly correlated metrics into a unified framework, achieving superior alignment with subjective evaluations.

- Design of the No-Reference Metric (NRM), leveraging neural networks to generalize quality assessment to scenarios without pristine reference images.

- Comprehensive validation of the proposed metrics against traditional methods and subjective Mean Opinion Scores (MOS), demonstrating significant improvements in correlation, robustness, and adaptability.

The proposed metrics effectively address the limitations of existing approaches, offering robust solutions for evaluating facial images and steganographic distortions. Their ability to bridge the gap between objective computations and human perception marks a significant advancement in the field of IQA.

8.2 Future Work

While this thesis has made substantial contributions, several avenues for future research remain:

8.2.1 Hybrid Metrics

Future work could explore hybrid approaches that combine the strengths of Full-Reference and No-Reference metrics. Such methods could leverage reference images when available while maintaining the flexibility to operate without references.

8.2.2 Generalization Across Datasets

The current metrics were validated on a specific dataset. Future studies should test their performance on larger, more diverse datasets to evaluate their generalizability across different image domains and distortion types.

8.2.3 Adapting to Emerging Distortion Types

As steganography and other encoding techniques evolve, new distortion types will emerge. Extending the proposed metrics to handle these distortions will be critical for maintaining their relevance.

8.2.4 Real-Time Applications

The computational complexity of learning-based metrics, such as the NRM, limits their use in real-time scenarios. Future research could focus on optimizing these models for real-time performance without sacrificing accuracy.

8.2.5 Integration with Human-Perception Models

Incorporating models from neuroscience and psychology into IQA frameworks could further enhance their alignment with human perception. For instance, integrating attention mechanisms to prioritize salient image regions could improve metric performance for facial images.

8.3 Final Remarks

This thesis represents a step toward more perceptually aligned IQA metrics, addressing critical gaps in the evaluation of facial images and steganographic distortions. By combining insights from subjective evaluations, objective computations, and neural network-based methods, the proposed metrics offer a robust foundation for future advancements in the field of IQA. Continued research in this area will not only refine these metrics but also expand their applicability to emerging challenges in digital imaging.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[40] J. Fridrich, "Steganography in digital media: Principles, algorithms, and applications," *Steganography in Digital Media*, Jan. 2010. DOI: 10.1017/CBO9781139192903.

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

[42] M. Tancik, B. Mildenhall, and R. Ng, *Stegastamp: Invisible hyperlinks in physical photographs*, Jun. 2020.

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

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

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

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

[47] X. Liu *et al.*, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *ICCV*, 2017.

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

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

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

[51] L. Chen, C. Pan, and Y. Fang, "Blind image quality assessment with pseudo-mos learning," *IEEE Signal Processing Letters*, vol. 28, pp. 1090–1094, 2021.

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

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

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

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

[56] L. Ma, S. Wang, G. Shi, D. Zhao, and W. Gao, "Psnr-b: A novel peak signal-to-noise ratio based on edge preservation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 537–549, 2012. DOI: 10.1109/JSTSP.2012.2204259.

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

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

[59] T. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1793–1807, 2013.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[83] N. Venkatanath, D. Praneeth, M. Panchagnula, S. R. M., L. K., and S. Chatterjee, "Blind image quality evaluation using perception based features," in *Proc. National Conf. Communications (NCC)*, 2015, pp. 1–6.

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

[85] W. Malpica and A. C. Bovik, "Range Image Quality Assessment by Structural Similarity," in *Encyclopedia of Multimedia*, B. Furht, Ed., Springer, 2008, pp. 755–757.

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

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

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

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

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

[91] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007.

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

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

[94] 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.

[95] 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.

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

[97] 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.

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

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

[100] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[101] Z. Wang and A. C. Bovik, "Perceptual quality metric using a joint model of signal fidelity and naturalness," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, IEEE, vol. 3, 2002, pp. III–III.

[102] N. Venkatanath, D. Praneeth, M. Bh, S. Channappayya, and S. Medasani, "Blind image quality evaluation using perception based features," in *National Conference on Communications (NCC)*, IEEE, 2015, pp. 1–6.

[103] J. Chen, K. Ma, S. Wang, W. Zeng, and Z. Wang, "Blind image quality assessment via score distribution prediction," *IEEE Transactions on Image Processing*, vol. 31, pp. 2496–2511, 2022.

[104] J. P. Robinson, R. Stevenson, N. Parde, A. Sarkar, K. W. Bowyer, and M. C. King, "Face recognition: Too bias, or not too bias?" *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 28–39, 2020.

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

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

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

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

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

[110] J. Hernández-Ortega, J. Galbally, J. Fierrez, and F. Alonso-Fernandez, "Faceqnet v2: Quality assessment for face recognition systems," in *Proceedings of the International Conference on Biometrics (ICB)*, IEEE, 2019, pp. 1–6.

Appendix A

Sample Appendix