

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

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.

Setembro 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

Co-Supervisor:

João de Sena Baptista Pimentel Marcos

Jury:

João Pedro de Almeida Barreto

João de Sena Baptista Pimentel Marcos

Vítor Manuel Mendes da Silva

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

Coimbra, September 2025

Acknowledgments

Em primeiro lugar, gostaria de agradecer ao Professor Nuno Gonçalves por sempre me ter acompanhado durante a licenciatura, mestrado e em projetos de investigação. Por ter aceitado o meu pedido de orientação e por ter persistido comigo durante estes últimos dois anos, sem nunca desistir de mim. Muito obrigado por toda a sua orientação, atenção, paciência, horas de reuniões e pela sua amizade. Levá-lo-ei no coração para a vida. Ao meu coorientador, Professor João Marcos, que também me acompanhou de perto no mestrado, trazendo uma boa dinâmica de trabalho e amizade. Pelos ensinamentos, dentro e fora da sala de aula, e pela partilha de experiências, sou-lhe muito grato, por mais que teime que o trate por "tu", acho que simplesmente não consigo. Por estes dois projetos de investigação, que tanto me ensinaram, o meu mais profundo obrigado.

A ti, Coimbra, à Universidade de Coimbra e ao DEEC, por me fornecerem as condições de estudo e de vida, pelas amizades, vivências e aprendizagens que, orgulhosamente, me moldaram na pessoa que hoje sou. Obrigado a todos os docentes que contribuíram para a minha formação e crescimento pessoal, em particular aos Professores Diogo Júdice, Paulo Coimbra e Lúcia Martins. Não podia deixar de agradecer à Dona Graça e ao Senhor Carlos, dois rostos que fazem o DEEC ser o que é. Obrigado por tudo.

I would like to thank Professor Shahrukh Athar of McMaster University, Ontario, whose previous work provided the foundation for this dissertation and who generously shared his insights with me. This work would not have been possible without his contributions.

Professor César, Professora Adélia, consegui. Os tempos de escola já lá vão, mas não me esqueci de vocês. Tinha de reservar um cantinho para os meus professores do ensino secundário, a si, Professora Adélia, por me ter chamado à razão quando era preciso, e a si, Professor César, por ter sido um porto seguro mesmo após tantos anos sem nos vermos. Obrigado por tudo.

A todos os meus colegas do ISR, pelos ensinamentos e horas de debugging. Obrigado Farhad, Luiz e Aniana por terem codificado as imagens usadas nesta dissertação. E ao VisTeam, pelo apoio triplicado na avaliação das imagens. Se a métrica tem a qualidade que tem, certamente é graças a vocês. Um obrigado especial ao Allan, pela paciência, e ao Diogo, pelo passinho de

dança no NL.

Aos meus colegas da TWEVO. Miguel, Eurico, João, Carreira, Carias e Xavier, os vossos ensinamentos ajudaram-me a crescer tanto em tão pouco tempo. Obrigado por me terem recebido de braços abertos e por me mostrarem o vosso esforço e dedicação por detrás desta empresa. Ao Carlos, à Mónica e ao Márcio, por me terem dado a oportunidade de trabalhar com esta equipa. Márcio, continuo a achar que és tu que me estás a dever a paçoca...

NEEEC, meu querido NEEEC, já disse, quando perdi uma casa, ganhei outra. Quando entrei no curso, não estava à espera de encontrar uma segunda família, mas encontrei-a em vocês. Desde os convívios aos jogos infindáveis de pingue-pongue, desde as milhentas horas de "estudo" às canecas no Reis. Obrigado por me terem ensinado que o DEEC é muito mais que um departamento, é um lar, onde devemos reivindicar pela nossa casa e ajudar o colega do lado.

Não podia deixar de falar dos nossos queridos vizinhos. O Somos Pólo II, mostrou-me que a amizade não se delimita por departamentos. A vossa força desafiante mostrou-me que juntos somos mais fortes. Um agradecimento especial ao Pedro Campos, Diogo Martinho, Roman, Henrique, Gonçalo Lourenço, Diogo Vieira e Raimundo. Obrigado por me terem integrado tão bem na equipa. Fizemos, indubitavelmente, o melhor MEGA que o Polo II já viu.

Flórido, Casca, Zé, eina... acho que não preciso de dizer muito mais. Se o mandato foi curto, a amizade que construímos é para a vida. Sei que houve reuniões difíceis, "Onde é que estão os chocolates?", "Flórido, eu meti a mão na urna!", "Casca, e essa gala, vai dar lucro?", "Desculpa, Zé, esqueci-me de fazer a ata". Mas nem todas as reuniões foram más, as do Reis corriam bem (pelo que me lembro). Obrigado por me terem convidado para este projeto. Sou-vos muito grato.

Por falar em Reis, um agradecimento especial a todos os que tornaram estes últimos anos em Coimbra, os mais memoráveis. Aos meus antecedentes eletrotécnicos, Paredes, Senhor Zé, pelas histórias do "no meu tempo" e conselhos de vida. Tiago, Tina, Joana, e Aldinhas, eu disse não me ia esquecer de vocês, fora as lições de moral, obrigado pelo convívio e por me fazerem sentir em casa. Carlos Grande, gostava de chegar à tua idade e de ter metade da tua boa disposição. Obrigado por me teres feito sorrir tantas vezes. A si Senhor António, pela partilha de conhecimento. À Tininha, que apesar de ainda não ter ganho nada nas raspadinhas, foi a minha sorte grande. Um especial abraço ao Tiago, Bryan, João Nuno, e Diogo, sempre foram um exemplo exímio de uma amizade perfeita. À Flávia e ao Gonçalo, que quando faltavam, tinham justificação, pois estavam a escrever o regulamento do jogo da moeda, obrigado por todos os momentos que me proporcionaram. Aos meus amigos do coração, Edwin e Brito, juro que não vos queria apartar, obrigado por me aturarem tanto tempo.

Déh e Duarte, juntos somos os três vices que sabem o que é a verdadeira luta. Obrigado pelo carinho, horas de conversa, boleias e apoio emocional. Se me disserem que gostam tanto de mim como eu gosto de vocês, estariam a mentir! Amo-vos do fundo do coração.

Família de praxe, até à Marysol, exclusive. Paulino, lá sabia eu que, mesmo a 2000 km de distância, me ias fazer chorar! Obrigado pelas turras na cabeça (exceto no Bot). Gosto muito de ti e espero que a tua vida te sorria tanto como tu me fazes sorrir. Sei que vais ser um grande engenheiro, tal como o Padrinho, obrigado por me teres escolhido (vamos ignorar o Casca). Marysol, estava a brincar! Tu desafias a paciência de qualquer um. Espero ter tido um impacto positivo na tua vida, as minhas pancadinhas sempre foram com carinho. André Cena, obrigado por teres depositado em mim a responsabilidade de ser teu padrinho. Espero que eu tenha estado à altura do desafio, continua a ser o grande homem que és.

A todos os que fizeram parte do (melhor) Carro da Queima, um especial obrigado ao meu amigo Carlos Faria, João Domingos, Rui Guimarães, Gonçalo Santos, João Magano, Gonçalo Dias e Miguel Branco. Mesmo longe, trago-vos no coração. Obrigado por me terem proporcionado um dos melhores dias da minha vida. Espero que tenha sido assim para vocês também.

Aos Amontoados, a ti Andreia, Tiago, Eduardo, João, Bernardino e Catarina, por terem feito parte da minha integração no curso e da minha vida. Eduardo, ainda estou à espera que faças a laboratorial de SM! Obrigado por teres estado lá nos melhores e piores momentos. Tiago, és uma força da natureza, o Iron Man está ao virar da esquina, não tenho a menor dúvida de que te vais tornar um. Para mim serás sempre o meu irmãozinho. Quando cortarem a fita, berra e grita, na tua vida só tu ditas o guião.

Miguel Ávila, Daniel Agostinho, tinha de haver um parágrafo vosso. A palavra amizade não chega para descrever o que existe entre nós. Desde os assaltos às bancas de shots no 24 às noites de boémia ao som do GrandSons, do Reis até Vigo, de Anadia até Viana. A tripla não dorme, nem deixa que Coimbra adormeça. Obrigado por terem estado do meu lado este último ano. Sei que isto é só o começo. Um brinde a vocês, sei que brevemente estaremos juntos a celebrar da mesma maneira de sempre. Jaci, Swarovski, não me esqueci de vocês, mas o parágrafo já estava grande demais. Um beijinho grande às duas. Obrigado por me aturarem e desculpem lá qualquer coisa!

Augusto, Cunha, Vasco, Pedro, para quando um pedi-tascas? Pedro, ainda te continuas a atrasar para a escola? Cunha, os 20 aqui são mais difíceis. Vasco, quando é que voltas a Portugal? Augusto, já entraste em Direito? Tinha de desabafar... Obrigado por terem feito parte da minha vida. Estarei sempre aqui para vocês. Danny, my man. I know you are far away, it's about time I came to visit you. It's only fair. Undoubtedly one of the most unlikely

friendships one could ever make, you are the sole reason I believe in destiny. Love you, bro.

Inês, obrigado por me teres aturado, por me teres ajudado a crescer, por me teres ensinado tanto e por me teres feito tão feliz, tanto em Coimbra, como em Milão, e agora em Leiria. Onde quer que esteja, se estiver contigo, estou em casa. Obrigado por estes últimos anos, foste a melhor coisa que me aconteceu. Obrigado por tudo o que és e por tudo o que me fazes ser. Amo-te.

A ti, Madrinha, pela mão pesada, e a vocês Rafael e Luís, por me terem acompanhado de perto durante esta jornada. Não foi fácil, vocês já o devem saber, mas está feito, espero ter feito jus às vossas expectativas. Prima João, anseio pelo dia em que te deixe espantar com uma criação minha, obrigado pela frontalidade e feedback.

Mãe, Pai, Nuna. Mãe, obrigado por tudo. Sei que ainda não me tornei no grande homem que tanto desejavas, mas estou a trabalhar para isso. Quem tem uma mãe tem tudo, mas quem tem uma mãe como tu não pode pedir por mais nada na vida. Amo-te incondicionalmente. Pai, sei que nem sempre foi fácil. Sei que nestes últimos seis anos passei muito pouco tempo contigo. Obrigado pelo carinho e por me teres transmitido o bichinho da política. Do teu filhote réré, amo-te. Nuna, minha maninha, este é um momento especial e não te podia deixar de fora. Temos muito mais em comum do que tu pensas. O mano ama-te, mesmo que sejas parvinha às vezes.

Deonilde, tinha de guardar o melhor para o fim. Espero um dia vir a ser tudo aquilo que dizes que sou... mas, por agora, espero que estejas orgulhosa do teu Neto. Ainda me lembro de te ligar e contar quantas frequências e trabalhos tinha para a próxima semana, e tu, na tua inocência, respondias: “O teu trabalho é estudar... não te pedem mais nada, pois não?”. Nunca deixes de rezar por mim.

Obrigado a todos vós, que me ajudaram a chegar até aqui. Obrigado a todos os que contribuiram para o estudo desta nova métrica. Esta dissertação é tão vossa como minha.

Abstract

Image quality is a fundamental requirement for reliable biometric recognition and secure identity systems. Yet, existing evaluation methods often fail against distortions that are imperceptible to the human eye but critically affect recognition, such as those introduced by steganography. They also reproduce demographic biases, leading to uneven treatment across age, gender, and skin tone. This dissertation introduces a pipeline for creating new image quality metrics through a progressive path from full-reference fusion to no-reference prediction. The first stage combines complementary full-reference measures into a single fused metric that reflects unbiased human perception more faithfully than any individual method. This produces metrics such as pMOS-SSIM+, which show superior alignment with subjective opinion scores across benchmark datasets. The second stage uses the fused metric as a proxy ground truth to supervise no-reference predictors. This step yields task-specific models, such as pMOS-StegaFace, designed for steganographically distorted facial images, or pMOS-NoisyTID, which generalizes to diverse content and distortion types. Results confirm that the fused metrics consistently outperform established baselines, and that the no-reference models trained from them detect subtle degradations while maintaining strong correlation with human judgments. The pipeline also integrates fairness-aware debiasing of subjective scores, ensuring consistent performance across demographic groups. By uniting perceptual validity, biometric utility, and operational security, this work replicates subjective opinion scores from a small annotated subset, allowing for the bridging of full-reference performance with no-reference applicability.

Resumo

A qualidade de imagem é um requisito fundamental para um reconhecimento biométrico fiável e para sistemas de identidade seguros. No entanto, os métodos de avaliação existentes tendem a falhar perante distorções impercetíveis ao olho humano, mas com um grande impacto no reconhecimento, como distorções introduzidas pela esteganografia. Além disso, correm o risco de propagar enviesamentos demográficos, podendo levar a tratamento desigual entre faixas etárias, géneros e tons de pele. Esta dissertação apresenta uma pipeline para a criação novas métricas de qualidade de imagem através de um processo que começa na fusão de métricas de referência completa (faz uso da imagem original como referência) até à predição sem referência (não requer uma imagem de referência). Numa primeira etapa, combinámos as avaliações de métricas de referência completa numa única métrica fundida, que melhor reflete a perceção humana do que qualquer outra métrica individualmente. Este processo produziu métricas como a pMOS-SSIM+, que mostra um alinhamento superior com as avaliações subjetivas em bases de dados com diferentes tipos de imagens. Numa segunda etapa, utilizámos a métrica fundida para propagar estas avaliações de modo a treinar preditores sem referência. Este passo gera modelos específicos dependendo da tarefa, como a pMOS-StegaFace, especializada para imagens faciais distorcidas por esteganografia, ou a pMOS-NoisyTID, que generaliza a diferentes tipos de imagens e distorções. Os resultados confirmam que as métricas fundidas superam consistentemente o estado da arte e que os modelos sem referência treinados a partir delas detetam degradações subtis mantendo uma forte correlação com a opinião humana. A pipeline integra ainda um desenviesamento às opiniões subjetivas, garantindo um desempenho consistente entre grupos demográficos. Ao unir validade percetual, utilidade biométrica e segurança operacional, este trabalho reproduz anotações subjetivas a partir de um pequeno subconjunto anotado, permitindo estabelecer a ponte entre o desempenho de métricas de referência completa e a aplicabilidade das métricas sem referência.

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

Agostinho da Silva

Contents

Acknowledgements iii

Abstract vii

Resumo ix

List of Acronyms xvii

List of Figures xxiii

List of Tables xxv

1 Introduction 1

1.1 Contextualization . . . 1

1.2 Problem Statement . . . 3

2 Related Work 5

2.1 Subjective Image Quality Assessment . . . 5

2.1.1 Assessment Protocols . . . 5

2.1.2 Crowdsourcing and Remote Studies . . . 6

2.1.3 Benchmark Datasets . . . 6

2.2 Objective Image Quality Assessment . . . 8

2.2.1 Computational Domains . . . 9

2.2.2 Spectral Representations . . . 9

2.3 Full-Reference IQA Metrics . . . 10

2.3.1 Error-based . . . 12

2.3.2 Spectral-based . . . 13

2.3.3 Structural Similarity . . . 13

2.3.4 Information-theoretic . . . 13

2.3.5 Deep Learning Methods . . . 14
2.4 No-Reference IQA . . . 14
2.4.1 Classical Approaches . . . 15
2.4.2 Facial Image Quality Assessment . . . 17
2.5 Fusion-based IQA Models . . . 18
2.6 Image distortion types . . . 19
2.6.1 Lossy compression . . . 19
2.6.2 Blur and Detail Loss . . . 20
2.6.3 Additive Noise . . . 20
2.6.4 Resampling and Resolution Variations . . . 20
2.6.5 Photometric and Color Distortions . . . 20
2.6.6 Transmission and Packet-Level Artifacts . . . 20
2.6.7 Authentic Mixtures . . . 21
2.7 Steganography . . . 21
2.7.1 StegaStamp . . . 22
2.7.2 CodeFace . . . 22
2.7.3 RiemStega . . . 22
2.7.4 StampOne . . . 22

3 Methodology **25**
3.1 Dataset . . . 25
3.2 Collection of Subjective Scores . . . 27
3.3 Debiasing Subjective Scores . . . 30
3.4 Evaluation of Full-Reference Metrics . . . 31
3.5 Full-Reference Fusion Metric . . . 31
3.6 pseudo-Mean Opinion Score Generation . . . 32
3.7 NR Metric from FR Fusion . . . 32

4 Results **35**
4.1 Face Research Lab London Dataset Pipeline . . . 35
4.1.1 Subjective Test Outcomes . . . 35
4.1.2 Correlation of FR Metrics with MOS . . . 36
4.1.3 Full-Reference Fusion . . . 38
4.1.4 No-Reference Model . . . 40
4.2 Tampere Image Database 2013 Pipeline . . . 41

4.2.1 Subjective Test Outcomes . . . 43
4.2.2 Correlation of FR Metrics with MOS . . . 43
4.2.3 Full-Reference Fusion . . . 45
4.2.4 No-Reference Model . . . 49

5 Conclusion and Future Work **51**

A Subjective Screening Procedures **67**
A.1 Confidence Intervals . . . 67
A.2 Observer Screening . . . 67
A.2.1 Outlier Detection . . . 68
A.2.2 Rejection Rule . . . 68
A.3 Summary . . . 68

B Full Reference Metrics **69**
B.1 Error-based Metrics . . . 69
B.2 Spectral-based Metrics . . . 70
B.3 Similarity-based Metrics . . . 71
B.4 Information-theoretic Metrics . . . 77
B.5 Learning-based Metrics . . . 77

C Regression Models for Metric Fusion **81**
C.1 Linear Models . . . 81
C.2 Ensemble Methods . . . 82
C.3 Boosted Trees . . . 83

D Hardware and Software Environment **85**
D.1 Hardware . . . 85
D.2 Software . . . 85

List of Acronyms

AGGD	Asymmetric Generalized Gaussian Distribution
ANOVA	Analysis of Variance
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
BAPPS	Berkeley-Adobe Perceptual Patch Similarity
BEF	Blocking Effect Factor
BRISQUE	Blind/Referenceless Image Spatial Quality Evaluator
C-SSIM	Color Structural Similarity
CatBoost	Categorical Boosting
CNN	Convolutional Neural Network
CSIQ	Categorical Subjective Image Quality
CV	Cross-Validation
CW-SSIM	Complex Wavelet Structural Similarity
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DISTS	Deep Image Structure and Texture Similarity
D_λ	Spectral Distortion Index
DMOS	Difference Mean Opinion Score
DSSIM	Structural Dissimilarity Index
DS	Double Stimulus

DWT	Discrete Wavelet Transform
ERGAS	Relative Global Dimensional Error of Synthesis
ESIM	Edge Strength Similarity
FaceQnet	Facial Quality Network
FIQA	Facial Image Quality Assessment
FR-IQA	Full-Reference Image Quality Assessment
FRGC	Face Recognition Grand Challenge
FRLL	Face Research Lab London
FR	Full-Reference
FSIMc	Feature Similarity Index with color
FSIM	Feature Similarity Index
G-SSIM	Gradient-based Structural Similarity
GANs	Generative Adversarial Networks
GGD	Generalized Gaussian Distribution
GMSD	Gradient Magnitude Similarity Deviation
GMS	Gradient Magnitude Similarity
GM	Gradient Magnitude
GSM	Gradient Scale Mixtures
HVS	Human Visual System
HYDICE	Hyperspectral Digital Imagery Collection Experiment
IAA	Image Aesthetic Assessment
ICAO	International Civil Aviation Organization
IEC	International Electrotechnical Commission
IFC	Information Fidelity Criterion

IJB-C	IARPA Janus Benchmark C
IQA	Image Quality Assessment
ISO	International Organization for Standardization
ITU-R BT.500-15	International Telecommunication Union - Radiocommunication Sector Recommendation Broadcasting service (television) 500-15
IVC	Image and Video Communications
IW-SSIM	Information-Weighted Structural Similarity
JPEG	Joint Photographic Experts Group
KADID-10k	Konstanz Artificially Distorted IQA Database
KonIQ-10k	Konstanz Image Quality
L-SSIM	Luminance Structural Similarity
LFW	Labeled Faces in the Wild
LightGBM	Light Gradient Boosting Machine
LIVE	Laboratory for Image & Video Engineering
LPF	Low-Pass Filter
LPIPS	Learned Perceptual Image Patch Similarity
MAE	Mean Absolute Error
MagFace	Magnitude-Aware Face
mANOVA	Multivariate Analysis of Variance
MMF	Multi-Method Fusion
MOS	Mean Opinion Score
MRTDs	Machine-Readable Travel Documents
MS-Celeb-1M	Microsoft Celeb One Million
MS-FSIM	Multiscale Feature Similarity Index
MS-SSIM	Multiscale Structural Similarity Index

MSCN	Mean Subtracted Contrast Normalization
MSE	Mean Squared Error
NIQE	Natural Image Quality Evaluator
NR-FIQA	No-Reference Face Image Quality Assessment
NR-IQA	No-Reference Image Quality Assessment
NR	No-Reference
NSS	Natural Scene Statistics
OFIQ	Open Source Face Image Quality
OLS	Ordinary Least Squares
ORM	Object-Relational Mapping
PC	Phase Congruency
PIPAL	Perceptual Image Processing Algorithms
PIQE	Perception-based Image Quality Evaluator
PLCC	Pearson Linear Correlation Coefficient
pMOS-NoisyTID	pseudo-Mean Opinion Score for Noisy TID2013 Images
pMOS-SSIM+	pseudo-Mean Opinion Score based on enhanced SSIM family metrics
pMOS-StegaFace	pseudo-Mean Opinion Score for Steganographic affected Face Images
pMOS	pseudo-Mean Opinion Score
PSM	Perceptual Similarity Metric
PSNR-B	Peak Signal-to-Noise Ratio with blocking effect
PSNR	Peak Signal-to-Noise Ratio
R-SSIM	Regional Structural Similarity
RBF	Radial Basis Function
RMSE	Root Mean Squared Error

ROI	Region of Interest
ROSIS	Reflective Optics System Imaging Spectrometer
RR	Reduced-Reference
S-CIELAB	Spatial CIELAB
SAM	Spectral Angle Mapper
SCC	Spatial Correlation Coefficient
SER-FIQ	Stochastic Embedding Robustness for Face Image Quality
SHAP	SHapley Additive exPlanation
SIQAD	Screen IQA Database
SLTD	Stolen and Lost Travel Documents
SNR	Signal-to-Noise Ratio
SPAQ	Smartphone Photography Attribute and Quality
SPD	Symmetric Positive Definite
SPIQ	Semantic Perceptual Image Quality
SR-SIM	Spectral Residual Similarity
SRCC	Spearman Rank Order Correlation Coefficient
SSIM	Structural Similarity Index
SS	Single Stimulus
STN	Spatial Transformer Network
SVD	Singular Value Decomposition
SVR	Support Vector Regression
TID	Tampere Image Database
TPE	Tree-structured Parzen Estimator
UQI	Universal Image Quality Index

VGG	Visual Geometry Group
VIFc	Visual Information Fidelity (color images)
VIFp	Visual Information Fidelity (pixel domain)
VIF	Visual Information Fidelity
VSNR	Visual Signal-to-Noise Ratio
W-SSIM	Wavelet Structural Similarity
WaDIQaM	Weighted Average Deep Image Quality Assessment Model
XGBoost	Extreme Gradient Boosting

List of Figures

3.1 Reference images from the MOS set, comprising the selected subjects from the FRLL dataset. 26
3.2 Steganographically distorted facial images from each method. 26
3.3 Django-based web application created for the Single Stimulus test sessions. . . . 27
3.4 Conceptual and physical database schemas implemented in the Django web application. 29

4.1 Scatter plots of MOS versus 40 individual FR-IQA metrics. The red curve is a local smoothing of the trend. 37
4.2 SVD-based analysis of the FR metric space. The variance curve selects k at the chosen threshold, and the Picard plot corroborates stability near that truncation. 38
4.3 SHAP summary plot for $k = 6$ on the best CatBoost fusion model. 39
4.4 pMOS-StegaFace predictions on train and test sets. 40
4.5 Reference images from the TID dataset. 42
4.6 Scatter plots of MOS versus 40 individual FR-IQA metrics. The red curve is a local smoothing of the trend. 46
4.7 SVD-based analysis of the FR metric space. The variance curve selects k at the chosen threshold, and the Picard plot corroborates stability near that truncation. 47
4.8 SHAP summary plot for $k = 12$ on the best CatBoost fusion model. 47
4.9 Scatter plot of pMOS-SSIM+ predictions versus human MOS on the pMOS set. The red line is the linear fit. 48
4.10 pMOS-NoisyTID predictions on train and test sets. 49
4.11 TID2013 reference images from the test set. 50

List of Tables

2.1 Summary of the 40 FR-IQA metrics used 10
2.2 Summary of the 6 NR-IQA metrics used 15

3.1 Partitioning of the steganographically distorted FRLL dataset. 25
3.2 Demographic and non-demographic profile of the observers. 28
3.3 ANOVA and mANOVA results for observer and image attributes. Before debiasing, several factors show statistically significant effects on MOS, p-value < 0.05. After residualization, all main effects show no significant impact, confirming the effectiveness of the debiasing procedure. 30
3.4 Training configuration of pretrained backbones for NR metric learning. 33

4.1 Per-method descriptive statistics on the MOS subset. 36
4.2 Individual FR-IQA metrics versus MOS on the unbiased train set. 38
4.3 Fusion regressors at $k = 6$ on the unbiased split. 39
4.4 Backbone comparison for training pMOS-StegaFace. Report 5-fold mean PLCC, SRCC, MSE, and MAE. 40
4.5 pMOS-StegaFace versus NR-IQA baselines on the test split. 41
4.6 Partitioning of the TID dataset. 43
4.7 Per-distortion descriptive statistics on the TID2013 dataset. 44
4.8 TID2013 FR fusion versus individual FR-IQA metrics on the MOS train set. . . 48
4.9 Backbone comparison for training pMOS-NoisyTID. Report 5-fold mean PLCC, SRCC, MSE, and MAE. 49
4.10 pMOS-NoisyTID versus NR-IQA baselines. 50

1 Introduction

1.1 Contextualization

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

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

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

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

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

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

The International Civil Aviation Organization (ICAO) [19] and the International Organization for Standardization (ISO) / International Electrotechnical Commission (IEC) 29794–5 standard [20] establish guidelines for image quality in Machine-Readable Travel Documents (MRTDs). In this context, the Open Source Face Image Quality (OFIQ) is the reference implementation for the ISO / IEC 29794–5 international standard. These guidelines ensure uniform image conditions (e.g., lighting, focus, and resolution) and consistency across datasets [21]. While these regulations establish a technical baseline, they do not account for perceptual biases and demographic variability in FIQA.

FIQA systems have been shown to produce different quality scores depending on a person's age, gender, or skin tone [22], [23]. 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 publicly available datasets, which leads to lower quality scores for those individuals [24]. This can cause face recognition systems to perform worse for some groups than others, raising serious concerns about fairness.

The ethical implications of these biases are profound. Political regulations, such as the European Convention on Human Rights (Article 14) [25], the Universal Declaration of Human Rights (Article 7) [26], the General Data Protection Regulation (Article 22) [27], and emerging AI governance frameworks, such as the European Artificial Intelligence Act (2024) [28] and proposals in the Artificial Intelligence Civil Rights Act (2024) of the U.S. Congress [29], 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 [24], [30]. 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 [31].

Although visually unobtrusive, these alterations can degrade biometric features and compromise recognition performance. No-Reference Image Quality Assessment (NR-IQA) approaches, while not requiring references, are generally not designed to detect such imperceptible, task-relevant distortions.

This issue is especially relevant for MRTDs, where enhanced security is critical. Incorporating steganographic imagery, such as hidden tamper-evident or authentication data, helps protect against increasingly sophisticated threats to identity documents. The necessity for such measures is underscored by growing criminal activity involving IDs. Reports from the European Border and Coast Guard Agency consistently highlight document fraud as one of the most serious threats at EU external borders [32]. Similarly, Europol identifies fraudulent IDs as a key enabler of organized crime and migrant smuggling [33]. INTERPOL's Stolen and Lost Travel Documents (SLTD) database currently contains more than 138 million records of lost or stolen travel documents, illustrating the global scale of the problem [34]. In the United States alone, identity fraud losses reached an estimated $43 billion in 2023 [35].

Steganographically embedding cryptographic signals in passport portraits can therefore serve as a complementary line of defense alongside chip-based authentication. This approach directly addresses known threats such as morphing attacks [36], photo substitution [19], and stolen-document misuse [34], while aligning with ICAO's emphasis on secure and standardized travel document images [19].

1.2 Problem Statement

Despite substantial advances in FIQA, existing methods are not designed to capture subtle but security-critical degradations, particularly those introduced by steganographic embedding. Steganography alters pixel statistics in ways that are imperceptible to human observers yet capable of undermining biometric feature extraction. Conventional FIQA approaches, whether based on handcrafted features [37], supervised learning from quality annotations [38], [39], [40], or ranking-based formulations [41], are typically trained on distortion categories represented in standard datasets (e.g., blur, noise, compression). Consequently, they fail to generalize to covert perturbations that do not visibly degrade an image but can critically impair recognition accuracy.

A second limitation lies in fairness. Empirical studies show that FIQA models often inherit demographic biases from their training data, leading to systematically lower quality scores for specific subgroups [22]. In high-stakes applications such as MRTDs, governed by ICAO and ISO / IEC standards [19], [20], these disparities raise profound ethical and legal concerns. At the

same time, evidence from border security agencies highlights that document fraud, morphing attacks, and illicit ID use remain prevalent, underscoring the necessity of robust and equitable quality control mechanisms.

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

2 Related Work

2.1 Subjective Image Quality Assessment

Subjective image quality assessment is grounded in human visual perception and remains the gold standard for evaluating visual fidelity. Standardized by ITU-R BT.500-15 [13], these methodologies rely on controlled laboratory conditions, calibrated displays, and predefined rating procedures to ensure reproducibility and validity.

2.1.1 Assessment Protocols

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

Rating scales can be numerical categorical or non-categorical. Numerical categorical scales range from 1 to 5, 1 to 11, or 0 to 100, and are used to express either quality (excellent to bad) or impairment (imperceptible to very annoying). Non-categorical are usually a variant of a categorical method, for example, the assessor assigns each image to a point on a line drawn between two semantic labels. Both forms must result in a distribution of numbers for each condition. The method of analysis used depends upon the type of judgment and the information required (e.g. ranks, central tendency, psychological "distances").

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

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

where N is the number of observers. In double-stimulus protocols, the DMOS is defined as the difference between the MOS of the reference and that of the distorted image:

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

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

2.1.2 Crowdsourcing and Remote Studies

While laboratory-based studies remain the benchmark for reliability, they are costly, time-consuming, and limited in participant diversity. To overcome these constraints, recent works have turned to crowdsourcing platforms such as Amazon Mechanical Turk, Figure Eight, and Prolific [42], [43]. These frameworks enable the collection of tens of thousands of ratings across heterogeneous populations at a fraction of the cost of laboratory tests.

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

However, uncontrolled environments introduce noise: participants use uncalibrated displays, ratings might be inconsistent, and some responses are malicious or inattentive. To mitigate these issues, robust post-screening strategies are employed, including consistency checks, gold-standard reference images, and statistical filtering [42], [44]. Despite these limitations, crowdsourced IQA datasets have proven critical for training modern deep learning-based models, as they provide both scale and diversity unavailable in controlled laboratory conditions.

2.1.3 Benchmark Datasets

Several benchmark datasets have been curated to support the development and evaluation of IQA models. Early datasets such as Laboratory for Image & Video Engineering (LIVE) [45], [46], [47], Categorical Subjective Image Quality (CSIQ) [48], and Tampere Image Database (TID) [49], [50] follow ITU-R BT.500-15-compliant protocols, providing high-quality MOS or

DMOS annotations across a wide variety of synthetic distortion types. These databases remain the most widely used for evaluating traditional full-reference IQA metrics. Tampere Image Database (TID)2008 [49] created the accepted baseline for reporting PLCC and SRCC values of FR metrics. TID2013 [50] extended its previous dataset with additional distortion types and levels, further challenging IQA models.

The Image and Video Communications (IVC) [51] dataset further contributed subjective quality scores for compression-related distortions, while Screen IQA Database (SIQAD) [52] targeted screen content images, extending IQA research to non-natural imagery. Konstanz Artificially Distorted IQA Database (KADID-10k) [43] scaled this approach to 10,000 distorted images, enabling large-scale training. The Waterloo Image Aesthetic Assessment (IAA) database [53] emphasized robustness testing through massive-scale synthetic distortions and helped reveal model overfitting issues to specific artifact distributions.

Konstanz Image Quality (KonIQ-10k) [44] introduced 10,073 images collected from the web with authentic distortions and over 1.2 million quality ratings, making it one of the largest IQA resources. Smartphone Photography Attribute and Quality (SPAQ) [54] specifically addressed smartphone photography, collecting 11,125 images and crowdsourced perceptual scores to capture the variability of mobile imaging pipelines. Perceptual Image Processing Algorithms (PIPAL) [55] targeted perceptual quality in image restoration, including outputs from generative models. Berkeley-Adobe Perceptual Patch Similarity (BAPPS) [56] provided a large-scale pairwise comparison dataset for learning perceptual similarity directly from human judgments.

Beyond general-purpose IQA, domain-specific datasets remain critical. Remote sensing benchmarks such as Reflective Optics System Imaging Spectrometer (ROSIS) [57], Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [58], and Hyperspectral Digital Imagery Collection Experiment (HYDICE) [59] provide hyperspectral imagery where distortions affect both perceptual and functional quality, but they lack standardized MOS or DMOS annotations. These datasets illustrate the difficulty of extending IQA methodologies into application-driven contexts.

Microsoft Celeb One Million (MS-Celeb-1M) [60] and Labeled Faces in the Wild (LFW) [61] are large-scale face datasets primarily designed for recognition tasks, but they have been repurposed for FIQA research, they contain a wide range of web-scrapped, real-world faces. Face Recognition Grand Challenge (FRGC) [62] combines controlled and uncontrolled scenarios for recognition evaluation. The IARPA Janus Benchmark C (IJB-C) [63] dataset further expanded this scope with images and videos captured in unconstrained environments, providing a chal-

lenging benchmark for both recognition and quality assessment.

The Face Research Lab London (FRLL) [64] represents a distinct contribution to IQA benchmarking, focusing on high-quality facial ICAO compliant imagery. It contains 102 unique subjects photographed under controlled studio conditions, with consistent illumination, wearing plain white t-shirts, and neutral, frontal poses. In addition to meeting passport-style requirements, the dataset also includes self-reported age, gender, and ethnicity. Attractiveness ratings (on a 1-7 scale from "much less attractive than average" to "much more attractive than average") for the neutral front faces, provided by 2513 raters aged 17-90, are also included.

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

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

2.2 Objective Image Quality Assessment

Objective IQA refers to the automatic estimation of visual quality using computational models. Unlike subjective approaches that rely on human ratings, objective IQA enables consistent and scalable evaluations, making it essential for applications such as compression, transmission, restoration, and enhancement [65], [66], [67]. Over the past decades, numerous methods have been proposed, shaped by different assumptions about the Human Visual System (HVS) and the statistical properties of natural images. Surveys and comparative studies [1], [68], [69], [70] summarize this evolution, showing a progression from simple error-based fidelity measures to advanced perceptual and learning-based frameworks. Although no metric fully replicates human judgment, objective IQA provides a reproducible and low-cost alternative that supports large-scale benchmarking and real-time quality monitoring.

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

2.2.1 Computational Domains

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

- Spatial domain: Early metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Peak Signal-to-Noise Ratio (PSNR) operate directly on pixel intensities by measuring absolute differences or signal-to-noise ratios [65]. Despite their simplicity, they fail to account for perceptual masking and structural consistency.

- Frequency domain: Transform-based methods use the Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), or wavelets to capture energy distributions across spatial frequencies. Examples include Information Fidelity Criterion (IFC) [71] and Visual Signal-to-Noise Ratio (VSNR) [72], which exploit properties of the HVS such as contrast sensitivity.

- Gradient and saliency domains: These methods emphasize structural edges and perceptually important regions. The Gradient Magnitude Similarity Deviation (GMSD) [73] measures gradient consistency, while saliency-based models [74] weight distortions according to visual attention maps.

- Information-theoretic domain: Some metrics interpret quality as the amount of shared information between reference and distorted images. Visual Information Fidelity (VIF) [75] and related approaches build on mutual information, quantifying how much perceptual information can be extracted reliably from a degraded signal.

- Deep feature domain: More recent approaches leverage convolutional neural networks or vision transformers to learn perceptual embeddings. Metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [76] and Deep Image Structure and Texture Similarity (DISTS) [70] correlate strongly with human judgments by comparing features extracted from pre-trained networks.

2.2.2 Spectral Representations

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

- Luminance-based: Many classical approaches, such as Structural Similarity Index (SSIM) [68], focus solely on the luminance channel (Y). This reflects the dominant role of brightness perception in the HVS and simplifies computation.

- Color-based: Extensions to RGB or alternative color spaces, including YCbCr and CIELab, enable the incorporation of chromatic information. Examples include Feature Similarity Index (FSIM) and its color extension Feature Similarity Index with color (FSIMc) [77], or Universal Image Quality Index (UQI) [78], which attempt to capture hue and saturation degradations alongside luminance.

- Multispectral and hyperspectral: In domains such as remote sensing, quality measures generalize to dozens or even hundreds of bands. Metrics like Spectral Angle Mapper (SAM) [79], Relative Global Dimensional Error of Synthesis (ERGAS) [80], and Spectral Distortion Index (D_λ) [51] ensure spectral consistency across wavelengths and are critical for preserving material properties.

This dual perspective highlights the diversity of IQA methodologies and provides a framework for their categorization. Building on this, Table 2.1 presents the 40 FR metrics and Table 2.2 summarizes the 6 NR metrics considered in this work. Each metric is characterized by its reference type, year of introduction, category domain, and tested databases, showing how the field has advanced from simple pixel-wise error measures to perceptually and semantically informed quality predictors.

2.3 Full-Reference IQA Metrics

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

Metric	**Year**	**Category**	**Tested Databases**
MSE [68]	2004	Error-based	LIVE, CSIQ, TID, KADID-10k, IVC, IAA
RMSE [68]	2004	Error-based	LIVE, CSIQ, TID, KADID-10k, IAA
MAE [68]	2004	Error-based	LIVE, CSIQ, TID, KADID-10k, IAA, PIPAL
PSNR [68]	2004	Error-based	LIVE, CSIQ, TID, KADID-10k, IAA, PIPAL
PSNR-B [81]	2011	Error-based	LIVE
SNR [68]	2004	Error-based	LIVE, CSIQ, TID2008
SAM [79]	1992	Spectral-based	AVIRIS, HYDICE, ROSIS
ERGAS [80]	2000	Spectral-based	AVIRIS, ROSIS
D_λ [80]	2000	Spectral-based	AVIRIS, ROSIS

(continuation)

Metric	**Year**	**Category**	**Tested Databases**
SSIM [68]	2004	Structural Similarity	LIVE, CSIQ, TID, TID
MS-SSIM [82]	2003	Structural Similarity	LIVE, CSIQ, TID, KADID-10k
DSSIM [68]	2004	Structural Similarity	LIVE, CSIQ, TID
IW-SSIM [83]	2010	Structural Similarity	LIVE, CSIQ, TID2013
CW-SSIM [84]	2009	Wavelet Similarity	LIVE, CSIQ, TID2013
G-SSIM [85]	2006	Gradient Similarity	LIVE, CSIQ, TID2008
R-SSIM [86]	2015	Gradient Similarity	—
L-SSIM [68]	2004	Structural Similarity	—
C-SSIM [87]	2017	Color Similarity	TID2013
W-SSIM [88]	2009	Wavelet Similarity	LIVE
GMS [73]	2014	Gradient Similarity	LIVE, CSIQ, TID2008
GMSD [73]	2014	Gradient Similarity	LIVE, CSIQ, TID2008
VSNR [72]	2007	Feature Similarity	LIVE, CSIQ, TID
FSIM [77]	2011	Feature Similarity	LIVE, CSIQ, TID2013, KADID-10k
MS-FSIM [77]	2011	Feature Similarity	LIVE, CSIQ
FSIMc [77]	2011	Feature Similarity	LIVE, CSIQ, TID2013, IVC
SR-SIM [89]	2013	Feature Similarity	LIVE, CSIQ
ESIM [90]	2006	Edge Similarity	LIVE
UQI [78]	2002	Structural Similarity	TID2008
SCC [91]	1995	Structural Similarity	—
IFC [71]	2005	Information	LIVE, TID2008
VIF [75]	2006	Information	LIVE, CSIQ, TID2008

(continuation)

Metric	Year	Category	Tested Databases
VIFp [75]	2006	Information	LIVE, TID
VIFc [75]	2006	Information	—
LPIPS-Alex [76]	2018	Deep Learning	BAPPS, LIVE, CSIQ
LPIPS-Squeeze [76]	2018	Deep Learning	BAPPS, LIVE, CSIQ
LPIPS-VGG [76]	2018	Deep Learning	BAPPS, LIVE, CSIQ
DISTS [70]	2020	Deep Learning	LIVE, CSIQ, KADID-10k
WaDIQaM [92]	2017	Deep Learning	LIVE
PSM [93]	2022	Deep Learning	BAPPS
SPIQ [94]	2022	Deep Learning	KADID-10k, KonIQ-10k

Full-Reference Image Quality Assessment (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 [68], [75]. 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.

A concise overview of each metric family, together with full notation, formal definitions, and implementation details for all 40 FR-IQA metrics in Table 2.1, is provided in Appendix B.

2.3.1 Error-based

Error-based metrics quantify distortion by computing direct numerical differences between the reference and distorted images [68]. Common examples include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Peak Signal-to-Noise Ratio with blocking effect (PSNR-B), and Signal-to-Noise Ratio

(SNR), which are mathematically simple and computationally efficient. However, they lack perceptual modeling and therefore correlate poorly with human opinion scores. Despite this limitation, they are still treated as FR-IQA methods and remain widely used as baselines.

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

2.3.2 Spectral-based

Spectral-based metrics evaluate the consistency of spectral information in multi-band or hyperspectral images [79], [80]. Each pixel is treated as a spectral vector, and quality is assessed by comparing structural relationships in the spectral domain. Representative measures include the Spectral Angle Mapper (SAM), which computes the angular difference between spectral vectors, the Relative Global Dimensional Error of Synthesis (ERGAS), which measures average relative errors across spectral bands, and D_λ, which quantifies inter-band distortion. These methods are widely used in remote sensing and image fusion, where preserving spectral integrity is critical.

2.3.3 Structural Similarity

Structural similarity metrics assess perceptual image quality by measuring structural or feature-based resemblance between the reference and distorted images. Rather than relying on pixel-wise errors, they capture patterns of spatial organization, texture, luminance, or phase alignment. Prominent examples include the Structural Similarity Index (SSIM) family and its extensions such as MS-SSIM, IW-SSIM, and CW-SSIM, as well as gradient- and edge-based measures like G-SSIM, GMS, and ESIM, and feature-driven approaches such as FSIM and SR-SIM. These methods often align more closely with the HVS than purely error-based metrics [68], [73], and remain central in modern FR-IQA research.

2.3.4 Information-theoretic

Information-theoretic approaches to IQA are grounded in Shannon's theory of communication [95], [96]. They model images as stochastic signals transmitted through a distortion channel, and quality is assessed by measuring how much mutual information between the reference and distorted images is preserved in the presence of perceptual noise. Representative metrics include the Information Fidelity Criterion (IFC), which quantifies shared mutual information between

reference and distorted coefficients, Visual Information Fidelity (VIF), which normalizes preserved information by that available in the reference, and its simplified or extended variants, VIFp and VIFc. These methods provide a principled view of perceptual fidelity and often correlate well with subjective judgments, but they are typically more computationally demanding than error- or similarity-based metrics.

2.3.5 Deep Learning Methods

Deep learning approaches to IQA exploit feature representations from convolutional neural networks pretrained on large-scale vision tasks. These feature spaces capture high-level perceptual attributes such as texture, structure, and semantics, which align more closely with human visual perception than pixel-wise comparisons [76], [97], [98]. Representative metrics include the Learned Perceptual Image Patch Similarity (LPIPS) family, which measures distances in deep feature space, Deep Image Structure and Texture Similarity (DISTS), which combines structure and texture similarity, Weighted Average Deep Image Quality Assessment Model (WaDIQaM), a data-driven patch-based model, Perceptual Similarity Metric (PSM), a shift-tolerant variant of LPIPS, and Semantic Perceptual Image Quality (SPIQ), which leverages self-supervised pretraining for quality-aware representations. These methods have shown strong correlation with human opinion scores, particularly for complex and perceptually salient distortions, though they typically require higher computational resources.

2.4 No-Reference IQA

NR-IQA estimates perceptual quality directly from a single image without requiring a pristine reference. Three modeling strategies are commonly distinguished. First, classical opinion-unaware methods rely on natural scene statistics or perceptually motivated features and operate without training on opinion scores, providing simple and interpretable baselines [99], [100]. Second, task-driven deep approaches exploit the robustness of embeddings from pretrained networks to derive quality scores without supervision, thereby aligning quality with representational stability rather than explicit perceptual labels [40]. Third, supervised regressors adapt recognition backbones to annotated labels, explicitly learning to map features to continuous quality scores [38], [39].

A complementary distinction concerns domain scope. Generic NR-IQA methods aim to capture perceptual fidelity across natural images, while face-specific approaches (FIQA) target biometric utility, where quality is linked to recognition performance rather than aesthetics. The latter group has gained traction with the emergence of embedding-based models that quantify

image suitability for reliable face recognition. Considering both modeling strategies and domain scope enables a nuanced comparison between general-purpose perceptual estimators and specialized, task-aligned FIQA metrics.

Table 2.2: Summary of the 6 NR-IQA metrics used

Metric	**Year**	**Category**	**Tested Databases**
BRISQUE [101]	2012	Classical (opinion-aware)	LIVE, CSIQ, TID2013, KADID-10k
NIQE [99]	2012	Classical (opinion-unaware)	LIVE, CSIQ, TID2013, KonIQ-10k
PIQE [100]	2015	Classical (perception-based)	LIVE, CSIQ, TID2013, KADID-10k
FaceQnet [38]	2019	Face-specific (supervised)	VGGFace2, BioSecure, FRGC
SER-FIQ [40]	2020	Face-specific (unsupervised)	LFW, VGGFace2, FRGC
MagFace [39]	2021	Face-specific (supervised)	MS-Celeb-1M, IJB-C, LFW

2.4.1 Classical Approaches

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

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [101] is an opinion-aware method grounded in Natural Scene Statistics (NSS). It first applies Mean Subtracted Contrast Normalization (MSCN) to locally standardize luminance:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + \epsilon}. \tag{2.3}$$

where, $\mu(x,y)$ and $\sigma(x,y)$ denote the local mean and standard deviation of pixel intensities in a neighborhood around location (x,y), and ϵ is a small constant that prevents division by zero. This normalization enhances structural regularities and makes subsequent natural scene statistics more discriminative for quality assessment.

From the MSCN field, BRISQUE extracts parametric NSS features at two scales (original and downsampled by 2). First, it fits a Generalized Gaussian Distribution (GGD) to the MSCN

coefficients to obtain shape and scale parameters. Second, it fits an Asymmetric Generalized Gaussian Distribution (AGGD) to four pairwise product maps that capture local dependencies:

$$P_h(x,y) = \hat{I}(x,y)\,\hat{I}(x{+}1,y), \tag{2.4}$$

$$P_v(x,y) = \hat{I}(x,y)\,\hat{I}(x,y{+}1), \tag{2.5}$$

$$P_{d1}(x,y) = \hat{I}(x,y)\,\hat{I}(x{+}1,y{+}1), \tag{2.6}$$

$$P_{d2}(x,y) = \hat{I}(x,y)\,\hat{I}(x{+}1,y{-}1). \tag{2.7}$$

Each AGGD fit contributes shape, mean, and left/right scale parameters. Concatenating the GGD and AGGD parameters over both scales yields a low-dimensional feature vector:

$$\mathbf{f}(I) \in \mathbb{R}^{36}. \tag{2.8}$$

A Support Vector Regression (SVR) with Radial Basis Function (RBF) kernel, trained on human opinion scores from LIVE, maps these features to a scalar quality prediction:

$$\mathrm{BRISQUE}(I) = \mathrm{SVR}_{\mathrm{RBF}}(\mathbf{f}(I)). \tag{2.9}$$

Lower BRISQUE values indicate better perceived quality.

Natural Image Quality Evaluator (NIQE) [99] is also grounded in NSS. It uses a statistical deviation model to quantify perceptual degradation. First, it applies MSCN to locally standardize luminance:

$$\hat{I}(x,y) = \frac{I(x,y) - \mu(x,y)}{\sigma(x,y) + \epsilon}, \tag{2.10}$$

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

$$\mathrm{NIQE}(\mathbf{f}_t) = \sqrt{(\mathbf{f}_t - \mu_r)^T \Sigma_r^{-1} (\mathbf{f}_t - \mu_r)}. \tag{2.11}$$

Lower NIQE values indicate statistics closer to the pristine manifold, while higher values indicate stronger deviation.

Perception-based Image Quality Evaluator (PIQE) [100] begins by dividing the input image into non-overlapping 16×16 blocks, denoted B_i, and identifying distorted blocks based on edge content and local variance thresholds. For each block, it computes perceptual features using the gradient magnitude:

$$S_i = \frac{1}{|B_i|} \sum_{(x,y)\in B_i} \sqrt{\left(\frac{\partial I}{\partial x}\right)^2 + \left(\frac{\partial I}{\partial y}\right)^2}, \tag{2.12}$$

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

$$N_i = \text{std}\left(B_i - \text{LPF}(B_i)\right). \tag{2.13}$$

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

$$\text{PIQE} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} w_i \cdot (\alpha S_i + \beta N_i + \gamma C_i), \tag{2.14}$$

where w_i are confidence weights and α, β, γ are scaling factors for sharpness, noise, and contrast, respectively. The reported score is typically scaled to $[0, 100]$, with larger values reflecting worse perceived quality.

2.4.2 Facial Image Quality Assessment

FIQA refers to the estimation of face image quality with respect to its biometric utility, most often in recognition, verification, or detection pipelines. While traditionally distinct from generic NR-IQA, recent work has highlighted the value of perceptual cues in predicting biometric utility [16], [17]. Factors such as resolution, pose, occlusion, blur, compression, and illumination jointly affect both human perception and algorithmic performance, motivating hybrid approaches that bridge perceptual and task-specific definitions of quality.

Traditional FIQA methods relied on handcrafted features such as sharpness, symmetry, or inter-eye distance [102]. Recent models leverage deep face embeddings extracted from recognition networks [40], [103]. These models typically apply a regression or ranking loss to map features to quality scores correlated with recognition accuracy.

FIQA is an inherently task-dependent quality problem. Unlike general-purpose IQA, it must account for the characteristics of the recognition model and dataset. Furthermore, FIQA methods often correlate poorly with perceptual quality metrics, as an image may appear visually high-quality but remain unsuitable for matching due to pose or occlusion [22]. This motivates the need to explore perceptually grounded IQA models tailored to face data, and to investigate how traditional metrics can be adapted, fused, or compared against biometric-specific quality indicators.

Facial Quality Network (FaceQnet) [38] formulates FIQA as regression trained on proxy targets derived from recognition performance rather than MOS. For each subject, an ICAO-compliant reference image $\mathbf{x}_{\text{ref}}$ is selected. A fixed face matcher $F(\cdot, \cdot)$ computes a similarity score $s(\mathbf{x}) = F(\mathbf{x}, \mathbf{x}_{\text{ref}})$ between a candidate image $\mathbf{x}$ and its reference. Scores are normalized to $[0, 1]$ to form targets $\tilde{s}(\mathbf{x})$. A regression network r_θ takes either the image or a recognition embedding as input and predicts $\hat{q} = r_\theta(\mathbf{x})$. Training minimizes

$$\mathcal{L}_{\text{FaceQnet}} = \frac{1}{N} \sum_{n=1}^{N} \left(r_\theta(\mathbf{x}_n) - \tilde{s}(\mathbf{x}_n)\right)^2. \tag{2.15}$$

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

Stochastic Embedding Robustness for Face Image Quality (SER-FIQ) builds on the intuition that high-quality facial images produce consistent embeddings under stochastic dropout perturbations, while low-quality images result in more variable and unstable representations. Given a face image, a pre-trained face recognition model with dropout enabled generates T stochastic embeddings $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T\} \in \mathbb{R}^d$. The pairwise cosine similarities between these embeddings are then computed:

$$\text{sim}(\mathbf{e}_i, \mathbf{e}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\|\|\mathbf{e}_j\|}, \quad \forall i, j \in \{1, \ldots, T\}, i < j. \tag{2.16}$$

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

$$\text{SER-FIQ} = \frac{1}{\binom{T}{2}} \sum_{i<j} \text{sim}(\mathbf{e}_i, \mathbf{e}_j), \tag{2.17}$$

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

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

$$\mathcal{L}_{\text{MagFace}} = -\log \frac{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))}}{e^{s \cdot (\cos(\theta_y + m(\|\mathbf{e}\|)))} + \sum_{j \neq y} e^{s \cdot \cos(\theta_j)}}, \tag{2.18}$$

where θ_y is the angle between $\mathbf{e}$ and the weight vector of the correct class, s is a scaling factor, and $m(\|\mathbf{e}\|)$ is a dynamic margin that increases with quality. The final MagFace quality score is proportional to the magnitude $\|\mathbf{e}\|$, constrained during training to lie within a fixed interval $[l, u]$ for regularization.

2.5 Fusion-based IQA Models

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

Early supervised fusion combined normalized metric scores linearly. [69] introduced the Multi-Method Fusion (MMF) framework, where a regression model learns weights that map

a vector of FR-IQA scores to MOS. Non-linear learners extend this principle by modeling interactions between metrics and mitigating collinearity. In a biometric setting, [37] demonstrated that tree ensembles trained on ISO-compliant facial images improve correlation with utility-focused targets. Subsequent studies confirmed that carefully designed fusion outperforms standalone metrics in Pearson Linear Correlation Coefficient (PLCC) [104] and Spearman Rank Order Correlation Coefficient (SRCC) [105] against human judgments [22].

When large-scale MOS is scarce, weak supervision turns to proxy signals and relative constraints, the simplest way to achieve this constists into fusing multiple FR-IQA outputs into proxy scores via averaging or reliability-weighted fusion [106]. Other methods rely on semi-supervised and positive unlabeled objectives within FR-IQA [107]. A complementary path replaces absolute scores with pairwise or listwise orderings induced by controlled degradations [41]. Cascaded predictors trained on these proxy metrics can bootstrap quality estimation when subjective ground truth is limited [108].

We refer the reader to Appendix C for concise per-method descriptions and discussion.

2.6 Image distortion types

Image distortions in IQA are typically organized into operational families that mirror failures across capture, processing, compression, transmission, and display, with taxonomies grounded in vision science and formal subjective testing protocols [2], [10], [12], [13]. Core classes include lossy compression artifacts from transform quantization and entropy coding, blur from optical defocus or motion, stochastic noise from sensor and electronics, geometric and resampling effects such as aliasing and interpolation blur, and photometric or color shifts induced by exposure, tone mapping, and white-balance errors. Real acquisition pipelines often produce authentic mixtures of these mechanisms, especially in mobile, webcam, and surveillance scenarios [3], [14], [15].

2.6.1 Lossy compression

Lossy compression artifacts arise from transform quantization and entropy coding to reduce bitrate. Joint Photographic Experts Group (JPEG) [109] is dominant for images and induces blocking, ringing, and loss of high frequency detail, while modern codecs vary the artifact profile but share similar roots in quantization and in loop filtering [5], [6]. These artifacts are ubiquitous in social media uploads, cloud messaging, and browser pipelines where repeated encode-decode cycles may accumulate damage. The effect is perceived as block boundaries, staircase edges, haloing near strong transitions, and texture washing out fine facial features [1], [10].

2.6.2 Blur and Detail Loss

Blur covers optical defocus from finite depth of field and focusing errors, Gaussian like point spread from lens or sensor integration, and motion blur caused by camera or subject movement. These effects smear edges, reduce acutance, and degrade recognizability of small structures [12], [65]. Real world drivers include handheld shooting, long exposures in low light, rolling shutter readout, and surveillance cameras with slow shutters or compression integrated stabilization [3], [47].

2.6.3 Additive Noise

Additive noise includes sensor read noise, signal dependent shot noise, thermal noise, and coarse impulse like contamination from faulty pixels or transmission induced bit flips. It manifests as grain, color speckle, or salt and pepper dots that mask texture and create false edges [4], [65]. Noise dominates in dim scenes, small pixel sensors, and high ISO capture and interacts non-linearly with denoising and sharpening in the camera pipeline [3]. Lab datasets simulate Gaussian, Poisson, and impulse processes across standard deviations or rates [43], [47], [50].

2.6.4 Resampling and Resolution Variations

Resampling and resolution variations introduce aliasing, moiré, interpolation blur, and haloing when images are downscaled or upscaled without ideal prefiltering. Repeated resampling in social platforms or display adaptation pipelines compounds these effects [53], [65]. Edge directed sharpening and unsharp masking can further create overshoot undershoot ripples, perceived as halos around high contrast structures [10].

2.6.5 Photometric and Color Distortions

Photometric and color distortions cover exposure errors, gamma mis calibration, global and local contrast changes, white balance shifts, channel noise imbalance, and saturation clipping. These arise from auto exposure and auto white balance failures, tone mapping, and display pipeline mismatches [3], [110]. Perceptually, these alter scene realism, skin tone fidelity, and local visibility of detail. Benchmark datasets include luminance and contrast change variants, color shifts, and mean intensity shifts to probe metric sensitivity [43], [50].

2.6.6 Transmission and Packet-Level Artifacts

Transmission and packet-level artifacts occur when compressed bit streams traverse unreliable channels. Packet loss, bit errors, or jitter map to block corruption, freezing, or tearing in images

and video, especially over wireless links or congested networks [7], [8]. Although less common in still images than in streaming video, partial downloads and progressive JPEG decoding can leave visible blocky regions or coarse approximations that degrade perceived quality [5], [45].

2.6.7 Authentic Mixtures

Authentic mixtures combine several mechanisms in the wild: for example, a low light smartphone selfie uploaded to a social network may exhibit high ISO noise, heavy denoising and sharpening, aggressive JPEG compression, slight motion blur, and color cast. In surveillance, long exposure blur, compression, and downsampling can occur [42], [44], [54]. These combinations motivate evaluating IQA models beyond single distortions and under realistic capture conditions, using subjective methodologies aligned with standards such as ITU-R BT.500-15 [13] and comprehensive performance studies of metric behavior across datasets and artifact families [1], [2].

2.7 Steganography

Steganography has evolved significantly with the rise of digital media and advanced compression algorithms, leading to a variety of embedding methods and detection strategies. Classical steganographic methods can be broadly categorized into spatial, frequency, and adaptive domain techniques. Spatial domain methods directly manipulate pixel values, most notably through least significant bit substitution [31], to encode binary data within an image. Frequency domain methods operate on transformed representations of the image, using tools such as the DCT or Discrete Wavelet Transform (DWT), embedding data in coefficients less sensitive to compression [111]. Adaptive methods further refine the previous approaches by analyzing local characteristics of the image, dynamically choosing to embed regions to minimize perceptual distortion [112].

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

Recent advances employ deep learning to design robust, end-to-end steganographic pipelines that simulate print-scan degradations during training. We highlight four of these methods that represent the current state of the art in printer-proof steganography for natural and facial images. The printer-proof steganography methods used were based on Generative Adversarial Networks (GANs) [113] to encode and decode information. Various results can be obtained

depending on the method used.

2.7.1 StegaStamp

StegaStamp [114] introduced the first deep learning-based pipeline for robust steganography under real-world distortions. It jointly trains an encoder-decoder architecture with simulated perturbations, such as blur, color shift, projective warps, and JPEG compression, to mimic the print-scan process. A random bitstring is embedded into the image, and the decoder learns to retrieve it despite the aforementioned degradations. The training loss combines a cross-entropy bit loss, an L_2 reconstruction loss, and a perceptual LPIPS loss to preserve visual quality. StegaStamp showed that it is possible to reliably recover short messages (e.g., 56 bits) from physically printed and scanned images, setting a benchmark for printer-proof robustness.

2.7.2 Code Face

Code Face [115] extends StegaStamp with a focus on facial images used in ID documents. It introduces facial-specific modules such as a Spatial Transformer Network (STN) for geometric alignment and a perceptual identity-preserving loss based on FaceNet embeddings. This ensures that the stego face remains visually and biometrically consistent with the original. The network is trained adversarially, with a discriminator encouraging naturalness and a decoder optimizing message recovery under print-scan distortions. Code Face proves to be effective for embedding messages in high-resolution facial portraits, making it suitable for privacy-preserving biometric IDs.

2.7.3 RiemStega

RiemStega [116] innovates by introducing a Riemannian manifold-based loss. Instead of comparing pixel values, it represents images through Symmetric Positive Definite (SPD) covariance matrices and minimizes the affine-invariant Riemannian distance between the cover and stego image descriptors. This statistical approach preserves global structural and texture patterns more robustly under degradation. The method maintains high visual fidelity while increasing message recoverability post-printing, and demonstrates strong performance on both generic and facial image datasets.

2.7.4 StampOne

StampOne [117] incorporates frequency-domain awareness by applying DWT and gradient maps to both image and message during encoding. The model balances frequency bands via a learned

attention mechanism and includes frequency-domain discriminators in its loss functions. This results in encoded images with spectral properties closely matching those of the original, enhancing resilience to printer-related distortions such as color shifts and sensor noise. StampOne outperforms prior methods in both visual similarity and decoding accuracy, especially for high-frequency facial features.

3 Methodology

3.1 Dataset

Our dataset is derived from the publicly available FRLL [64] set, comprising 102 full-color, 1350×1350 frontal, ICAO-compliant facial images. Self-reported age, gender and ethnicity are included. Attractiveness ratings (on a 1–7 scale from "much less attractive than average" to "much more attractive than average") for the faces from 2513 people (ages 17–90) are included as well. The FRLL dataset was selected for its high-fidelity ICAO-compliant images, controlled acquisition conditions, and extensive demographic annotations, which are crucial for analyzing perceptual quality and potential demographic biases in subjective ratings. Each image was encoded using four printer-proof steganographic methods, each applied at nine different intensity levels, $x = 0.6 + 0.1k$ for $k \in \{0, \ldots, 8\}$, yielding a total of 3,672 distorted images.

Table 3.1: Partitioning of the steganographically distorted FRLL dataset.

Subset	**Subjects**	**Images**	**Purpose**
MOS train set	12	432	Train FR fusion model
MOS test set	3	108	Final evaluation
pMOS set	87	3,132	Generate proxy MOS labels
NR train set	99	3,564	MOS train + pMOS
Total	**102**	**3,672**	—

The dataset was partitioned to support both supervised learning and independent evaluation. A small, demographically diverse subset of 15 identities, shown in Figure 3.1 was annotated with MOS through subjective tests. This set was split into a training portion (12 identities, Figure 3.1a) used to fit the fusion model, and a disjoint test portion (3 identities, Figure 3.1b) held out for final evaluation. The remaining 87 identities formed the unlabeled set, to which pseudo-Mean Opinion Score (pMOS) were later assigned. Together, the annotated and pseudo-labeled subsets were used to train the NR-IQA model. A summary of these partitions, including the

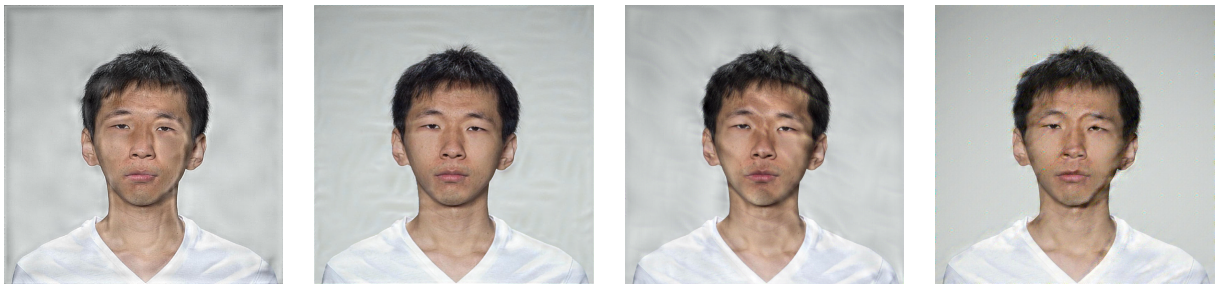
number of subjects and images per subset, is presented in Table 3.1.



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

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

The steganographic distortions are visible in Figure 3.2, which shows examples of distorted facial images from each method. The distortions are applied to the original images, and the resulting stego images are used for both subjective evaluation and training of the NR-IQA model. The selected steganographic methods represent diverse classes of distortion processes (geometric, color-based, local texture perturbation) commonly encountered in practical steganographic applications. The chosen intensity levels ensure a broad perceptual range from barely noticeable to clearly visible distortions.



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

Figure 3.2: Steganographically distorted facial images from each method.

3.2 Collection of Subjective Scores

We followed the ITU-R BT.500-15 [13] recommendation and adopted the Single Stimulus method. The test was implemented using a custom Django [118] web application seen in Figure 3.3. Prior to the test session, participants signed an informed consent form and filled out a registration form providing demographic and environmental information such as age, gender, education level, country of origin, ethnicity, type of device being used, and where they were located at the time of the test session. After this, observers were shown 10 randomly selected sample images. These examples were designed to help them understand the variation in quality across the dataset.

In the subjective assessment phase, each image was shown individually, with no time limit. Ratings were submitted using a labeled slider, and automatic saving ensured session robustness.



Figure 3.3: Django-based web application created for the Single Stimulus test sessions.

Each image in the MOS set was evaluated approximately 30 times by human observers, resulting in over 14,000 ratings. We had around 200 participants, each session lasted about 22 minutes and included roughly 70 evaluations. Table 3.2 summarizes the demographic profile of the participant pool. Following the session, outlier observers were identified and removed using both Kurtosis-based and correlation-based post-screening methods described in ITU-R BT.500-15, resulting in the exclusion of one participant.

The database was implemented using the Django web framework, which provides an Object-Relational Mapping (ORM) layer that directly maps Python model classes to database tables.

Table 3.2: Demographic and non-demographic profile of the observers.

(a) Age group

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

(b) Gender

Gender	Count
Male	143
Female	59

(c) Ethnicity

Ethnicity	Count
White	177
Latino	8
Black	6
Multiracial	3
Other	7

(d) Education Level

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

(e) Country of Origin

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

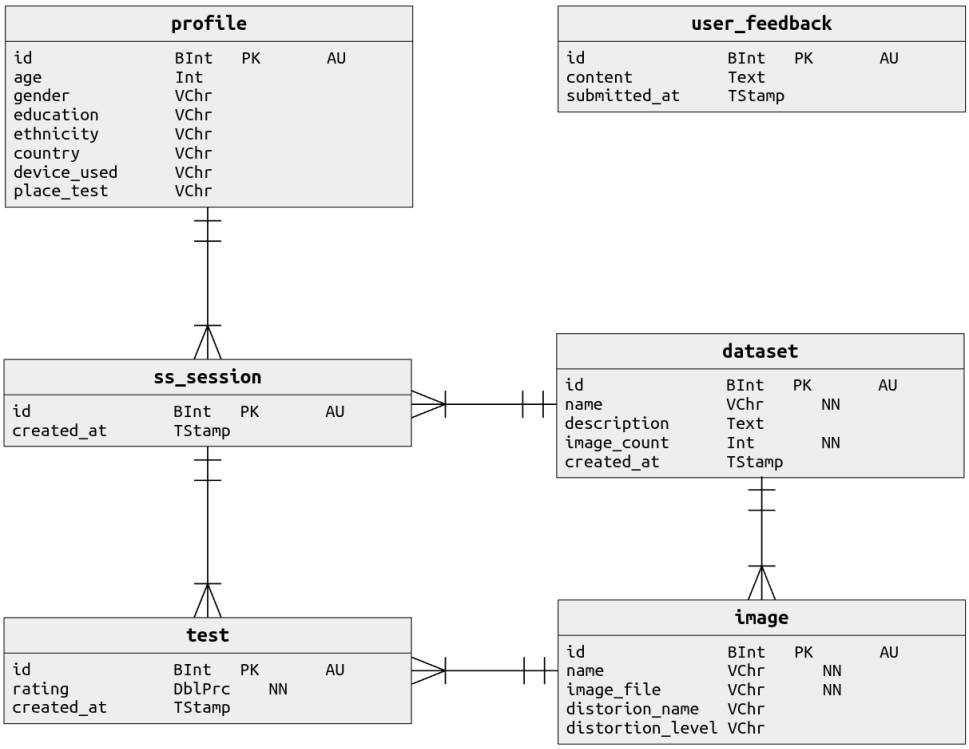
(f) Device Used

Device	Count
Laptop	94
Phone	65
Desktop	42
Other	1

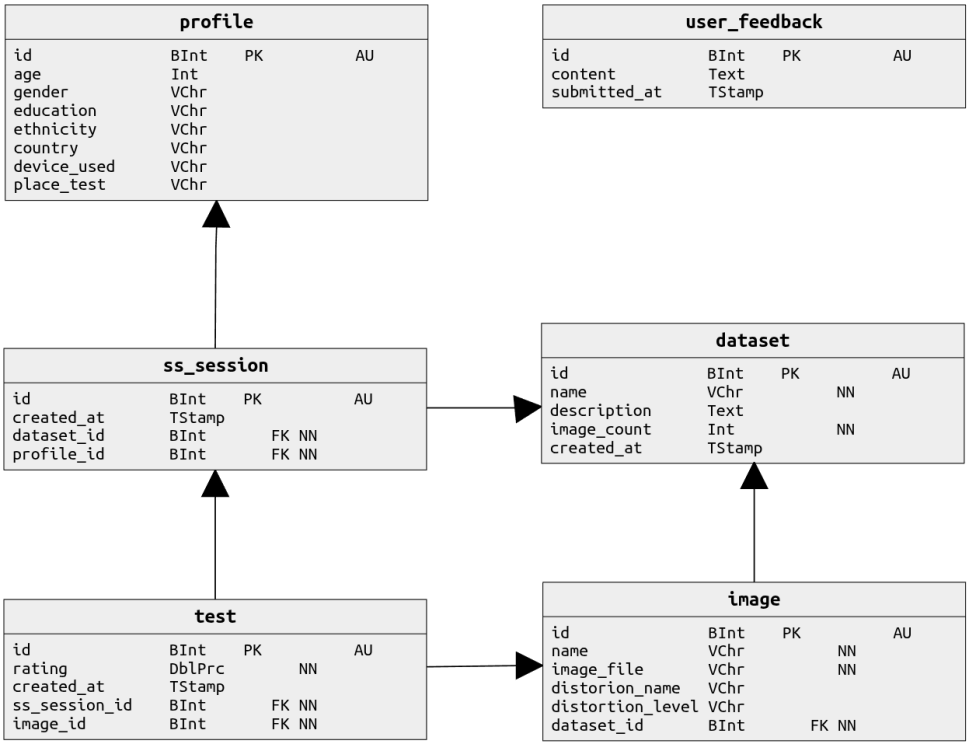
The ORM also enforces data integrity constraints and simplifies querying for statistical analysis. The database management system used was PostgreSQL.

The database schema was designed to ensure structured storage and traceability of all subjective quality evaluation data. Figure 3.4a presents the conceptual data model, which defines the main entities and their relationships. The profile entity stores participant metadata, including age, gender, education, ethnicity, country of origin, and device used. The session models individual tests and is linked to both the participant profile and the dataset being evaluated. The image entity encodes each image's filename, associated distortion type, and distortion level. The test entity stores individual subjective ratings, including the precise timestamp of submission, linked to both the session and the image presented. An auxiliary feedback entity captures optional, anonymous, observer comments and critiques.

The physical implementation of this schema, shown in Figure 3.4b, is realized as a relational database with explicit foreign key constraints to ensure referential integrity. The dataset and profile fields in session formally enforce the linkage to specific datasets and participants. The test table connects each subjective rating to its session and image, enabling consistent aggregation of ratings into a statistical descriptor such as MOS and confidence intervals per image and per



(a) Conceptual schema.



(b) Physical schema.

Figure 3.4: Conceptual and physical database schemas implemented in the Django web application.

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

3.3 Debiasing Subjective Scores

To correct for demographic bias in the subjective scores, we followed a procedure inspired by prior work on bias correction in perceptual tasks [119], where we applied a residualization method based on linear modeling. An Ordinary Least Squares (OLS) regression was fit to the MOS values, using observer and image attributes, and their pairwise interactions as categorical predictors. The fitted bias components were subtracted from the original scores, and the residuals were mean-centered to preserve the global score distribution. As shown in Table 3.3, several factors exhibit statistically significant effects on the MOS prior to residualization, notably observer and subject ethnicity. After applying the residualization procedure, these effects disappear, as confirmed by an Analysis of Variance (ANOVA) [120] and Multivariate Analysis of Variance (mANOVA) [121] test, showing no significant impact from any individual factor. The corrected MOS labels are then used as ground truth in all supervised stages of the pipeline to ensure fairness and reduce the influence of socially conditioned priors.

Table 3.3: ANOVA and mANOVA results for observer and image attributes. Before debiasing, several factors show statistically significant effects on MOS, p-value < 0.05. After residualization, all main effects show no significant impact, confirming the effectiveness of the debiasing procedure.

Factor	**p-value**	**p-value (residualized)**
Subject attractiveness	0.3173	0.3173
Observer gender	0.022	0.9930
Observer ethnicity	8.44×10^{-4}	1
Subject gender	1.60×10^{-3}	0.9722
Subject ethnicity	7.36×10^{-3}	1
Observer gender $\times$ Subject gender	0.6417	0.6417
Observer ethnicity $\times$ Subject ethnicity	0.0582	0.0582

3.4 Evaluation of Full-Reference Metrics

We compute 40 FR-IQA scores for each distorted image and compare them against the corresponding MOS. To assess alignment with human perception, we evaluate both linear and monotonic relationships using PLCC and SRCC. These measures provide a ranking of metrics, IQA score, by their mean correlation. Negative correlations are kept as is, PLCC tests linear fit, while SRCC tests monotone ordering. Adding magnitudes would treat them as interchangeable, thus, rewarding inconsistent behavior. If a metric's polarity is reversed, the downstream fusion model can learn an appropriate sign or transformation during training.

Beyond pairwise correlation, we analyze the intrinsic dimensionality of the FR feature space using Singular Value Decomposition (SVD). Cumulative variance helps to choose the number of retained metrics k at a target threshold, and the Picard plot [122] validates stability by contrasting singular values with the projected response. This analysis yields a compact subset of normalized, polarity-aligned metrics, which will form the input to the fusion stage.

3.5 Full-Reference Fusion Metric

Building on the selected FR-IQA features, we construct a fusion-based metric that learns to predict perceptual quality directly from heterogeneous objective scores. Each distorted image is represented by its k-dimensional feature vector, and the corresponding human MOS serves as the regression target.

A diverse set of regressors is trained on these features:

- Linear: Linear Regression [123], Ridge [124], Bayesian Ridge [125], ElasticNet [126].
- Ensembles: Random Forest [127], Extra Trees [128], Gradient Boosting [129], Histogram Gradient Boosting [130].
- Boosted trees: XGBoost [131], LightGBM [132], CatBoost [133].

Each regressor is tuned with a Tree-structured Parzen Estimator (TPE) [134] search implemented in Optuna [135], using Successive Halving [136] for early stopping, and evaluated by 5-fold cross-validation. Models are ranked by the mean of PLCC and SRCC between predictions and MOS. Per-model outputs include predictions, scores, tuned parameters, and trained checkpoints.

For tree-based models we compute SHapley Additive exPlanation (SHAP) [137], highlighting which FR metrics drive the fused prediction and whether their contributions are consistent across

the dataset. The result is a fusion-based FR metric that consistently outperforms any single score and serves as both a standalone contribution and the basis for pseudo-label generation.

3.6 pseudo-Mean Opinion Score Generation

The best validated fusion model is then applied to the unlabeled portion of the distorted dataset to produce pMOS labels. Each image is passed through the selected top-k metrics, normalized with the same scaler used in training, and scored by the fusion model.

This procedure extends perceptual supervision from the small human-annotated subset to the entire dataset, producing a complete set of labels for downstream use. These pseudo labels enable two main objectives:

The first objective is FR-IQA at scale when human MOS is scarce. When a pristine reference is available but exhaustive subjective scoring is not feasible, the fusion metric provides pMOS for all distorted versions of a reference image. This makes it possible to analyze distortion sensitivity curves, compare distortion methods across payload levels, and generate robust statistical summaries without collecting new subjective scores. In practice, this establishes a reproducible way to extend human perception to large datasets with minimal annotation effort.

The second objective is supervision for training NR-IQA models. With a large quantity of labeled images available, we can train NR regressors that predict perceptual quality directly from distorted images without a reference. The pseudo labels provide the scale and coverage needed to fit modern deep backbones, and the resulting models operate in fully blind conditions, enabling quality assessment of images where no reference is available.

3.7 NR Metric from FR Fusion

We train on two complementary image sets: a labeled set containing the ≈10% of images annotated with human MOS, and a pseudo labeled set with the remaining distorted images scored by our FR fusion into pMOS.

We train three pretrained backbones with a 1-unit linear regressor head using Smooth-L_1 loss ($\beta = 1$), AdamW [138], cosine learning-rate scheduling, batch size of 31, mixed-precision on GPU, and ImageNet normalization. Table 3.4 summarizes the training configuration. There was no use of data augmentation algorithms, since the distortions themselves provide sufficient variability.

We adopt 5-fold Cross-Validation (CV). The merged dataset is split into 5 disjoint folds $\{F_1, \ldots, F_5\}$ at the image level, stratification not enforced. For split i, training uses $D_{\text{train}} =$

Table 3.4: Training configuration of pretrained backbones for NR metric learning.

Backbone	**Input**	**Epochs**	**Learning rate**	**Weight decay**	**Drop rate**
ResNet-50 [139]	224^2	50	1×10^{-3}	1×10^{-4}	0
EfficientNet-B1 [140]	240^2	60	1×10^{-3}	1×10^{-5}	0.2
ViT-B/16 [141]	224^2	60	3×10^{-4}	0.05	0.1

$D \setminus F_i$ and validation uses $D_{\text{val}} = F_i$. All hyperparameters are kept fixed. We monitor validation MAE after every epoch and apply early stopping with patience p epochs. For each fold we retain the checkpoint with the lowest validation MAE. Let m_i denote a metric on fold i, we report the mean and dispersion across the 5 folds:

$$\bar{m} = \frac{1}{5}\sum_{i=1}^{5} m_i, \quad \hat{\sigma} = \sqrt{\frac{1}{4}\sum_{i=1}^{5}(m_i - \bar{m})^2}, \quad \text{CI}_{95}: \ \bar{m} \pm 1.96\,\frac{\hat{\sigma}}{\sqrt{5}}. \tag{3.1}$$

Final test set results are reported on a disjoint MOS-only set. Full hardware and software environment details are in Appendix D.

4 Results

This chapter reports the experimental validation of the proposed framework. The FRLL dataset is used as the primary benchmark, providing high-quality facial images with controlled steganographic distortions and subjective opinion scores. It allows us to test the alignment of individual FR metrics with human judgments, assess the benefits of fusion, and explore the use of pseudo-MOS to supervise NR models, when image content is relatively homogeneous, and distortions are subtle.

We further evaluate the framework on the TID2013 database, which includes a wide variety of reference images and synthetic distortions. Unlike the face-focused FRLL set, this benchmark spans diverse content and impairment types, providing a complementary test of generalization and robustness. Crucially, since ground-truth MOS is available for all distorted images, TID2013 allows us to directly test the alignment of pMOS with true human scores, offering an independent validation of the fusion stage.

4.1 Face Research Lab London Dataset Pipeline

The evaluation begins with subjective test outcomes, then examines the correlation of individual FR metrics with human perception, advances to the construction and validation of the fusion-based FR metric, and finally demonstrates how pMOS supervision supports the training of NR models.

4.1.1 Subjective Test Outcomes

We computed per-method descriptive statistics over the MOS subset. Table 4.1 reports the mean MOS and the mean of per-image standard deviations (mean std) for each steganographic method. Since all methods were evaluated under the same conditions and distortion levels, these aggregates enable a direct comparison of their perceptual impact. The results indicate a similar performance across methods, with Code Face achieving slightly higher mean MOS.

Table 4.1: Per-method descriptive statistics on the MOS subset.

Method	**mean MOS**	**mean std**
Code Face	63.10	20.69
RiemStega	42.89	19.98
StampOne	46.28	19.79
StegaStamp	43.16	19.45

4.1.2 Correlation of FR Metrics with MOS

To quantify alignment, we calculate both PLCC and SRCC, producing a ranking by their mean correlation. Figure 4.1 visualizes the scatter plots across metrics, showing their varying consistency with human perception. Several metrics exhibit strong linear or monotonic trends with human opinion, while others align weakly or negatively. Some metrics, such as MS-SSIM and PSNR, exhibit strong monotonic trends with MOS, while others like UQI or VIF capture only partial aspects of perceptual quality. Deep learning-based metrics, such as LPIPS, DISTS, and error-based metrics, like MSE, RMSE, show weaker or even negative alignment, often failing to separate high- from low-quality images. Other metrics, such as D_λ, SR-SIM and IW-SSIM, don't show clear trends, indicating not to be reliable indicators of perceptual quality in this context, for either facial images or facial distortions.

This heterogeneity highlights that no single FR-IQA metric consistently matches human judgments across all distortion types, which motivates combining complementary metrics into a fused predictor.

The cumulative explained variance from SVD from Figure 4.2a shows that, in order to explain 68%, 95%, and 99.7% of the variance, we would need to retain $k = 2$, $k = 6$, and $k = 14$ metrics, respectively. We chose $k = 6$ as it provides a robust trade-off between expressiveness and stability. This dimensionality reduction ensures that the fusion model leverages the most informative components of the metric space, while avoiding overfitting and instability associated with higher dimensions. The complementary Picard plot in Figure 4.2b further validates this choice by contrasting the decay of singular values with the magnitude of the response projections. Both curves decrease in parallel up to approximately the sixth component, beyond this point, the singular values decay more rapidly than the projections, revealing that additional components would amplify noise rather than contribute meaningful variance. This joint evidence confirms that truncating at $k = 6$ achieves a balance between expressive power and robustness, while avoiding instability from ill-conditioned directions.

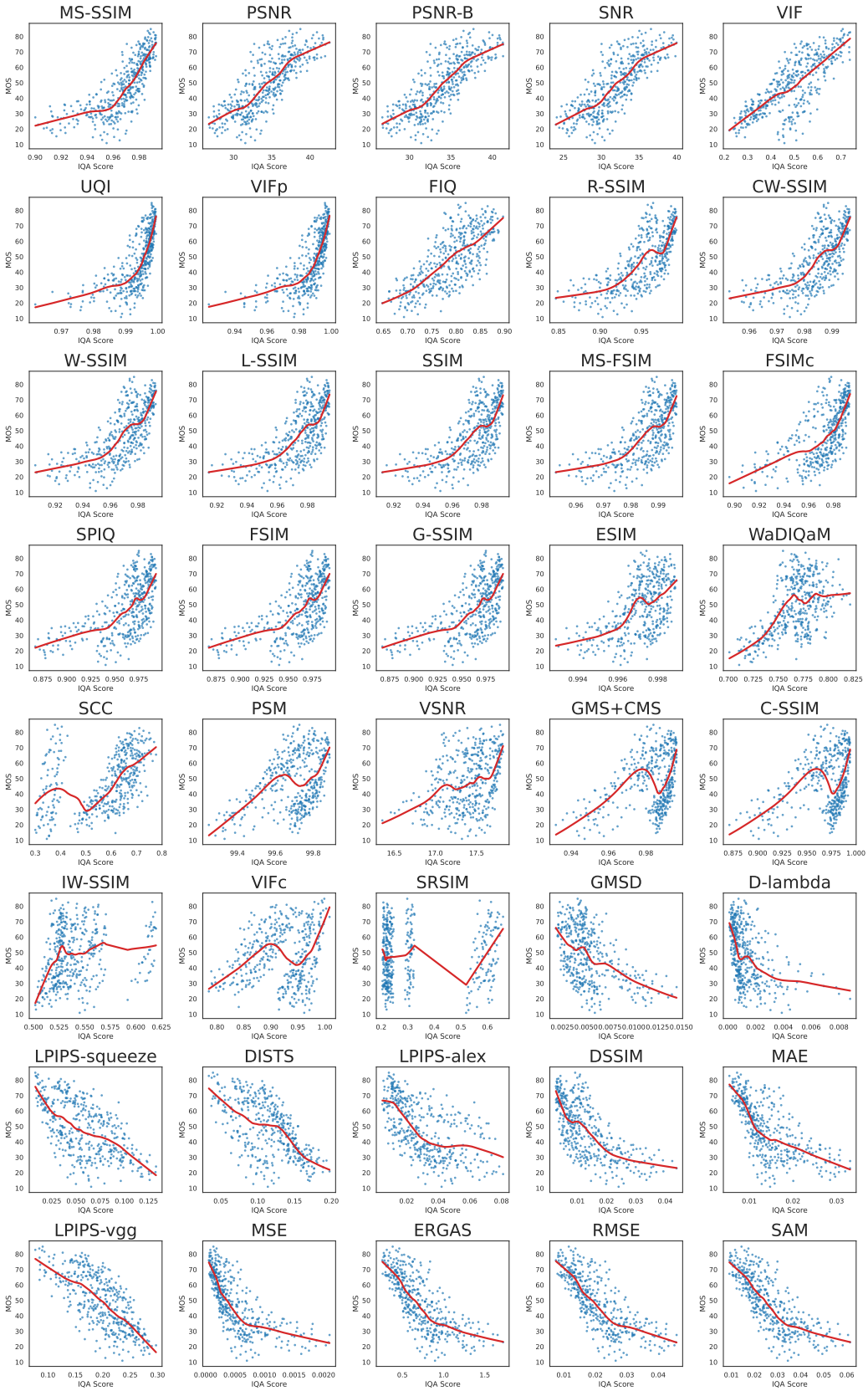
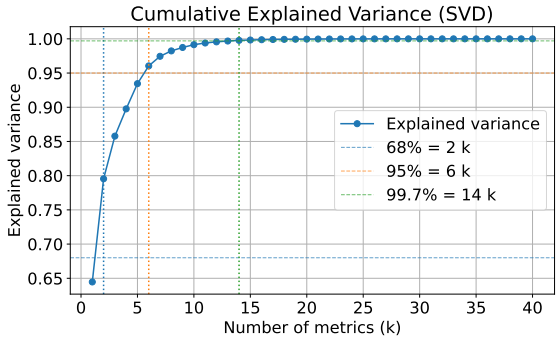
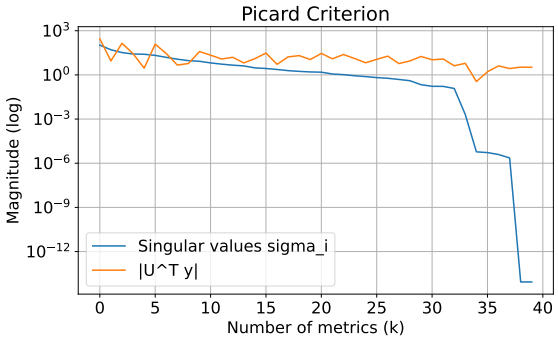


Figure 4.1: Scatter plots of MOS versus 40 individual FR-IQA metrics. The red curve is a local smoothing of the trend.



(a) Cumulative explained variance from SVD. Vertical markers indicate k at 68, 95, and 99.7 percent.



(b) Picard plot showing singular values and the magnitude of the target projections.

Figure 4.2: SVD-based analysis of the FR metric space. The variance curve selects k at the chosen threshold, and the Picard plot corroborates stability near that truncation.

The individual metric results in Table 4.2 shows the PSNR achieves the highest linear correlation, PLCC = 0.8009, and the lowest error scores, reflecting its sensitivity to pixelwise differences in relatively controlled distortions. In contrast, MS-SSIM shows the strongest monotonic correlation, SRCC = 0.8489, confirming its ability to preserve rank order consistency even when absolute scores differ. Variants such as PSNR-B and SNR perform close to PSNR, while VIF lags behind despite its information-theoretic basis. Interestingly, UQI achieves an SRCC = 0.8085, but with weaker linear correlation and elevated error.

Table 4.2: Individual FR-IQA metrics versus MOS on the unbiased train set.

Metric	**PLCC**	**SRCC**	**MSE**	**MAE**
MS-SSIM	0.7712	**0.8489**	2542.00	47.27
PSNR	**0.8009**	0.8132	**420.70**	**16.57**
PSNR-B	0.7996	0.8126	434.80	16.84
SNR	0.7931	0.8047	493.90	18.08
VIF	0.7723	0.7670	2584.00	47.75
UQI	0.6865	0.8085	2540.00	47.24

4.1.3 Full-Reference Fusion

The results in Table 4.3 show a clear gain from fusing multiple FR metrics compared to relying on any single one. The fusion models trained at $k = 6$ consistently exceed 0.81 in both PLCC and SRCC, while reducing error scores by nearly an order of magnitude.

Among regressors, ensemble tree methods dominate: CatBoost achieves the best overall

performance, closely followed by GradientBoosting and XGBoost. The stability across these ensembles suggests that the fusion metric is robust to the choice of regressor, provided sufficient model complexity. Linear baselines (Ridge, ElasticNet, BayesianRidge, Linear) trail behind, confirming that non-linear interactions among metrics are important.

Table 4.3: Fusion regressors at $k = 6$ on the unbiased split.

Model	**PLCC**	**SRCC**	**MSE**	**MAE**
CatBoost	**0.8932**	**0.8950**	**62.28**	**6.046**
GradientBoosting	0.8901	0.8905	63.98	6.100
XGBoost	0.8891	0.8915	64.55	6.157
ExtraTrees	0.8871	0.8892	65.65	6.131
RandomForest	0.8861	0.8879	66.25	6.094
LightGBM	0.8838	0.8858	67.66	6.313
HistGradientBoosting	0.8760	0.8790	71.94	6.507
Ridge	0.8131	0.8353	105.71	7.941
ElasticNet	0.8129	0.8354	106.55	7.983
BayesianRidge	0.8145	0.8302	103.64	7.761
Linear	0.8147	0.8284	103.55	7.754

The best fusion model is CatBoost with 3,000 iterations, learning rate 0.0118, depth 10, and ℓ_2 leaf regularization 1.005, selected by Optuna. To interpret the fitted model, we compute SHAP values, shown in Figure 4.3. The analysis reveals that MS-SSIM and UQI provide the strongest contributions to the fused prediction, confirming their strong alignment with human MOS observed in the single-metric analysis. VIF, PSNR-B, SNR, and PSNR have mixed positive and negative impacts depending on the distorted image.

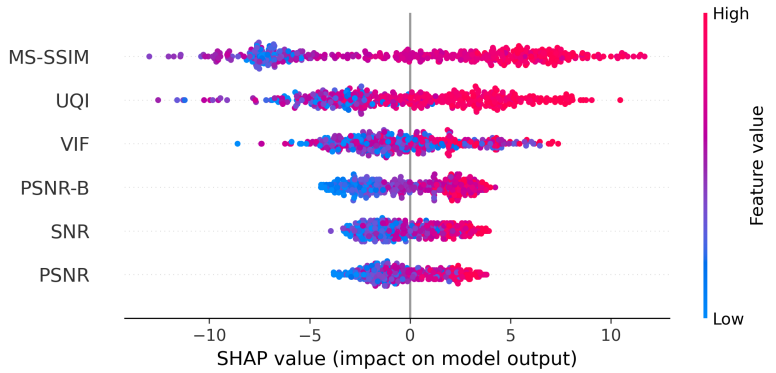


Figure 4.3: SHAP summary plot for $k = 6$ on the best CatBoost fusion model.

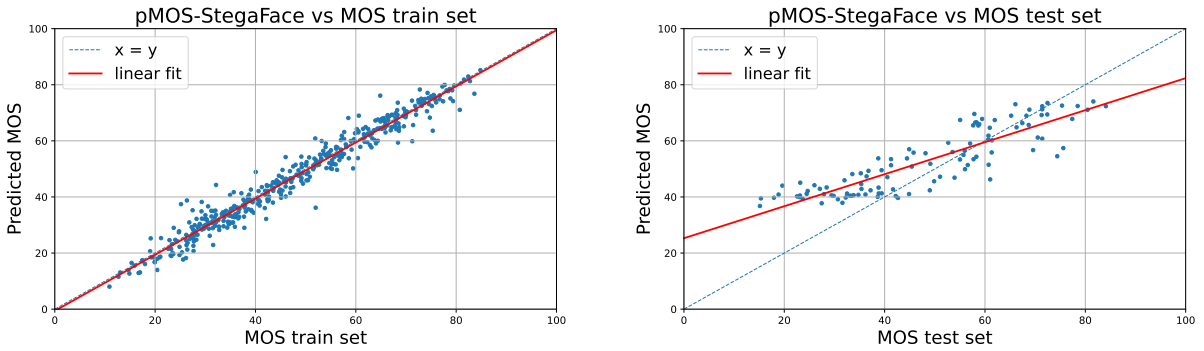
4.1.4 No-Reference Model

Based on the backbone comparison in Table 4.4, we selected ResNet-50 as the foundation for our NR-IQA model. The resulting metric, pseudo-Mean Opinion Score for Steganographic affected Face Images (pMOS-StegaFace), is named to reflect its reliance on pMOS from the FR fusion and its focus on facial datasets distorted by steganography.

Table 4.4: Backbone comparison for training pMOS-StegaFace. Report 5-fold mean PLCC, SRCC, MSE, and MAE.

Backbone	**PLCC**	**SRCC**	**MAE**	**MSE**
ResNet-50	**0.9765**	**0.9743**	$\mathbf{7.48 \times 10^{-4}}$	**0.0194**
EfficientNet-B1	0.9509	0.9512	1.77×10^{-3}	0.0298
ViT-B/16	0.9705	0.9685	9.19×10^{-4}	0.0216

Figure 4.4 shows two scenarios. On the MOS train set with ground-truth MOS we obtain PLCC = 0.9810, SRCC = 0.9800, MSE = 1.24×10^{-3}, MAE = 0.0257. On the disjoint test set we obtain PLCC = 0.8747, SRCC = 0.8768, MSE = 0.0108, MAE = 0.0847, confirming generalization to unseen human labels and alignment with the pseudo supervision.



(a) Predicted MOS vs ground-truth MOS from the train set.

(b) Predicted MOS vs ground-truth MOS from the disjoint test set.

Figure 4.4: pMOS-StegaFace predictions on train and test sets.

Table 4.5 benchmarks our learned NR-IQA metric, pMOS-StegaFace, against widely used no-reference baselines. pMOS-StegaFace delivers the strongest linear and rank correlations while also achieving the lowest error.

Table 4.5: pMOS-StegaFace versus NR-IQA baselines on the test split.

Metric	**PLCC**	**SRCC**	**MSE**	**MAE**
pMOS-StegaFace (ours)	**0.8747**	**0.8768**	**0.0118**	**0.0847**
NIQE	0.7536	0.7431	6859	82.62
PIQE	0.2993	0.3114	3297	57.05
BRISQUE	0.1234	0.5678	9101	23.45
FaceQnet	0.4567	0.4789	5123	34.12
SER-FIQ	-0.1648	-0.1542	123.5	9.230
MagFace	-0.6095	-0.6362	4437	66.24

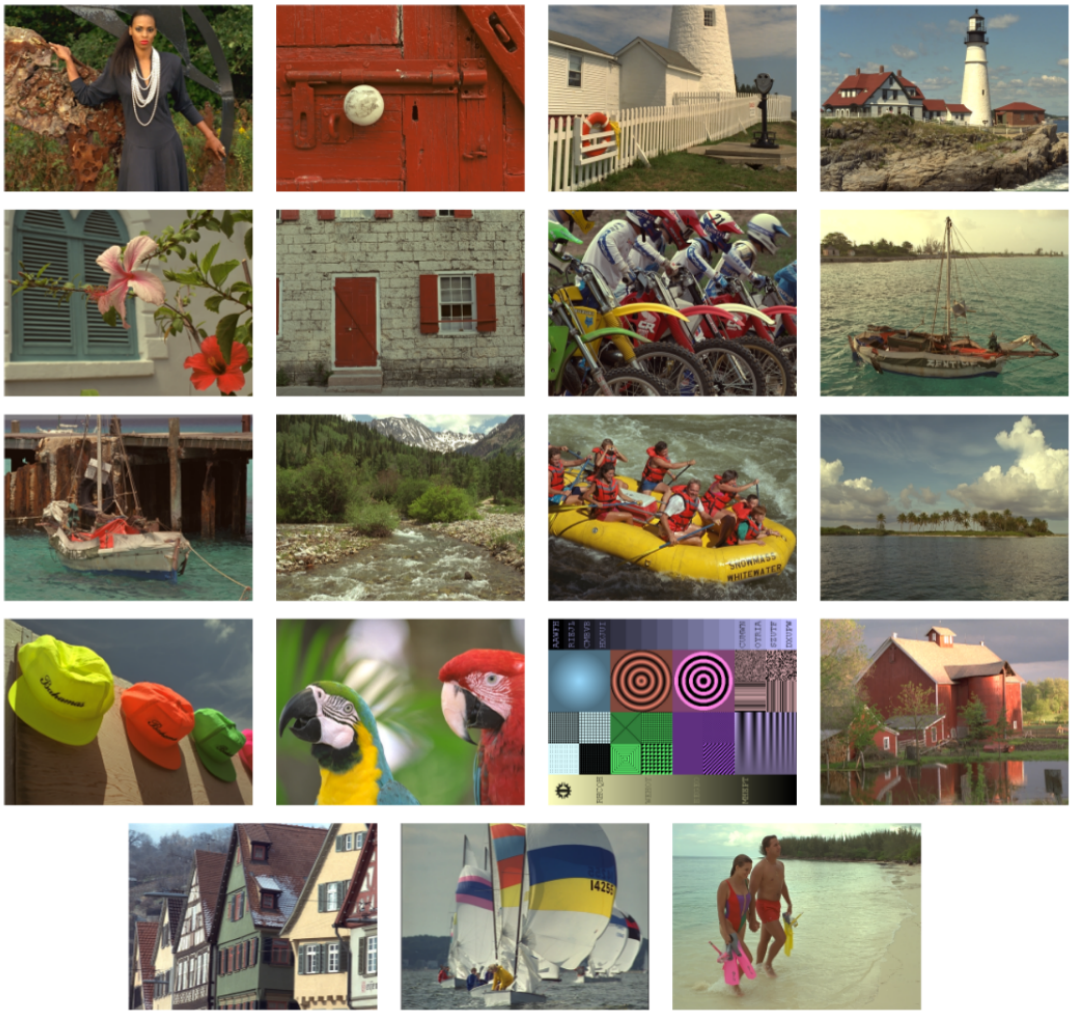
4.2 Tampere Image Database 2013 Pipeline

We repeated the FR fusion and the NR training on the TID2013 database. We followed a very similar pipeline as in section 4.1.

The TID2013 database [50] contains 3,000 distorted images derived from 25 reference images, shown in Figure 4.5, with 24 distortion types at 5 levels of intensity. The dataset covers a wide spectrum of distortion types. Noise-related degradations include additive Gaussian noise, noise in color components, spatially correlated noise, masked noise, high-frequency noise, impulse noise, quantization noise, and multiplicative Gaussian noise. Blur and restoration artifacts are represented by Gaussian blur and image denoising. Compression-related distortions comprise JPEG compression, JPEG2000 compression, and their associated transmission errors. Additional artifacts include non-eccentric pattern noise, local block-wise distortions, mean shift, and contrast changes. Color-related impairments are captured through changes in saturation, image color quantization with dithering, and chromatic aberrations. Finally, mixed and compound degradations are represented by comfort noise, lossy compression of noisy images, and sparse sampling with reconstruction.

A total of 971 observers performed 524,340 comparisons of visual quality of distorted images or 1,048,680 evaluations of relative visual quality in image pairs.

We split the dataset in a similar way we did previously with FRLL, as seen in Table 4.6, but this time we have the true MOS labels from the distorted set, so we can draw conclusions on the performance of our FR fusion model on the distorted set, as well as the performance of the NR model trained on mixed supervision. We randomly select 3 reference images to serve as the training set for the FR fusion model, and left 3 reference images for the final evaluation of the NR model. The remaining 19 reference images are used to generate pMOS labels for the



(a) pMOS set.



(b) MOS train set.



(c) MOS test set.

Figure 4.5: Reference images from the TID dataset.

unlabeled set.

Due to the nature of the images, we skipped the debiasing step.

Table 4.6: Partitioning of the TID dataset.

Subset	**Reference**	**Images**	**Purpose**
MOS train set	3	360	Train FR fusion model
MOS test set	3	360	Final evaluation
pMOS set	19	2,280	Generate proxy MOS labels
NR train set	22	2,640	MOS train + pMOS
Total	**25**	**3,000**	—

4.2.1 Subjective Test Outcomes

We computed the per-method descriptive statistics over the dataset. Table 4.7 reports the mean MOS and the mean of per-image standard deviations (mean std) for each distortion type. It reveals clear differences in perceived quality and in rating consistency across distortion families. Mean shift achieves the top mean MOS, followed closely by masked noise and chromatic-dominant additive noise, indicating that uniform photometric changes are judged as relatively less bothersome. By contrast, artifacts that disrupt local structure rank lowest, local block-wise distortions, sparse sampling and reconstruction, and JPEG2000-related degradations occupy the bottom tier.

Variability also differs by distortion. Local block-wise distortions show the smallest mean standard deviation, indicating consistent judgments across images. The largest variability appears for sparse sampling and reconstruction, JPEG/JPEG2000 compression, image denoising, and Gaussian blur, which are cases where perceived quality depends strongly on content and strength.

This shows that the HVS is most forgiving of uniform brightness shifts, and toughest on artifacts that break structure. Some distortions are judged very consistently, but others, like compression, denoising, blur, and sparse sampling, depend a lot on the image content and how strong the effect is.

4.2.2 Correlation of FR Metrics with MOS

We calculated the PLCC and SRCC between each metric and MOS for each distorted image and compare them against the corresponding MOS. Figure 4.6 shows heterogeneous patterns across metrics. MS-SSIM is again the best performing metric, followed by some of its variants,

Table 4.7: Per-distortion descriptive statistics on the TID2013 dataset.

Distortion	**mean MOS**	**mean std**
Additive Gaussian noise	52.82	8.558
Additive noise (chromatic > luminance)	59.69	6.91
Spatially correlated noise	41.35	9.14
Masked noise	59.53	6.70
High frequency noise	53.39	12.38
Impulse noise	44.26	6.67
Quantization noise	45.22	10.86
Gaussian blur	45.83	14.36
Image denoising	49.13	18.80
JPEG compression	48.76	18.44
JPEG2000 compression	41.50	19.73
JPEG transmission errors	50.26	13.48
JPEG2000 transmission errors	40.59	11.09
Non-eccentricity pattern noise	56.72	9.82
Local block-wise distortions	39.80	4.41
Mean shift (intensity shift)	61.73	7.13
Contrast change	58.53	14.48
Change of color saturation	47.94	6.35
Multiplicative Gaussian noise	51.50	8.41
Comfort noise	53.18	11.95
Lossy compression of noisy images	46.40	13.42
Color quantization with dither	49.45	12.69
Chromatic aberrations	54.66	14.66
Sparse sampling and reconstruction	40.98	21.85

CW-SSIM and IW-SSIM. The PSNR and MSE families clearly show a strong regime changes, where different distortion types occupy different quality bands. Some metrics, such as VIF and VIFc, show a heavy heteroscedasticity, where variance in MOS is very high for small VIF scores, the same phenomena happens for VIFp, but at the higher end of the spectrum. Deep learning-based metrics, such as LPIPS and DISTS, show, again, stronge negative correlations.

We chose $k = 12$ as it represents 95% explained variance, as shown in Figure 4.7a. This choice doesn't come as a surprise, as the TID2013 dataset contains a much wider variety of distortion types, and thus a larger number of metrics is needed to capture the relevant perceptual dimensions. The Picard plot in Figure 4.7b again confirms that truncating at $k = 12$ retains stable components, as the singular values and projection magnitudes decrease in parallel up to this point.

4.2.3 Full-Reference Fusion

The best fusion model was, again, CatBoost, this time with 1,500 iterations, learning rate 0.0881, depth 5, and ℓ_2 leaf regularization 5.628, selected by Optuna.

In the SHAP summary, from Figure 4.8, the dominant drivers are CW-SSIM, PSNR, MS-SSIM, W-SSIM and IW-SSIM, which show the widest SHAP spreads. Their direction matches intuition, high scores cluster with positive values while low scores align with negative values, meaning larger SSIM-family scores and PSNR raise predicted quality and small values penalize it. The long, asymmetric negative tails for CW-SSIM and IW-SSIM indicate non-linear penalties, modest drops in these metrics trigger strong downward corrections that help separate truly poor images. Although several SSIM variants are important simultaneously, their clouds are not identical, each variant captures a slightly different structural aspect, so the model exploits complementary, partially redundant information rather than any single SSIM.

We also observe occasional inversions, blue points with positive SHAP or red points with negative SHAP. That's a sign of feature interactions, like a high PSNR paired with poor structure, which the CatBoost fusion can handle better than a simple one-direction mapping. In practice, most of the lift comes from combining structure-aware SSIM variants with PSNR for overall noise/blur, then letting specialist cues, like ESIM, SPIQ, L-SSIM, and PSM, tidy up the tricky edge cases.

The FR fusion model was evaluated on the MOS train set and compared with the top $k = 12$ individual FR metrics in Table 4.8. We will refer to the fusion model as pseudo-Mean Opinion Score based on enhanced SSIM family metrics (pMOS-SSIM+), as it relies heavily on SSIM-inspired metrics, and requires a reference image to compute pMOS. The fusion model produced

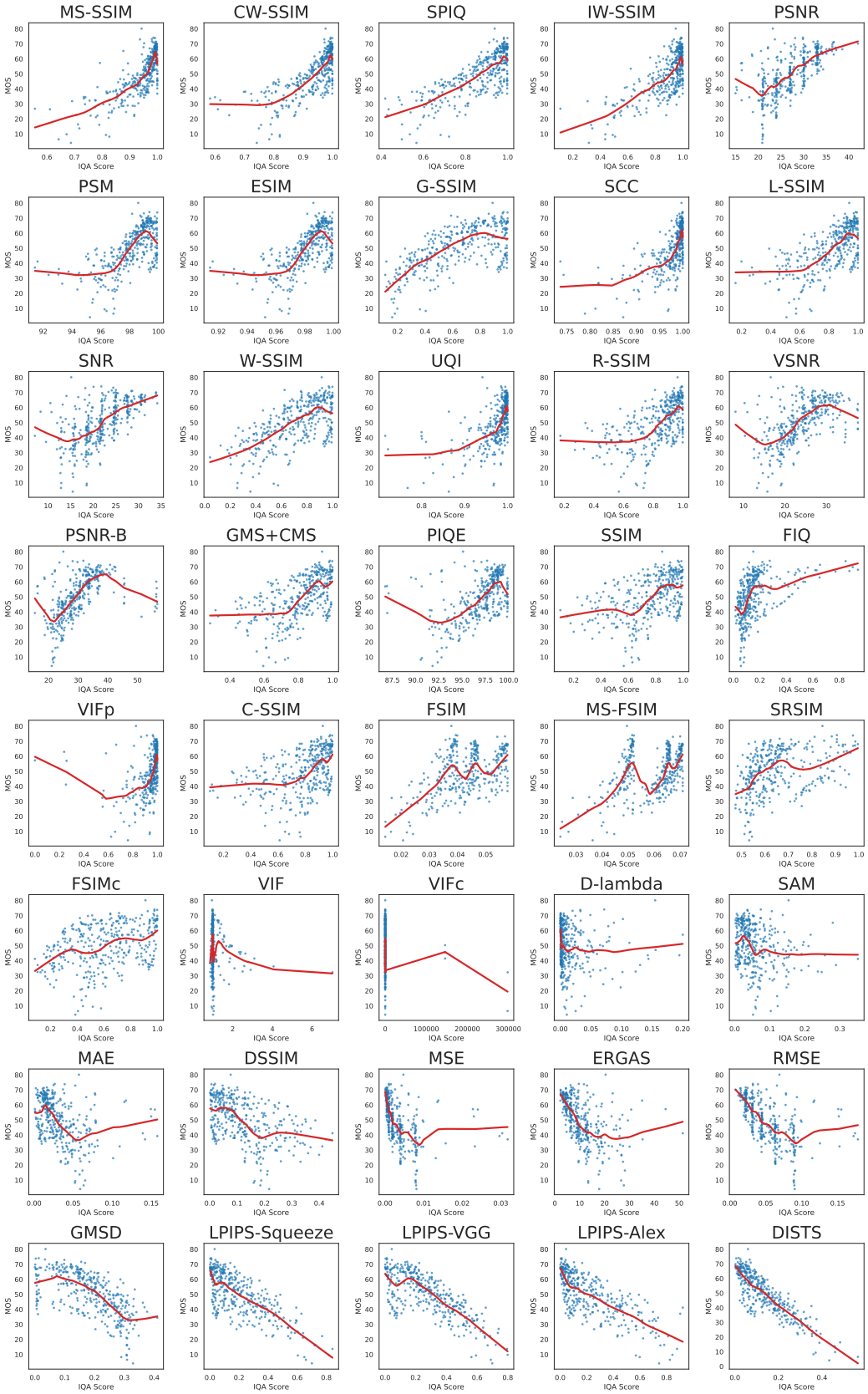
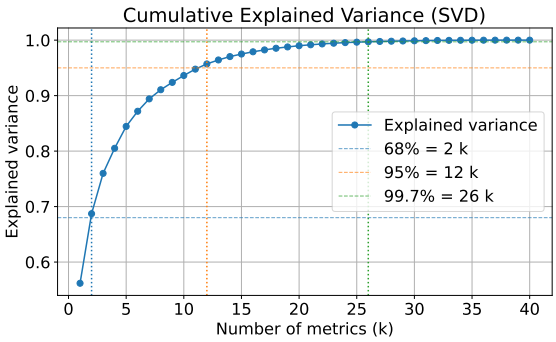
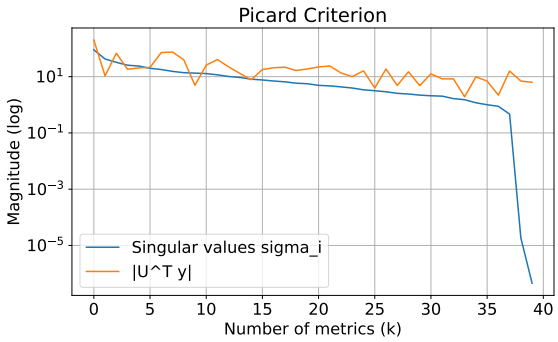


Figure 4.6: Scatter plots of MOS versus 40 individual FR-IQA metrics. The red curve is a local smoothing of the trend.



(a) Cumulative explained variance from SVD. Vertical markers indicate k at 68, 95, and 99.7 percent.



(b) Picard plot showing singular values and the magnitude of the target projections.

Figure 4.7: SVD-based analysis of the FR metric space. The variance curve selects k at the chosen threshold, and the Picard plot corroborates stability near that truncation.

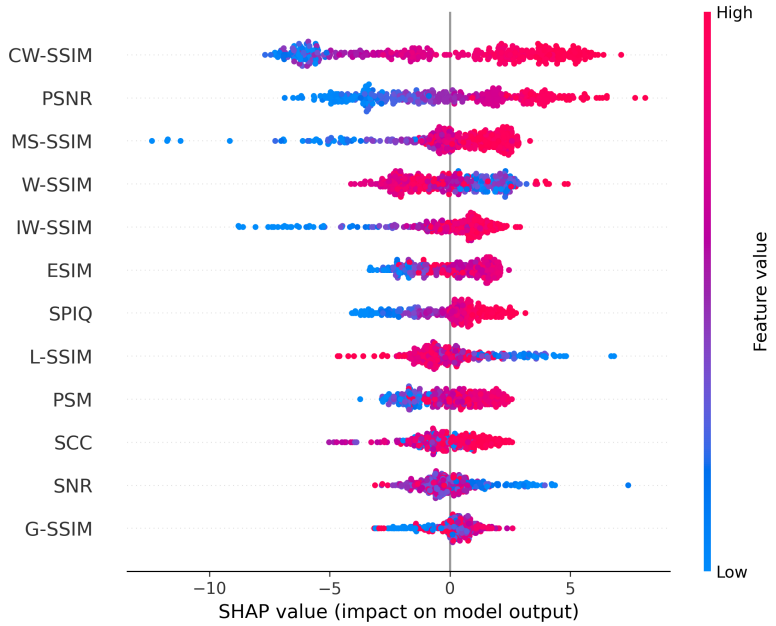


Figure 4.8: SHAP summary plot for $k = 12$ on the best CatBoost fusion model.

pMOS for 2,280 distorted images. pMOS-SSIM+ attains a PLCC = 0.9264 and SRCC = 0.9174, substantially outperforming every single FR-IQA metric.

Table 4.8: TID2013 FR fusion versus individual FR-IQA metrics on the MOS train set.

Metric	**PLCC**	**SRCC**	**MSE**	**MAE**
pMOS-SSIM+ (ours)	**0.9264**	**0.9174**	**39.69**	**4.863**
MS-SSIM	0.7640	0.7550	2564	48.76
CW-SSIM	0.7488	0.7376	2565	48.78
SPIQ	0.7418	0.7580	2569	48.83
IW-SSIM	0.6808	0.7119	2568	48.82
PSNR	0.6776	0.6619	651	23.33
PSM	0.6482	0.6483	2455	47.95
ESIM	0.6482	0.6483	2562	48.74
G-SSIM	0.6400	0.6528	2584	49.01
SCC	0.6379	0.5999	2562	48.74
L-SSIM	0.6366	0.6387	2573	48.88
SNR	0.6351	0.6139	1002	29.74
W-SSIM	0.6268	0.6453	2582	48.98

TID2013 is fully annotated with MOS, this allows us to directly study the correlation between pMOS-SSIM+ and ground-truth MOS, thus validating the fusion stage independently of the NR learner. On the pMOS set (n = 2,280), pMOS-SSIM+ aligned strongly with human MOS, obtaining a PLCC of 0.8926, an SRCC of 0.8827, an MSE of 39.69, and an MAE of 4.863. Figure 4.9 confirms a strong alignment and shows that the fusion model generalizes well to unseen reference images and distortion types.

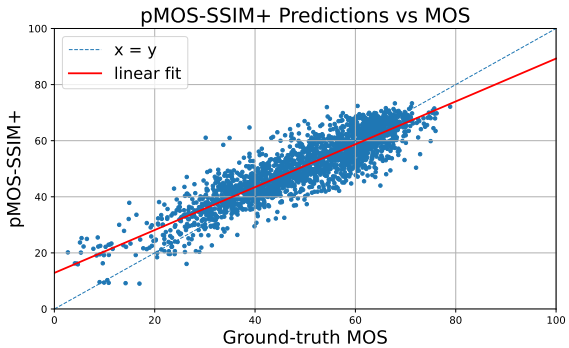


Figure 4.9: Scatter plot of pMOS-SSIM+ predictions versus human MOS on the pMOS set. The red line is the linear fit.

This is a strong indication that human perception can be effectively modeled by combining multiple FR metrics, and that the fusion model can produce reliable pseudo-labels, and eventually, train NR models, even when the reference images are very different from each other in content and are affected by various types of distortions.

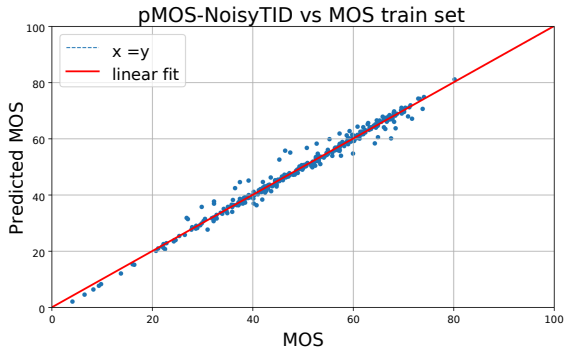
4.2.4 No-Reference Model

This time, the differences in performance were clearly evident. Table 4.9 shows that ResNet-50 was the best suited to be our NR-IQA model backbone. The resulting metric, pseudo-Mean Opinion Score for Noisy TID2013 Images (pMOS-NoisyTID), follows the same nomenclature as pMOS-StegaFace, indicating a pMOS regressor trained on Noisy distortions from the TID2013 dataset.

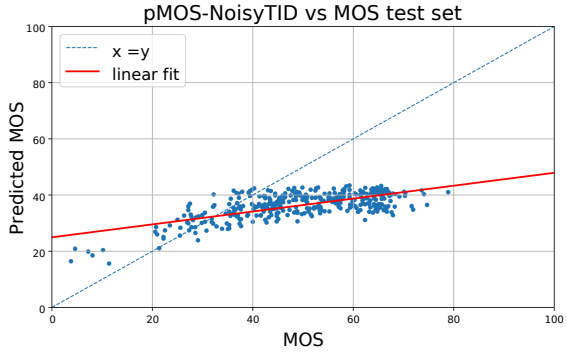
Table 4.9: Backbone comparison for training pMOS-NoisyTID. Report 5-fold mean PLCC, SRCC, MSE, and MAE.

Backbone	**PLCC**	**SRCC**	**MAE**	**MSE**
ResNet-50	**0.9484**	**0.9430**	$\mathbf{1.55 \times 10^{-3}}$	**0.0295**
EfficientNet-B1	0.7388	0.6870	6.7×10^{-3}	0.0645
ViT-B/16	0.8556	0.8296	3.94×10^{-4}	0.0454

Figure 4.10 shows two scenarios. On the MOS train set with ground-truth MOS the model fits almost perfectly, achieving a PLCC = 0.9924, SRCC = 0.9911, MSE = 3.03×10^{-4} and MAE = 0.0257. On the disjoint test set, however, correlations drop sharply PLCC = 0.6860, SRCC = 0.5591, MSE = 0.0305, MAE = 0.1460. This is a classic sign of overfitting, the model memorized the training set but failed to generalize to unseen images and distortion types.



(a) Predicted MOS vs ground-truth MOS from the train set.



(b) Predicted MOS vs ground-truth MOS from the disjoint test set.

Figure 4.10: pMOS-NoisyTID predictions on train and test sets.

We further evaluated the pMOS-NoisyTID metric on individual groups of reference images from the test set. Results varied: for Figure 4.11a, correlations remained relatively strong (PLCC = 0.836, SRCC = 0.796), for Figure 4.11b, performance declined (PLCC = 0.731, SRCC = 0.609), and for Figure 4.11c, correlations dropped further (PLCC = 0.683, SRCC = 0.557). These differences highlight the sensitivity of the model to content and distortion type.



(a) (b) (c)

Figure 4.11: TID2013 reference images from the test set.

Table 4.10 compares pMOS-NoisyTID predictions to standard NR-IQA metrics, showing clear improvements. There's a clear gap between standard NR-IQA baselines and the proposed pMOS-NoisyTID. Traditional blind metrics such as NIQE, PIQE and BRISQUE exhibit weak correlation with human judgements on TID2013, whereas pMOS-NoisyTID aligns substantially better with subjective quality, achieving PLCC = 0.6860, and SRCC = 0.5591. These results support our two-stage approach, pMOS fusion followed by NR model training, as a more reliable and generalizable route to NR-IQA, capturing perceptual quality variations in general images beyond the facial domain of FRLL.

Table 4.10: pMOS-NoisyTID versus NR-IQA baselines.

Metric	**PLCC**	**SRCC**	**MSE**	**MAE**
pMOS-NoisyTID (ours)	**0.6860**	**0.5591**	**0.0305**	**0.1460**
NIQE	-0.3096	-0.2262	2599	49.02
PIQE	-0.4102	-0.3643	2224	43.88
BRISQUE	-0.5055	-0.4444	1829	37.34

5 Conclusion and Future Work

This dissertation introduced a data-efficient pipeline for building new image quality metrics that scales from a small set of human-rated distortions to large unlabeled image sets, and that can either stop at a full-reference fusion metric or continue to a no-reference predictor, each optimized for the target application.

The pipeline begins with a compact but carefully curated subjective study. Human scores are screened and debiased to minimize observer and content effects. From there, a diverse set of full-reference measures is harmonized and fused into a single perceptual target. This stage yielded pMOS-SSIM+, a fused metric that aligns with human opinion more closely than any individual method and that can be adopted immediately wherever references are available.

Crucially, the same fused metric enables expansion beyond the limited human-rated subset. By assigning pseudo-labels to large collections without references, the pipeline supervises the learning of no-reference models that inherit the perceptual fidelity of the fusion while gaining real-world reach. Two outcomes illustrate the range of this second stage: pMOS-StegaFace, specialized for steganographically affected facial images, and pMOS-NoisyTID, designed to generalize across diverse content and distortion types. Both surpass established no-reference baselines, detect subtle degradations that conventional approaches often miss, and maintain strong agreement with human judgments. When needed, fairness safeguards are integrated at the data and model levels so that quality estimates remain consistent across demographic groups.

Future work follows four paths. First, extend the set of human-rated distortions to cover additional subtle or task-critical mechanisms, and study how many ratings are needed for the fusion to transfer reliably. Second, strengthen generalization of the no-reference stage with improved training and data augmentation, targeting challenging capture conditions and authentic mixtures of artifacts. Third, integrate calibrated uncertainty and active sampling so that the system can request new human ratings only where they add the most value, further reducing annotation cost. Fourth, complete end-to-end deployment studies in biometric workflows, including threshold setting, early-warning triggers for tampering, and continuous monitoring for fairness across demographic groups.

By unifying a compact subjective core, a robust fusion target, and scalable no-reference learning, this work delivers three concrete metrics: pMOS-SSIM+, pMOS-StegaFace, and pMOS-NoisyTID, and more importantly, a general pipeline to build the next ones. This foundation enables more reliable, fair, and security-aware image quality control in practice, raising the standard for how future systems screen and maintain image quality in alignment with human perception.

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[21] Federal Office for Information Security (BSI), "Open source face image quality (ofiq) – short report," BSI, Tech. Rep. Version 1.0, 2024. [Online]. Available: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/OFIQ/OFIQ_1_0_short_report.pdf.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[42] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE transactions on image processing*, vol. 25, no. 1, pp. 372–387, 2015.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[60] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, Springer, 2016, pp. 87–102.

[61] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[62] P. J. Phillips et al., "Overview of the face recognition grand challenge," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 947–954.

[63] B. Maze et al., "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 international conference on biometrics (ICB)*, IEEE, 2018, pp. 158–165.

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

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

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

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

[68] 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.

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

[70] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2022. DOI: 10.1109/TPAMI.2020.3045810.

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

[72] 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.

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

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

[75] 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.

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

[77] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011. DOI: 10.1109/TIP.2011.2109730.

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

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

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

[81] 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.

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

[83] 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.

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

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

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

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

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

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

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

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

[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 Transactions on image processing*, vol. 27, no. 1, pp. 206–219, 2017.

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

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

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

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

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

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

[99] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

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

[101] 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, 2012. DOI: 10.1109/TIP.2012.2214050.

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

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

[104] K. Pearson, "Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, no. 187, pp. 253–318, 1896.

[105] C. Spearman, "The proof and measurement of association between two things.," 1961.

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

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

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

[109] A. M. Leger, T. Omachi, and G. K. Wallace, "Jpeg still picture compression algorithm," *Optical Engineering*, vol. 30, no. 7, pp. 947–954, 1991.

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

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

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

[113] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

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

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

[116] A. Cruz, G. Schardong, L. Schirmer, J. Marcos, F. Shadmand, and N. Gonçalves, "Riemstega: Covariance-based loss for print-proof transmission of data in images," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2025, pp. 7572–7581.

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

[118] Django Software Foundation, *Django*, version 2.2, May 5, 2019. [Online]. Available: `https://djangoproject.com`.

[119] V. Chernozhukov et al., *Double/debiased machine learning for treatment and structural parameters*, 2018.

[120] R. A. Fisher, "Statistical methods for research workers.," 1934.

[121] H. Hotelling et al., "The generalization of student's ratio," 1931.

[122] J. B. Conway, *Functions of one complex variable II*. Springer Science & Business Media, 2012, vol. 159.

[123] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and simple linear regression," *Radiology*, vol. 227, no. 3, pp. 617–628, 2003.

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

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

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

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

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

[129] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[130] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *International Conference on Analysis of Images, Social Networks and Texts*, Springer, 2019, pp. 39–50.

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

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

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

[134] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems*, vol. 24, 2011.

[135] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[136] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *Journal of Machine Learning Research*, vol. 18, no. 185, pp. 1–52, 2018.

[137] S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, 2017. arXiv: 1705.07874 [cs.AI]. [Online]. Available: https://arxiv.org/abs/1705.07874.

[138] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1711.05101.

[139] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

[140] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[141] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[142] 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.

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

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

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

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

Appendix A

Subjective Screening Procedures

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

A.1 Confidence Intervals

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

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

The 95% confidence interval is then given by:

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

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

A.2 Observer Screening

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

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

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

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

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

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

For each observer i, two counters are defined:

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

A.2.1 Outlier Detection

If the distribution is normal:

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

If the distribution is non-normal:

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

A.2.2 Rejection Rule

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

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

where L is the total number of presentations.

A.3 Summary

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

Appendix B

Full Reference Metrics

B.1 Error-based Metrics

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

The most basic error-based metric is the Mean Squared Error (MSE) [68], which quantifies the average squared difference between corresponding pixels in the reference and distorted images:

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

The Root Mean Squared Error (RMSE) [68] is a direct extension of the mean squared error, obtained by taking its square root:

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{B.2}$$

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

The Mean Absolute Error (MAE) [68] measures the average absolute difference between the reference and distorted image pixels:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{B.3}$$

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

The Peak Signal-to-Noise Ratio (PSNR) [68] expresses the fidelity between a reference and a distorted image using a logarithmic decibel (dB) scale. It is defined as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX^2}{\text{MSE}}\right) \tag{B.4}$$

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

The Peak Signal-to-Noise Ratio with blocking effect (PSNR-B) [81], [142] was introduced to address one of the main weaknesses of PSNR, its insensitivity to compression artifacts such as blocking. PSNR-B extends the original formula by adding a Blocking Effect Factor (BEF) to the denominator:

$$\text{PSNR-B} = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE} + \text{BEF}} \right) \tag{B.5}$$

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

The Signal-to-Noise Ratio (SNR) [68] quantifies the proportion of signal power relative to noise power between a reference image and its distorted counterpart. It is computed as:

$$\text{SNR} = 10 \cdot \log_{10} \left(\frac{\sum x_i^2}{\sum (x_i - y_i)^2} \right) \tag{B.6}$$

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

B.2 Spectral-based Metrics

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

The Spectral Angle Mapper (SAM) [79] is one of the earliest spectral similarity measures. It computes the angular difference between the spectral vectors of the reference and distorted images, treating each pixel as a vector in the spectral domain:

$$\text{SAM}(x, y) = \cos^{-1} \left(\frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \right). \tag{B.7}$$

By focusing on the spectral angle rather than magnitude, SAM is insensitive to brightness variations and primarily reflects spectral shape consistency, making it a standard in hyperspectral image analysis.

The Relative Global Dimensional Error of Synthesis (ERGAS) [80], [143] measures the average relative root mean squared error across spectral bands:

$$\text{ERGAS} = 100 \cdot \frac{h}{l} \cdot \sqrt{\frac{1}{B} \sum_{i=1}^{B} \left(\frac{\text{RMSE}_i}{\mu_i} \right)^2}. \tag{B.8}$$

It accounts for differences in spatial resolution between reference and distorted images through the $\frac{h}{l}$ factor. Lower values indicate better spectral fidelity, and it is widely adopted in image fusion and pansharpening tasks to evaluate global relative errors.

The D_λ metric [80] evaluates spectral distortion by averaging inter-band distances:

$$D_\lambda = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{\substack{j=1 \\ j \neq i}}^{B} d(r_i, r_j). \tag{B.9}$$

Here, $d(r_i, r_j)$ is a distance between spectral bands i and j. D_λ emphasizes how well the relationships between spectral bands are preserved, providing a complementary perspective to per-band error metrics like ERGAS. Together, these measures capture both absolute spectral distortions and relative inter-band consistency.

B.3 Similarity-based Metrics

Similarity-based metrics assess perceptual image quality by measuring structural or feature-based resemblance between the reference and distorted images. Rather than relying on pixel-wise errors, they capture patterns of spatial organization, texture, luminance, or phase alignment. These methods often align more closely with the HVS than purely error-based metrics [68], [73].

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

The Structural Similarity Index (SSIM) [68] is a perceptual metric designed to evaluate image quality by comparing structural information between a reference and a distorted image. Unlike traditional error-based metrics, SSIM models luminance, contrast, and structure in a local window. Given two image patches x and y, SSIM is computed as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{B.10}$$

SSIM values range from 0 to 1, where 1 indicates perfect similarity.

The Multiscale Structural Similarity Index (MS-SSIM) [82] extends SSIM by computing the structural similarity at multiple scales, capturing both fine and coarse details. This is achieved by iteratively downsampling the input images and evaluating luminance, contrast, and structure

components at each level. The final score is a weighted geometric mean of the similarity scores across scales:

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \tag{B.11}$$

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

The Structural Dissimilarity Index (DSSIM) [68] is a reformulation of SSIM that expresses dissimilarity rather than similarity. It is derived directly from the SSIM value by:

$$\text{DSSIM}(x, y) = \frac{1 - \text{SSIM}(x, y)}{2} \tag{B.12}$$

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

The Information-Weighted Structural Similarity (IW-SSIM) [83] modifies MS-SSIM by incorporating local information content as a weighting mechanism. The intuition is that regions of an image with higher information density (e.g., textures or edges) are more perceptually important and should contribute more to the quality score. IW-SSIM applies a MS-SSIM computation, weighted by the local mutual information between reference and distorted images:

$$\text{IW-SSIM}(x, y) = \sum_i w_i \cdot \text{SSIM}_i(x, y) \tag{B.13}$$

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

The Complex Wavelet Structural Similarity (CW-SSIM) [84] adapts the SSIM formulation to the complex wavelet domain, which is more robust to small geometric distortions such as translations and rotations. Instead of comparing pixel intensities directly, CW-SSIM compares the phase and magnitude of complex wavelet coefficients between reference and distorted images:

$$\text{CW-SSIM}(x, y) = \frac{2\sum_i |c_{x,i}||c_{y,i}| + K}{\sum_i |c_{x,i}|^2 + \sum_i |c_{y,i}|^2 + K} \cdot \frac{2\left|\sum_i c_{x,i} c_{y,i}^*\right| + K}{\sum_i |c_{x,i} c_{y,i}^*| + K} \tag{B.14}$$

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

The Gradient-based Structural Similarity (G-SSIM) [85] extends SSIM by incorporating image gradient information into the structural comparison. The motivation is that gradients

reflect edge strength and direction, which are more sensitive to visual structure than raw intensity values. G-SSIM modifies the structure term in the SSIM formula to include gradient magnitudes:

$$\text{G-SSIM}(x, y) = l(x, y)^{\alpha} \cdot c(x, y)^{\beta} \cdot s(\nabla x, \nabla y)^{\gamma} \tag{B.15}$$

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

The Regional Structural Similarity (R-SSIM) [86] focuses on evaluating image quality within a central region of interest. This approach assumes that viewers primarily attend to central areas of an image, and thus, distortions in these regions are more perceptually significant:

$$\text{R-SSIM}(x, y) = \text{SSIM}(x|_{\Omega_c}, y|_{\Omega_c}) \tag{B.16}$$

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

The Luminance Structural Similarity (L-SSIM) [68] extends SSIM by applying it over multiple small patches across the image and averaging the results. This local analysis allows for finer sensitivity to spatially localized distortions that may be overlooked by global quality metrics. Given a window ω_k centered at pixel k, L-SSIM is computed as:

$$\text{L-SSIM}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \text{SSIM}(x|_{\omega_k}, y|_{\omega_k}) \tag{B.17}$$

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

The Color Structural Similarity (C-SSIM) [87] extends the traditional SSIM formulation by incorporating chromatic information alongside luminance, contrast, and structure. It proposes a four-component model that quantifies image distortion through a combination of local comparison functions, including a color fidelity term derived from the Spatial CIELAB (S-CIELAB) color space [144].

$$\text{C-SSIM}(x, y) = l(x, y) \cdot C(x, y) \cdot S(x, y) \cdot \text{Cr}(x, y) \tag{B.18}$$

$$\text{Cr}(x, y) = 1 - \frac{1}{k} \Delta E(x, y) \tag{B.19}$$

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

The Wavelet Structural Similarity (W-SSIM) [88], [145] is a region-pooled extension of SSIM that emphasizes perceptually important areas of an image. Instead of computing a uniform average of local SSIM values, W-SSIM assigns a larger weight to a Region of Interest (ROI), such as the image center, and a smaller weight to the surrounding regions. The final score is given by:

$$\text{W-SSIM} = w \cdot \frac{\sum_x M(x)\, S(x)}{\sum_x M(x)} + (1 - w) \cdot \frac{\sum_x (1 - M(x))\, S(x)}{\sum_x (1 - M(x))} \tag{B.20}$$

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

The Gradient Magnitude Similarity (GMS) [73] quantifies local similarity of gradient magnitudes, which better reflect perceptual structure than raw intensities. Let m_x and m_y be gradient-magnitude maps of reference x and distorted y (e.g., from Sobel or Scharr). Define the per-pixel similarity

$$\text{GMS}_i = \frac{2\, m_x(i)\, m_y(i) + C}{m_x(i)^2 + m_y(i)^2 + C}, \tag{B.21}$$

with small $C > 0$. The image-level score is the spatial average

$$\text{GMS}(x, y) = \frac{1}{N} \sum_{i=1}^{N} \text{GMS}_i, \tag{B.22}$$

ranging in $[0, 1]$ where higher is better. GMS is sensitive to blur and edge losses while being less affected by uniform luminance changes.

The Gradient Magnitude Similarity Deviation (GMSD) [73] uses the same gradient-magnitude similarity map as GMS but measures the spatial variation of quality rather than its mean. After computing GMS_i and its mean $\overline{\text{GMS}}$, the final score is the standard deviation

$$\text{GMSD}(x, y) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(\text{GMS}_i - \overline{\text{GMS}}\right)^2}. \tag{B.23}$$

Lower values indicate better quality. GMSD emphasizes localized degradations by penalizing spatial nonuniformity in the similarity map.

The Visual Signal-to-Noise Ratio (VSNR) [72] is a perceptual metric that combines subthreshold detection and suprathreshold distortion modeling in the wavelet domain. Let $d = y - x$ be the distortion. After applying a contrast sensitivity function and contrast masking to the wavelet coefficients of d to obtain a perceptual distortion $\tilde{d}$, VSNR is defined as

$$\text{VSNR}(x, y) = 20 \log_{10} \left(\frac{\|x\|_2}{\|\tilde{d}\|_2} \right) \text{ dB}. \tag{B.24}$$

Larger values (in dB) indicate better quality. If the distortion is below detection thresholds, VSNR becomes very large, reflecting near-invisibility, otherwise it decreases as perceptual distortion grows.

The Feature Similarity Index (FSIM) [77] assesses image quality by leveraging perceptually meaningful features. It uses Phase Congruency (PC) as the primary feature, which reflects the significance of local structure independent of luminance and contrast, and Gradient Magnitude (GM) as a secondary feature to weight salient regions. The local similarity map $S_{\text{FSIM}}(x)$ is defined as:

$$S_{\text{FSIM}}(x) = \frac{S_{\text{PC}}(x) \cdot S_{\text{GM}}(x)}{\max(PC_{\text{ref}}(x), PC_{\text{dist}}(x)) + \epsilon} \tag{B.25}$$

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

A multi-scale variant, Multiscale Feature Similarity Index (MS-FSIM) [77], extends FSIM by evaluating feature similarity across a pyramid of S resolutions. At each scale s, phase congruency and gradient magnitude maps are computed, a local similarity $S_{\text{FSIM},s}(x)$ is formed as in FSIM, and an image-level score FSIM_s is obtained via PC-weighted averaging. The final MS-FSIM combines the per-scale scores with weights $\{w_s\}_{s=1}^{S}$:

$$\text{MS-FSIM} \;=\; \prod_{s=1}^{S} \left(\text{FSIM}_s\right)^{w_s} \quad \text{with} \quad \sum_{s=1}^{S} w_s = 1\,. \tag{B.26}$$

Pooling across scales improves sensitivity to distortions that manifest at different spatial frequencies, such as blur, ringing, and compression artifacts, while preserving FSIM's emphasis on perceptually salient structure.

The Feature Similarity Index with color (FSIMc) [77] is an extension of FSIM that integrates color information to better evaluate the perceptual quality of color images. In addition to the PC and GM features used in FSIM, FSIMc includes chrominance similarity components derived from the I and Q channels of the YIQ color space [110]. The final FSIMc score is given by:

$$\text{FSIM}_c = \frac{\sum_{x\in\Omega} S_{\text{FSIM}}(x) \cdot S_I(x) \cdot S_Q(x) \cdot \text{PC}_{\max}(x)}{\sum_{x\in\Omega} \text{PC}_{\max}(x)} \tag{B.27}$$

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

The Spectral Residual Similarity (SR-SIM) [89] leverages visual saliency derived from the spectral residual of the Fourier transform. The spectral residual highlights regions that are

more likely to attract human attention, and SR-SIM incorporates this information to weight structural similarity computations. The final score is defined as:

$$\text{SR-SIM} = \frac{\sum_{x\in\Omega} S(x) \cdot \text{Sal}_{\min}(x)}{\sum_{x\in\Omega} \text{Sal}_{\min}(x)}, \tag{B.28}$$

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

The Edge Strength Similarity (ESIM) [90] improves SSIM by emphasizing structural fidelity in edge regions, which are perceptually more sensitive to distortions. ESIM first extracts edge maps from the reference and distorted images using edge detectors such as Sobel or Canny, and then computes local SSIM values restricted to these edge pixels. The final score is defined as:

$$\text{ESIM} = \frac{1}{N} \sum_{x\in\Omega_{\text{edge}}} S(x), \tag{B.29}$$

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

The Universal Image Quality Index (UQI) [78] models image distortion as a combination of three factors: loss of correlation, luminance distortion, and contrast distortion. UQI is defined as:

$$\text{UQI} = \frac{4\sigma_{xy}\mu_x\mu_y}{\left(\sigma_x^2 + \sigma_y^2\right)\left(\mu_x^2 + \mu_y^2\right)}, \tag{B.30}$$

where μ_x and μ_y are the mean intensities of the reference and distorted images, σ_x^2 and σ_y^2 are their variances, and σ_{xy} is the covariance between them. The index ranges from -1 to 1, with higher values indicating greater similarity.

The Spatial Correlation Coefficient (SCC) [91] measures the linear correlation between the reference and distorted images. It is defined as:

$$\text{SCC} = \frac{\sum_{x\in\Omega}(I_r(x) - \mu_r)(I_d(x) - \mu_d)}{\sqrt{\sum_{x\in\Omega}(I_r(x) - \mu_r)^2} \cdot \sqrt{\sum_{x\in\Omega}(I_d(x) - \mu_d)^2}}, \tag{B.31}$$

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

B.4 Information-theoretic Metrics

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

The Information Fidelity Criterion (IFC) [71] quantifies the mutual information shared between the reference and distorted coefficients under the GSM model:

$$\text{Information Fidelity Criterion (IFC)} = \sum_k I(C_k; D_k \mid s_k)\,. \tag{B.32}$$

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

The Visual Information Fidelity (VIF) index [75] extends this framework by incorporating perceptual noise to account for the HVS. It is defined as the ratio between the mutual information preserved after distortion and the information available in the reference:

$$\text{VIF} = \frac{\sum_k I(C_k; D_k \mid s_k)}{\sum_k I(C_k; C_k + n_k \mid s_k)}. \tag{B.33}$$

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

Visual Information Fidelity (pixel domain) (VIFp) [75] is a simplified pixel-domain version that avoids multiscale wavelet decomposition. Although less precise, it significantly reduces computational complexity and is useful in real-time applications.

Visual Information Fidelity (color images) (VIFc) [75] extends the metric to color images by transforming them into the YUV color space. VIF is computed independently on each channel, and the final score is given by:

$$\text{VIFc} = \tfrac{1}{3}\left(\text{VIF}_Y + \text{VIF}_U + \text{VIF}_V\right). \tag{B.34}$$

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

B.5 Learning-based Metrics

Notation. Let $\phi_l(x)$ denote the activation (feature map) of image x at layer l of a pretrained Convolutional Neural Network (CNN) with spatial size $H_l \times W_l$ and C_l channels. Features are

channel-wise normalized, denoted by $\hat{\phi}_l(x)$. Channel weights are written as w_l, and similarity functions are applied element-wise over spatial positions (h, w).

Learned Perceptual Image Patch Similarity (LPIPS) [76] measures perceptual similarity by comparing deep features extracted from pretrained networks such as AlexNet, SqueezeNet, or Visual Geometry Group (VGG). The metric is defined as:

$$\mathrm{LPIPS}(x, y) = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h,w} \left\| \hat{\phi}_l(x)_{h,w} - \hat{\phi}_l(y)_{h,w} \right\|_2^2. \tag{B.35}$$

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

Deep Image Structure and Texture Similarity (DISTS) [70] combines structure and texture similarity in the deep feature space. For each layer l, structure similarity S_l is computed from normalized feature statistics, and texture similarity T_l is obtained from feature covariances. The final score is a convex combination:

$$\mathrm{DISTS}(x, y) = \sum_l \alpha_l S_l(x, y) + \beta_l T_l(x, y), \tag{B.36}$$

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

Weighted Average Deep Image Quality Assessment Model (WaDIQaM) [92] employs deep convolutional networks to learn a patch-based representation of image quality. In the full-reference setting, reference and distorted patches are passed through a siamese network [146], and local quality estimates are aggregated using a learned weighting mechanism:

$$\mathrm{WaDIQaM}(x, y) = \sum_p w_p \, q_p(x, y), \tag{B.37}$$

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

Perceptual Similarity Metric (PSM) [93] builds on LPIPS to make full-reference perceptual similarity robust to small, imperceptible misalignments. It replaces shift-sensitive elements in the feature extractor with anti-aliased strided convolutions, anti-aliased pooling, reduced strides, and appropriate padding, then aggregates layerwise feature distances like LPIPS. Let $F_l(x)$ be shift-tolerant features at layer l and $\hat{F}_l$ their channel-normalized maps, the score is

$$\mathrm{PSM}(x, y) = \sum_l w_l \cdot \frac{1}{H_l W_l} \sum_{h,w} |\hat{F}_l(x)_{h,w} - \hat{F}_l(y)_{h,w}|_2^2, \tag{B.38}$$

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

Semantic Perceptual Image Quality (SPIQ) [94] is an approach that first learns a quality-aware visual representation via self-supervised pretraining, then fits a lightweight regressor on top with limited MOS labels. Two augmented views of the same image are encouraged to produce consistent features during pretraining. At inference a regressor maps features to a scalar quality:

$$\mathrm{SPIQ}(x) = g(f(x)), \tag{B.39}$$

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

Appendix C

Regression Models for Metric Fusion

Fusion-based IQA learns a regression from heterogeneous metric scores to a single perceptual estimate. Let the dataset be $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with $\mathbf{x}_i \in \mathbb{R}^p$ the vector of p objective scores and $y_i \in \mathbb{R}$ the MOS. We first harmonize polarity so that larger metric values indicate better quality, optionally rescale each metric to zero mean and unit variance, and map targets to a fixed range (for example 0–100). We then estimate a regressor $f : \mathbb{R}^p \to \mathbb{R}$ by empirical risk minimization

$$\min_{\theta} \; \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, f(\mathbf{x}_i; \theta)\big) + \lambda \, \Omega(\theta),$$

using a squared or robust loss ℓ (for example Huber), and regularizer Ω if applicable. To handle redundancy, we either select the top k metrics by correlation with y or project $\mathbf{x}_i$ to a lower-dimensional subspace learned from the score matrix. The learned fusion f aggregates complementary cues and typically aligns with human judgments better than any single metric.

There are three useful categories for fusion on tabular metric vectors. Linear models assume an approximately additive mapping and control complexity via shrinkage or sparsity [123], [124], [125], [126]. Bootstrap aggregating and gradient boosting ensembles model interactions and are robust to heterogeneous scales [127], [128], [129], [130]. Modern boosted-tree frameworks optimize second order objectives with explicit regularization and efficient growth strategies [131], [132], [133].

C.1 Linear Models

Linear Regression [123] fits an affine predictor by minimizing squared residuals:

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

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

Ridge Regression [124] stabilizes linear fusion with L_2 shrinkage

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

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

Bayesian Ridge [125] is a probabilistic ridge with Gaussian prior and noise

$$\mathbf{w} \mid \alpha \sim \mathcal{N}(\mathbf{0},\, \alpha^{-1}\mathbf{I}), \quad y_i \mid \mathbf{x}_i, \mathbf{w}, \beta \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i + b,\, \beta^{-1}), \tag{C.3}$$

$$\mathbf{S} = \Big(\alpha \mathbf{I} + \beta \mathbf{X}^\top \mathbf{X}\Big)^{-1}, \quad \mathbf{m} = \beta\, \mathbf{S}\, \mathbf{X}^\top \mathbf{y}. \tag{C.4}$$

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

ElasticNet [126] promotes sparse yet stable linear fusion with mixed penalties

$$\min_{\mathbf{w},\,b} \sum_{i=1}^{n} \Big(y_i - \mathbf{w}^\top \mathbf{x}_i - b\Big)^2 + \alpha \left(\rho \|\mathbf{w}\|_1 + \tfrac{1-\rho}{2} \|\mathbf{w}\|_2^2\right), \tag{C.5}$$

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

C.2 Ensemble Methods

Random Forest [127] employs bagging regression trees to reduce variance and requires minimal preprocessing. Nodes choose splits that maximize variance reduction

$$\Delta = \mathrm{Var}(S) - \frac{|S_L|}{|S|} \mathrm{Var}(S_L) - \frac{|S_R|}{|S|} \mathrm{Var}(S_R), \tag{C.6}$$

and the fusion prediction averages T trees

$$\hat{y}(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}). \tag{C.7}$$

Extremely randomized trees [128] sample split thresholds at random for each feature before selecting the best split, which increases diversity and often lowers variance. Aggregation is identical to Random Forest, with similar variance reduction analysis.

Gradient Boosting [129] employs additive modeling to fit weak learners to pseudo residuals. For squared error, residuals are $r_{im} = y_i - f_{m-1}(\mathbf{x}_i)$. Each stage fits a tree b_m to $\{(\mathbf{x}_i, r_{im})\}$ and updates

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu\, b_m(\mathbf{x}), \tag{C.8}$$

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

Histogram Gradient Boosting [130] is a histogram variant of gradient boosting that bins each feature into B buckets and accumulates gradient and Hessian statistics per bin to evaluate splits efficiently. It approximates standard boosting while improving speed and memory on many-metric settings.

C.3 Boosted Trees

Extreme Gradient Boosting (XGBoost) [131] optimizes a second-order regularized objective with per-leaf closed-form weights:

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t), \qquad \Omega(f) = \gamma T + \tfrac{\lambda}{2} \sum_{j=1}^{L} w_j^2. \tag{C.9}$$

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

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \qquad \text{Gain} = \tfrac{1}{2} \left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} \right) - \gamma, \tag{C.10}$$

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

Light Gradient Boosting Machine (LightGBM) [132] builds trees leaf-wise (best-first) under depth or leaf constraints and uses gradient-based one-side sampling and exclusive feature bundling, selecting each split by maximizing a second-order regularized gain:

$$\text{Gain} = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda} - \gamma, \tag{C.11}$$

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

Categorical Boosting (CatBoost) [133] performs ordered boosting with symmetric trees to mitigate prediction shift by computing gradients on permutation prefixes and updating leaves with second-order estimates while optimizing a standard regression loss:

$$f_t(\mathbf{x}) = f_{t-1}(\mathbf{x}) + \nu \, b_t(\mathbf{x}), \tag{C.12}$$

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \tag{C.13}$$

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

Appendix D

Hardware and Software Environment

D.1 Hardware

CPU: AMD Ryzen 5 7600X, GPU: NVIDIA GeForce RTX 5060 Ti (16 GB), Host OS and kernel: Ubuntu 22.04.5 LTS (Jammy), Linux 6.8.0-79-generic

D.2 Software

Container and drivers

Docker: 28.4.0, Container image: `nvcr.io/nvidia/pytorch:25.06-py3`, OS (in-container): Ubuntu 24.04.2 LTS, Python: 3.12.3, CUDA runtime (in-container): 12.9, cuDNN: 9.1.2

Core Python stack

NumPy: 1.26.4, SciPy: 1.15.3, pandas: 2.2.3

Deep learning

PyTorch: 2.8.0a0 (NGC nv25.06), TorchVision: 0.22.0a0, timm: 1.0.19

IQA and image processing

PIQ: 0.8.0, sewar: 0.4.6, LPIPS: 0.1.4, scikit-image: 0.25.2, OpenCV (headless): 4.11.0.86, Pillow: 11.2.1

Fusion and explainability

scikit-learn: 1.6.1, CatBoost: 1.2.8, LightGBM: 4.6.0, XGBoost: 2.1.4, SHAP: 0.48.0

Hyperparameter search and logging

Optuna: 4.5.0, TensorBoard: 2.16.2

Application services for subjective tests

Django: 4.2.10, PostgreSQL: 15.14, psycopg2: 2.9.7, pgAdmin4: 6.8