Image Quality Assessment: Addressing the Data Shortage and Multi-Stage Distortion Challenges

by

Shahrukh Athar

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Simon Xianyi Yang
Professor, School of Engineering,
University of Guelph

Supervisor(s): Zhou Wang
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal Member: Oleg Michailovich
Associate Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal Member: Weihua Zhuang
Professor, Dept. of Electrical and Computer Engineering,
University of Waterloo

Internal-External Member: Alexander Wong
Associate Professor, Dept. of Systems Design Engineering,
University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Visual content constitutes the vast majority of the ever increasing global Internet traffic, thus highlighting the central role that it plays in our daily lives. The perceived quality of such content can be degraded due to a number of distortions that it may undergo during the processes of acquisition, storage, transmission under bandwidth constraints, and display. Since the subjective evaluation of such large volumes of visual content is impossible, the development of perceptually well-aligned and practically applicable objective image quality assessment (IQA) methods has taken on crucial importance to ensure the delivery of an adequate quality of experience to the end user. Substantial strides have been made in the last two decades in designing perceptual quality methods and three major paradigms are now well-established in IQA research, these being Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR), which require complete, partial, and no access to the pristine reference content, respectively. Notwithstanding the progress made so far, significant challenges are restricting the development of practically applicable IQA methods. In this dissertation we aim to address two major challenges: 1) The data shortage challenge, and 2) The multi-stage distortion challenge.

NR or blind IQA (BIQA) methods usually rely on machine learning methods, such as deep neural networks (DNNs), to learn a quality model by training on subject-rated IQA databases. Due to constraints of subjective-testing, such annotated datasets are quite small-scale, containing at best a few thousands of images. This is in sharp contrast to the area of visual recognition where tens of millions of annotated images are available. Such a data challenge has become a major hurdle on the breakthrough of DNN-based IQA approaches. We address the data challenge by developing the largest IQA dataset, called the Waterloo Exploration-II database, which consists of 3,570 pristine and around 3.45 million distorted images which are generated by using content adaptive distortion parameters and consist of both singly and multiply distorted content. As a prerequisite requirement of developing an alternative annotation mechanism, we conduct the largest performance evaluation survey in the IQA area to-date to ascertain the top performing FR and fused FR methods. Based on the findings of this survey, we develop a technique called Synthetic Quality Benchmark (SQB), to automatically assign highly perceptual quality labels to large-scale IQA datasets. We train a DNN-based BIQA model, called EONSS, on the SQB-annotated Waterloo Exploration-II database. Extensive tests on a large collection of completely independent and subject-rated IQA datasets show that EONSS outperforms the very state-of-the-art in BIQA, both in terms of perceptual quality prediction performance and computation time, thereby demonstrating the efficacy of our approach to address the data challenge.

In practical media distribution systems, visual content undergoes a number of degradations as it is transmitted along the delivery chain, making it multiply distorted. Yet, research in IQA has mainly focused on the simplistic case of singly distorted content. In many practical systems, apart from the final multiply distorted content, access to earlier degraded versions of such content is available. However, the three major IQA paradigms (FR, RR, and, NR) are unable to take advantage of this additional information. To address this challenge, we make one of the first attempts to study the behavior of multiple simultaneous distortion combinations in a two-stage distortion pipeline. Next, we introduce a new major IQA paradigm, called degraded reference (DR) IQA, to evaluate the quality of multiply distorted images by also taking into consideration their respective degraded references. We construct two datasets for the purpose of DR IQA model development, and call them DR IQA database V1 and V2. These datasets are designed on the pattern of the Waterloo Exploration-II database and have 32,912 SQB-annotated distorted images, composed of both singly distorted degraded references and multiply distorted content. We develop distortion behavior based and SVR-based DR IQA models. Extensive testing on an independent set of IQA datasets, including three subject-rated datasets, demonstrates that by utilizing the additional information available in the form of degraded references, the DR IQA models perform significantly better than their BIQA counterparts, thereby establishing DR IQA as a new paradigm in IQA.

Acknowledgements

Words cannot express how grateful I am to my supervisor, Dr. Zhou Wang, for being such an exceptional, supportive, and patient mentor throughout my Ph.D. studies. Thank you so very much Dr. Wang, for taking me on as a Ph.D. student, for having continual faith in me, for supporting me through the many ups and downs that I went through over the past many years, and for giving me the flexibility to explore different research directions. If it was one thing that I could always count on during my Ph.D. studies, it was Dr. Wang's support, which gave me immense peace of mind. Through the countless meetings and discussions that we had, not only did Dr. Wang teach me how to do good research that is vital for the advancement of our field, but he also tremendously refined my writing and presentation skills to effectively communicate research findings. Even though Dr. Wang is one of the most accomplished professors at the University of Waterloo, he is a down-to-earth person who is fully invested in his students' success and does not impose his will on them. His flexible attitude allowed me to also pursue my passion of teaching besides research. It has been an honor and privilege working with you Dr. Wang.

I am honored to have Dr. Simon Xianyi Yang, Dr. Oleg Michailovich, Dr. Weihua Zhuang, and Dr. Alexander Wong as my thesis committee members, and thank them for the time they took to review my thesis and for giving valuable advice and suggestions to further improve my work.

I have had the opportunity to work alongside truly amazing graduate students and am grateful to all members of the Image and Vision Computing Lab, past and present, including Abdul Rehman, Hojatollah Yeganeh, Kai Zeng, Jiheng Wang, Qingbo Wu, Shiqi Wang, Rasoul Mohammadi Nasiri, Kede Ma, Wentao Liu, Zhengfang Duanmu, Zhuoran Li, Zhongling Wang, Xinyu Guo, and Jinghan Zhou. I want to especially thank my friend Abdul Rehman for introducing me to Dr. Wang and for his immense guidance during the early years of my Ph.D. studies; Kede Ma, Wentao Liu, and Zhengfang Duanmu for their phenomenal advice, help, and support; and Zhongling Wang for working so diligently on a key project with me.

I am eternally grateful to my parents for instilling love for knowledge and education in me. Though they left this world more than a decade ago, yet the void that they left can never be filled. They had wanted so much that I pursue a doctoral degree, and they would have been so happy to see me accomplish this milestone. *Ammi* and *Abbu*, I miss you a lot.

I am so very grateful to my dear mentor and friend, Dr. Zartash Afzal Uzmi, for being a source of constant support in the last ten years. Dr. Uzmi's confidence in my abilities was a big motivation for me to go for a doctoral degree after having given up on the prospect. Even though we are thousands of kilometers apart, whenever I felt down, I knew that you, and thus my motivation, were just a call away.

I want to thank my dear wife Shahida and my loving daughter Maryam, for persevering with me for the last many years. Words cannot express how much I appreciate your love, help, support, encouragement, and most of all your patience. My wife is a symbol of hard work and brilliance, for she not only stood by me during my Ph.D., but also completed her own Ph.D., which is something truly remarkable. I am so lucky to have both of you and hope that I can become the husband and father that you deserve!

While I have named just a few people above, I want to thank all the people who made this thesis possible, thank you!

Dedication

I dedicate this thesis to
my late parents, Mr. Athar Kemal and Mrs. Nighat Athar,
my wife, Dr. Shahida Jabeen,
and my daughter, Maryam Athar.

Table of Contents

List of Tables

List of Figures

List of Acronyms

2D two-dimensional. 15, 28, 31, 35, 47, 96, 103, 110, 195

2stepQA two-step quality assessment. 155–157, 160, 167–169, 205, 206, 216, 217

3D three-dimensional. 15

ACR absolute category rating. 27

AGGD Asymmetric Generalized Gaussian Distribution. 54, 58

AWGN additive white Gaussian noise. 114

B-JPG Blur-JPEG. 160, 164–168, 170, 214–216, 218, 220, 221, 228

B-N Blur-Noise. 160, 165–168, 170, 214–216, 218–220, 228

BIQA blind image quality assessment. 2, 3, 5–7, 9, 52, 96, 98–101, 104–109, 114, 117–119, 131–138, 140–142, 144, 146, 148–151, 223–227

BIQI Blind Image Quality Index. 53, 54, 85, 94, 136, 138

BIRD Biggest Index Ranking Difference. 50

BLISS Blind Learning of Image quality using Synthetic Scores. 51, 117, 119–122

BRISQUE Blind/Referenceless Image Spatial Quality Evaluator. 54, 58, 85, 94, 136, 138, 156

CD Context-Dependent. 49

CF colorfulness. 28, 29, 31, 110, 179–181, 195

CF Context-Free. 49, 50, 68, 69

CID Color-Image-Difference measure. 42, 67, 68, 74, 80, 121, 124, 130, 160

CISI Combined Image Similarity Index. 48, 63, 74, 82, 83

CNN convolutional neural network. 56, 102, 105–108, 118, 133

CNNM Combined Neural Network Metric. 50, 63, 74, 81–83, 96, 118

CORNIA Codebook Representation for No-reference Image quality Assessment. 55, 56, 59, 85, 93–96, 118, 136, 138, 142, 156, 160, 164–166, 168, 206, 209, 215, 217

CR corrupted-reference. 155

CSF contrast sensitivity function. 38, 46

D-test pristine/distorted image discriminability test. 25

DCT discrete cosine transform. 38, 42, 67

DIP quality-discriminable image pair. 59, 94, 95, 118

dipIQ DIP inferred quality. 59, 93–96, 118, 136, 142, 160, 164, 166, 168, 206, 209, 215, 217

DMOS Difference Mean Opinion Score. 17, 20–22, 33, 47, 62, 158, 160, 161, 164, 167, 210

DNN deep neural network. 3, 6, 7, 9, 56, 57, 96–102, 104–109, 131–134, 136, 139–141, 143–146, 148, 149, 194, 223, 224, 226–229

DR Degraded Reference. 5, 8–10, 121, 126, 150, 151, 157, 161, 164, 167, 169–172, 178, 179, 181, 188–210, 214–222, 225–230

DSCQS double stimulus continuous quality scale. 25

DSS DCT Subbands Similarity. 42, 67, 74, 80, 96, 118, 121, 124, 130, 160

DVICOM Detail Virtual Cognitive Model. 46, 47, 74, 80, 124

DWT discrete wavelet transform. 38

EONSS End-to-end Optimized deep neural Network using Synthetic Scores. 7, 9, 98, 99, 109, 131–146, 148, 149, 224, 226, 227

ESSIM Edge Strength Similarity. 41, 42, 74, 80, 124

FD Final Distorted. 169–172, 178, 179, 181, 188–194, 197, 199, 200, 206, 209, 214–216, 229

FR Full-Reference. 2–9, 11–17, 19, 27, 35, 39, 40, 42–52, 56, 59, 63, 66–70, 72, 74, 78, 80–84, 88, 92, 95–97, 99, 105, 106, 112, 117–121, 124, 126, 128–131, 136, 140, 141, 148–152, 154, 156–158, 160, 161, 163, 166–168, 170–172, 178, 191–194, 196, 200, 202, 206–210, 214, 215, 217, 219–222, 224, 225, 227

FSIM Feature Similarity. 40, 41, 48–50, 52, 59, 68, 74, 80, 96, 105, 117, 118, 124, 130, 156, 160, 161, 164, 168, 172, 178, 179, 181, 200, 202, 206, 209

GDN generalized divisive normalization. 57, 132, 133

GGD Generalized Gaussian Distribution. 53, 54, 58

GMSD Gradient Magnitude Similarity Deviation. 42, 52, 59, 68, 74, 80, 117, 118, 124

GSIM Gradient Similarity. 41

GSM Gaussian Scale Mixtures. 43, 44

GWHGLBP Gradient-Weighted Histogram of Local Binary Pattern calculated on the Gradient map. 54, 55, 85, 94, 136, 138, 155

HFSIMc Hybrid Feature Similarity. 48, 63, 74, 82, 83, 96, 119

HOSA High Order Statistics Aggregation. 56, 85, 93, 94, 96, 136, 138

HVS human visual system. 3, 38, 39, 42–46, 105

ICA independent component analysis. 44

IFC Information Fidelity Criterion. 43, 49, 50, 68, 82

ILNIQE Integrated Local Natural Image Quality Evaluator. 58, 95, 136, 142

ILSVRC ImageNet Large Scale Visual Recognition Challenge. 102

IQA Image quality assessment. 2–17, 19, 22, 25–29, 31, 33–35, 37, 39–47, 49, 50, 52, 53, 55–57, 59–63, 67, 70–72, 78, 80, 82–85, 92, 94–96, 98–101, 103, 105–112, 114–119, 121, 124–126, 128, 130, 131, 133–135, 139, 140, 148–157, 159, 161, 164, 166–171, 178, 191–210, 214–229

IWSSIM Information content Weighted Structural Similarity. 40, 52, 67, 68, 74, 80, 81, 88, 95, 96, 118, 121, 124, 130, 136, 140, 160, 227

JPG-JPG JPEG-JPEG. 160, 164, 166, 168, 170, 214, 216, 218, 220, 221, 228

KRCC Kendall Rank-order Correlation Coefficient. 14, 61

L-test listwise ranking consistency test. 25

LBP Local Binary Pattern. 55, 60, 155

LPSI Local Pattern Statistics Index. 60, 95, 136

MAD Most Apparent Distortion. 45, 46, 49, 50, 68, 74, 80, 124

MAE Mean Absolute Error. 14, 133

MCSD Multiscale Contrast Similarity Deviation. 42, 68, 80, 124

MEON Multi-task End-to-end Optimized deep neural Network. 57, 85, 93, 94, 96, 136, 138–141, 143–145, 149

MMF Multi-Method Fusion. 49, 50, 68, 69, 74, 81–83, 96, 118, 124, 128, 129

MOS Mean Opinion Score. 12, 13, 17, 20, 21, 23, 26, 27, 33, 62, 72, 158–161, 164, 167, 210

MSCN Mean Subtracted Contrast Normalized. 54, 58, 228

MSE Mean Squared Error. 2, 35, 37, 39, 133

MSSSIM Multi-scale Structural Similarity. 39, 40, 42, 48–50, 59, 68, 118, 156, 160, 161, 167, 168, 206, 209, 214, 216

MUSIQUE Multiply and Singly distorted Image Quality Estimator. 118, 155

MVG multivariate Gaussian. 58

N-JP2 Noise-JPEG2000. 160, 164, 166, 168, 170, 214, 216, 218, 220, 228

N-JPG Noise-JPEG. 160, 163, 164, 166, 168, 170, 214, 216, 218, 220, 221, 228

NBJ-JPG Noise/Blur/JPEG-JPEG. 160, 163, 164, 166, 168, 205, 214, 216, 218, 220, 221, 228

NIQE Natural Image Quality Evaluator. 57, 58, 95, 136, 156, 160, 164–167, 206, 209, 216, 217

NQM Noise Quality Measure. 38, 49, 82

NR No-Reference. 2, 3, 5, 6, 8, 9, 11–17, 27, 35, 47, 52–60, 84, 85, 88, 92–96, 99, 135, 148, 150–152, 154–157, 160, 164–170, 193, 194, 200, 206–209, 215–217, 219–222, 224, 225, 227

NRSL No-Reference quality assessment using statistical Structural and Luminance features. 55, 85, 94, 136, 138

NSS Natural Scene Statistics. 14, 43, 44, 53, 54, 58, 67, 74, 80, 108, 124, 228

OA Opinion-Aware. 52–56, 85, 94, 136, 138

OCR Optical Character Recognition. 27

OU Opinion-Unaware. 52, 57–60, 94, 95, 136

P-test pairwise preference consistency test. 25

PCA principal component analysis. 58

PDF probability density function. 154

PLCC Pearson Linear Correlation Coefficient. 14, 15, 49, 50, 60–63, 70–72, 74, 78, 80–85, 88, 93–95, 124–130, 133–137, 139, 143, 144, 146, 160, 161, 163, 164, 167, 210, 226

PQR probabilistic quality representation. 107, 156

PR Pristine Reference. 169–172, 178, 181, 188, 190–193, 197, 199, 200, 206, 214, 215, 219, 220

PSNR Peak Signal-to-Noise Ratio. 2, 34, 35, 37–39, 49, 74, 78, 88, 92, 95, 124, 128, 136, 141, 156

QA quality assessment. 2, 12, 99

QAC Quality Aware Clustering. 59, 95, 117, 136

QASD Quality index with Adaptive Sub-Dictionaries. 45, 74, 80, 124

QoE quality-of-experience. 112

RAS RRF based Adjusted Scores. 52, 63, 66–69, 74, 81–84, 96, 119–121, 124, 128, 129, 224

RBF radial basis function. 68, 209

ReLU Rectified Linear Unit. 56, 57, 133

RFSIM Riesz-transform based Feature Similarity. 40, 48

RMSE Root-Mean-Squared Error. 14, 15

RR Reduced-Reference. 2, 3, 5, 8, 9, 12, 14, 15, 17, 35, 43, 99, 148, 150–152, 154, 157, 170, 221, 225, 227, 228

RRF Reciprocal Rank Fusion. 6, 7, 51, 52, 119–123, 129, 131, 166, 224

RTC relative total coverage. 31

SDR standard dynamic range. 14, 19

SDSP Saliency Detection by combining Simple Priors. 41

SFF Sparse Feature Fidelity. 44, 45, 50, 67, 68, 74, 80, 124

SFMS Sequential Forward Method Selection. 49, 50, 68, 69

SI spatial information. 28, 29, 31, 110, 179–181, 195

SISBLIM Six-Step Blind Metric. 59, 60, 95, 136, 142, 155

SNR Signal-to-Noise Ratio. 37

SQB Synthetic Quality Benchmark. 7–9, 115, 116, 120–122, 124, 126–131, 134, 148, 151, 158, 160, 161, 164, 166, 167, 199, 200, 209, 210, 215, 217, 222, 224, 225, 229

SRCC Spearman Rank-order Correlation Coefficient. 14, 15, 47, 61, 63, 66, 70–72, 80–85, 88, 93–95, 122, 123, 125–129, 133–137, 139, 143, 144, 146, 160, 161, 163, 164, 167, 210, 226

SRSIM Spectral Residual based Similarity. 41, 50, 68

SRVS Spectral Residual based Visual Saliency. 41

SSIM Structural Similarity. 39, 40, 49, 118, 171, 172

SVM Support Vector Machine. 49, 54

SVR Support Vector Regression. 8, 10, 49, 53–56, 68, 100, 106, 107, 155, 194, 195, 199, 209, 210, 217–220, 222, 225, 228

TEF total effective coverage. 31

UQI Universal image Quality Index. 39

VIF Visual Information Fidelity. 43, 44, 48–50, 52, 59, 68, 74, 80, 118, 124

VQA video quality assessment. 229, 230

VSI Visual Saliency-based Index. 41, 67, 68, 74, 80, 96, 118, 124, 128–130, 156

VSNR Visual Signal-to-Noise Ratio. 38, 49, 82

WA weighted average. 125, 127, 133, 136, 137, 139, 143, 144, 146

WaDIQaM-NR Weighted Average Deep Image Quality Measure for NR IQA. 56, 57, 85, 92, 94, 96, 136, 138, 139, 141, 149

WSNR Weighted Signal-to-Noise Ratio. 38

WSSI Wavelet Structural Similarity Index. 40

Chapter 1

Introduction

Advances in technology have enabled ever increasing and affordable connectivity, and the development of a multitude of mobile devices, leading to a well-connected world. An increasingly large proportion of the global population now accesses visual content through the Internet for various purposes such as communication, entertainment, education, sports, social media sharing, and so on. For example, YouTube has over 2 billion users and around one billion hours of video is watched daily [7]. Similarly, social media sharing platforms such as Facebook and Instagram have an enormous user base leading to millions of photos being uploaded on a daily basis. The subscriber base of streaming media platforms such as Netflix and Disney+ is also running into hundreds of millions. The trend of employees working remotely from their homes is on the rise in various industries, thereby increasing the use of videoconferencing tools such as Cisco Webex, Microsoft Teams, Zoom, Skype, etc. In academia, not only are various universities offering fully online degree and diploma programs, but massive open online course platforms, such as edX and Coursera, are offering thousands of courses fully online. Images and videos are fundamental to the success of such online education. It is projected that by 2022 the annual global IP traffic will reach 4.8 zettabytes per year, with videos constituting the vast majority of this traffic at an expected 82% [8].

Visual content undergoes a number of distortions during the processes of acquisition, storage, transmission under bandwidth constraints, and display, any of which can degrade its perceived quality. Given the important role that such content has come to play in our lives, perceptual image and video quality assessment, aiming to assess the quality of visual content as a *human* would perceive it, has become a fundamental problem that is pivotal for the design, optimization and evaluation of various image and video processing algorithms and systems. Image quality assessment (IQA) can be classified into *subjective* and *objective* quality assessment (QA). In subjective QA, humans are tasked to rate the visual quality of content. Since humans are the ultimate receivers of visual content, subjective QA is regarded as the most reliable way to quantify its perceptual quality. However, subjective QA is time consuming, expensive, cannot be embedded in algorithms for optimization purposes, and cannot be deployed in a large-scale and real-time manner. To address these issues, the goal of objective QA is to automatically predict the perceptual quality of visual content as perceived by humans. Traditional objective QA methods such as Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR), which have been used for decades, are found to have poor correlation with perceptual quality of images and videos [9]. Thus, the development of objective quality assessment algorithms that are known to have good correlation with the perceptual quality of content, has not only been the target of intense academic research but such methods have also been adopted and recognized by the industry [10].

1.1 Motivation

Objective IQA algorithms can be further categorized into three major frameworks or paradigms [11,12]: 1) Full-Reference (FR) IQA methods require complete access to the pristine or reference version of a distorted image to evaluate its quality; 2) Reduced-Reference (RR) IQA methods require partial access to the pristine reference image through certain extracted features; 3) No-Reference (NR) or blind image quality assessment (BIQA) methods evaluate the quality of a distorted image in the absence of its reference version. In the last two decades, significant progress has been made in the development of FR IQA algorithms and the performance of state-of-the-art training-free FR methods (such as but not limited to [13–16]) correlates well with human perception of quality while evaluating images afflicted with common distortion types. Notwithstanding these advances, the practical application of FR (and RR) methods remains limited because in real-world media delivery systems, access to pristine reference images is either extremely rare or altogether nonexistent especially at the end-user level. In such practical scenarios, NR IQA or BIQA is the only feasible option.

While a lot of work has been done on the development of general-purpose BIQA methods, their performance is still a considerable distance away from FR IQA and significant room for improvement exists to further enhance their proficiency (as we shall demonstrate in Chapter 2). This is understandable as BIQA is a much more difficult task owing to lack of access to the reference image. To fill the void left by the absence of the reference image, BIQA methods mostly rely on machine learning tools to learn a quality model. This usually involves extracting domain knowledge based features from a set of training images that belong to a subject-rated IQA dataset and then using these features and subjective ratings to train a quality model. However, this approach has shown only limited success (see Chapter 2), mainly because research about truly universal perceptual quality features and the human visual system (HVS) itself remains in its primitive stages. An alternative is to learn perceptually relevant quality features automatically. Machine learning approaches, such as deep neural network (DNN) based techniques, offer such a capability as they not only perform regression but can learn goal-oriented features, thereby offering end-to-end model development. Indeed, DNN based models have enjoyed tremendous successes in the area of visual recognition in the past decade [17]. However, such breakthroughs have not been witnessed in IQA, where DNN based models have offered only limited gains. This is because such models require an adequately large amount of training data. For example, the ImageNet database [18], which has been used widely in the area of visual recognition, has 14 million annotated images. On the contrary, subject-rated datasets in the IQA area are quite small-scale. For example, the largest well-known IQA dataset [19] has only 3,000 annotated images, while the largest subject-rated dataset [20] in IQA has only 12,000 images. Training DNN based models on such datasets leads to quite severe overfitting and generalization issues. Supplementing training data through data augmentation techniques has also witnessed limited success. Thus, the true potential of machine learning techniques, such as DNNs, has not been realized in the area of IQA due to the shortage of large-scale annotated data.

Although creating visual content for large-scale IQA datasets is not a problem, a major bottleneck arises when it comes to annotating such content with perceptually relevant quality labels against which model training can be done later. Such labels are usually assigned by conducting subjective tests, where human subjects come into a controlled laboratory environment to rate the quality of test images. This leads to considerable logistical constraints which means only a small number of images can be rated, thereby imposing severe limitations on the size of IQA datasets, which thus remain small-scale. For example, some large-scale IQA datasets do exist [21,22], however they do not have subjective quality labels and hence cannot be used for model training. Thus, alternative means to annotate new large-scale IQA datasets need to be found. In a few early efforts, researchers have used FR IQA methods to annotate large-scale training data because their performance has matured quite well. A few FR fusion-based methods have recently been proposed and claim to perform better than their constituent FR methods, and perhaps these can be used to annotate large-scale datasets. However, since there are quite a lot of FR methods, each claiming state-of-the-art performance, the choice of selecting one method over the other becomes difficult. The lack of a widespread and common test set makes comparing FR and fused FR methods even more difficult. In any given research area, large-scale performance evaluation surveys prove to be an invaluable resource as they independently compare a number of methods on a wide variety of common test data. However, such surveys in the area of IQA are either quite old missing significant recent developments, or the test data that they use is not diverse which means that their findings cannot be generalized. Thus, in the absence of such surveys, the development of alternative data annotation techniques, that utilize the current state-of-the-art FR or fused FR methods, remains missing.

In practical media delivery systems, visual content undergoes a number of degradations between the source and the final destination, which means that such content has been afflicted with multiple distortions or is multiply distorted. In such practical scenarios, apart from the final distorted version of visual content, its earlier degraded versions are also available at different points in the distribution chain, for instance, at the input and output of an encoder. However, the bulk of IQA research carried out so far and most datasets have focused on the simplified case of singly distorted content. Some datasets that have multiply or authentically distorted images have only recently been developed, but they are also small-scale in nature. In the absence of pristine reference images, FR and RR methods cannot be applied to multiply distorted content, and thus researchers have tried to design NR or BIQA methods for such content. However, these methods have also demonstrated only limited success. More importantly, none of the three major IQA paradigms (FR, RR, and NR) are capable of incorporating additional information about a final distorted image, available in the shape of its earlier degraded versions which can be regarded as *degraded references*, to determine its quality. Since the performance of BIQA methods remains limited, the development of a new paradigm that uses degraded references in the task of quality assessment, may enhance the objective quality prediction performance of IQA methods when evaluating multiply distorted content. However, very few efforts have been directed at the development of this new paradigm and it remains largely missing.

The various challenges mentioned above are hindering the development of robust and practically applicable IQA methods, and become the main motivation behind the work done in this thesis.

1.2 Objectives

The work in this thesis has two main objectives:

1. To address the data shortage challenge in IQA by developing a new very large-scale dataset, composed of both singly and multiply distorted images, and to develop an alternative mechanism to automatically quality-annotate the constituent images without relying on subjective testing.

2. To address the multi-stage distortion challenge by introducing a new IQA paradigm, which we refer to as Degraded Reference (DR) IQA, aiming to build objective quality assessment models that can predict the perceived quality of multiply distorted images when access is available to degraded reference images.

1.3 Contributions

The major contributions of this thesis can be grouped under the following three contexts.

Review and Performance Evaluation of IQA Algorithms

To address the shortcomings of existing surveys in the area of IQA, we conduct so far the most comprehensive review and performance evaluation study of 64 state-of-the-art IQA methods. Specifically, the performance of 43 FR, seven fused FR (22 versions), and 14 NR methods is evaluated on nine subject-rated IQA databases which include five singly and four multiply distorted datasets. A common set of test databases, enables us to make fair comparisons between IQA methods. The diversity of test data also allows for rigorous testing. To the best of our knowledge, this is the largest IQA performance evaluation study to-date and shall prove to be a beneficial resource for both new and seasoned researchers in this area for the foreseeable future. By comprehensively comparing FR and fused FR methods, this study allows us to determine that Reciprocal Rank Fusion (RRF) [23] based FR fusion outperforms all other fused and individual FR methods. Thus, it forms the basis of further work that we carry out to achieve the first objective of this thesis as stated in Section 1.2.

Addressing the Data Challenge

As mentioned in Section 1.1, BIQA methods, including those that employ DNNs, have achieved only limited success. Our study suggests that this is primarily due to the small-scale nature of available subject-rated IQA databases. While researchers have focused on the modeling aspect of the problem, the fundamental issue of the lack of large-scale annotated training data thus far has not received major attention. One major bottleneck in creating such large datasets is the lack of an automatic quality-annotation mechanism that assigns perceptual quality labels to dataset images without requiring ratings from humans. To address this annotated data shortage challenge, we make the following three main contributions:

1. We construct a very large-scale IQA dataset which we call the Waterloo Exploration-II database. This dataset has 3,570 pristine and more than 3.45 million distorted images, making it the largest IQA dataset to-date by a wide margin. This dataset includes both singly distorted images, belonging to three distortion categories, and multiply distorted images belonging to five distortion combinations. Another novelty of this dataset is that we use content adaptive distortion parameters to create distorted content so that the entire quality spectrum can be adequately represented, which is in contrast to the usual practice of using fixed distortion parameters to create IQA databases regardless of content.

2. We develop a novel data annotation mechanism, called Synthetic Quality Benchmark (SQB), to automatically assign perceptually relevant quality ratings to constituent images of an IQA dataset. This mechanism is based on RRF [23] and follows directly from our comprehensive performance evaluation study discussed earlier. Extensive testing of the SQB on nine subject-rated IQA databases reveals that it outperforms all other state-of-the-art FR and fused FR methods. We use SQB to quality-annotate the Waterloo Exploration-II dataset, thereby enabling its utilization for learning based model development.

3. To validate our approach of using large-scale synthetically annotated datasets to resolve the data challenge in IQA, we use the Waterloo Exploration-II database to train a DNN based BIQA method which we call End-to-end Optimized deep neural Network using Synthetic Scores (EONSS). Compared to other DNN based methods, we choose a simple architecture for EONSS as our focus is not on DNN model development but on establishing the impact of data on the performance of DNN based BIQA methods. Extensive testing of EONSS on nine subject-rated IQA databases reveals that it not only comprehensively outperforms existing DNN based BIQA methods, but also the very state-of-the-art in BIQA, thereby establishing the data shortage challenge as the major hurdle that limits existing learning based BIQA methods, and also the efficacy of our approach to address the challenge.

The three contributions stated above and the comprehensive performance evaluation study, help us in achieving the first objective of this thesis, mentioned in Section 1.2.

Degraded Reference Image Quality Assessment

As discussed in Section 1.1, the three major paradigms of FR, RR and NR IQA are unable to handle the practical scenario of evaluating the quality of a multiply distorted image when its earlier distorted version, which we termed as degraded reference, is also available. Such a scenario calls for the development of a new IQA paradigm, which we called DR IQA in Section 1.2. To develop DR IQA and hence address this practical multi-stage distortion challenge, we make the following main contributions:

1. Surprisingly, a comprehensive multiple distortions behavior analysis has remained largely missing thus far in the IQA literature. We make one of the first attempts to analyze the behavior of multiple simultaneous distortions on images. Specifically, we consider the case of a two-stage distortion pipeline and study the behavior of five distortion combinations. This analysis helps us in developing DR IQA models later.

2. We propose two scenarios for the DR IQA framework, where the first scenario considers the pristine reference images to be available in addition to the degraded references and the final distorted images, while the second scenario does not make such an assumption.

3. We construct two new databases specifically for the development of DR IQA models. Thus, they are referred to as DR IQA databases Version 1 (V1) and Version 2 (V2). Each of these datasets consists of 32,912 distorted images overall and contain both singly distorted degraded references and multiply distorted images. They are constructed in a manner similar to the Waterloo Exploration-II database, consist of three single distortion categories, five multiple distortion combinations, and use SQB for data annotation.

4. We develop three major DR IQA models, where 35 parameter settings are developed under the umbrella of each model depending upon various combinations. The first two models are distortion behavior based, where Model 1 follows directly from the multiple distortions behavior analysis and Model 2 follows from Model 1. We also develop Support Vector Regression (SVR) based models under the umbrella of Model

3 to ascertain if machine learning based tools can lead to better results. Extensive analysis on four multiply distorted datasets, which include degraded references, reveals that all three major models lead to more or less similar results and outperform the use of NR methods to directly evaluate the quality of multiply distorted images.

The four contributions mentioned above help us in achieving the second objective of this thesis, as mentioned in Section 1.2.

1.4 Thesis Outline

The rest of this thesis is organized as follows. Chapter 2 discusses the comprehensive review and performance analysis survey of IQA algorithms. It starts with a review of IQA databases that form the test set and analyzes their reference and distorted content. It then provides a review of FR, fused FR, and NR methods being evaluated. This is followed by the evaluation results for FR, fused FR, and NR methods along with associated analysis.

Chapter 3 begins with a discussion on the data challenge in IQA by first using the success of DNNs in the area of visual recognition as a case study and then discussing the challenges holding back similar successes in the area of BIQA. Next, the construction of the very large-scale Waterloo Exploration-II database is presented which is followed by a detailed discussion on the development and extensive testing of the synthetic quality benchmark (SQB) for data annotation. The development of a DNN based BIQA model, EONSS, along with its extensive performance analysis is discussed next as a means to validate our approach to addressing the shortage of annotated data in IQA.

Chapter 4 opens with a discussion on the limitations of FR, RR and NR IQA, and the multiply distorted nature of visual content in the real world. It provides a review of the associated literature and then evaluates the performance of some IQA methods to define a baseline against which the performance of DR IQA models can be evaluated. It then introduces DR IQA as a new paradigm by first providing a detailed analysis on the behavior of multiple distortions and then proposing two scenarios for the DR IQA framework. The construction of DR IQA databases V1 and V2 is discussed next. A detailed account about the development of DR IQA Model 1 based on the multiple distortions behavior analysis is provided, which is followed by the description of another distortion behavior based model (Model 2) and an SVR-based model (Model 3). This chapter is concluded by extensively discussing the performance of the DR IQA models, not only in comparison with the baseline but also with each other.

Finally, Chapter 5 concludes this thesis and points out promising future research directions.

Chapter 2

Review and Performance Evaluation of IQA Algorithms

Image quality assessment (IQA) algorithms aim to predict perceived image quality by human observers. Over the last two decades, a large amount of work has been carried out in the field. New algorithms are being developed at a rapid rate in different areas of IQA, but are often tested and compared with limited existing models using out-of-date test data. There is a significant gap when it comes to large-scale performance evaluation studies that include a wide variety of test data and competing algorithms. In this chapter we aim to fill this gap by carrying out the largest performance evaluation study so far. We test the performance of 43 full-reference (FR), seven fused FR (22 versions), and 14 no-reference (NR) methods on nine subject-rated IQA datasets, of which five contain singly distorted images and four contain multiply distorted content. We use a variety of performance evaluation and statistical significance testing criteria. Our findings not only point to the top performing FR and NR IQA methods, but also highlight the performance gap between them. In addition, we have also conducted a comparative study on FR fusion methods, and an important discovery is that rank aggregation based FR fusion is able to outperform not only other FR fusion approaches but also the top performing FR methods.

2.1 Introduction

Image quality assessment (IQA) can be broadly categorized into *subjective* and *objective* quality assessment (QA). In subjective QA, humans are tasked to evaluate the visual quality of content and the average of subjective ratings is termed as Mean Opinion Score (MOS). Subjective QA is usually regarded as the most reliable method of quantifying perceptual quality of content since in most cases such content is meant to be viewed by humans. However, subjective QA is time consuming, expensive, and cannot be embedded in image processing algorithms for optimization purposes. It is thus the goal of objective QA algorithms to automatically predict the quality of images as perceived by humans. Significant progress has been made in the last two decades in the design of objective QA methods and three major frameworks are now well-established in IQA research [11,12]: 1) Full-Reference (FR) IQA, 2) Reduced-Reference (RR) IQA, and 3) No-Reference (NR) or blind IQA. To evaluate the quality of a distorted image, FR methods require the complete availability of its pristine quality version termed as a reference image, while RR methods require access to certain features that have been extracted from the reference image. On the other hand, NR methods evaluate the quality of the distorted image in the absence of the reference image.

Since the beginning of this century, with the availability of subject-rated datasets, a large number of IQA methods belonging to all three frameworks (FR, RR, NR) have been proposed. These methods are tested on one or more subject-rated datasets and claim state-of-the-art performance. Given the large number of IQA methods that now exist, a number of challenges arise when it comes to selecting the top performing methods within and across different IQA frameworks for various purposes: 1) It can be respectively seen from Tables 2.3, 2.4, and 2.5 that different FR, fused FR, and NR methods are tested on different sets of subject-rated datasets, and thus straightforward performance comparison becomes difficult. 2) It is also evident from Tables 2.3, 2.4, and 2.5 that IQA methods are usually tested (and at times trained) on singly distorted subject-rated datasets that contain different distortion types, but typically, each distorted image has been afflicted with a single stage of distortion [19,24–30]. This is in contrast to real world media distribution systems where the same visual content can undergo a number of distortions, during the processes of acquisition, transmission, and storage, before reaching the end user. While some IQA datasets with multiply distorted images are now available [31–34], only a limited number of IQA methods have used some of them for testing purposes. 3) General-purpose NR methods, which either rely on handcrafted features or on end-to-end learning, require training which is usually done on subject-rated IQA datasets where MOS acts as ground truth. While such training requires the availability of a large amount of data, subject-rated datasets offer only a small amount of annotated data. For example, the largest well-known subject-rated singly distorted database has a total of 3,000 distorted images [19], while there are only 1,600 distorted images in the largest multiply distorted database [33]. The number of images in individual distortion categories is even smaller. Such constraints make it difficult to avoid model overfitting and raises questions about the generalizability of NR methods trained on these datasets (as will become evident later in this chapter). To circumvent these issues, large-scale annotated datasets are required that consist of thousands of pristine reference and hundreds of thousands if not millions of distorted images. These datasets should have a wide variety of distortions and distortion combinations along with appropriately selected distortion intensity levels that cover the entire range of the quality spectrum with adequate density. However, given the limitations of subjective testing, it is not possible to obtain quality ratings from humans for such large datasets. Clearly, alternative methods for annotating large-scale IQA datasets are desired. Since the area of FR IQA has matured quite well, one possible alternative is to replace subjective ratings with scores from reliable FR methods. In fact, a number of works in IQA literature have already used either FR scores [35–40] or fused FR scores [41] as replacement of subjective ratings. However, their choice of FR methods seems rather ad hoc as detailed analysis about method selection has not been provided. Essentially the following questions remain unanswered while using FR scores for annotating large-scale IQA datasets as alternatives to subjective ratings: i) Which FR method or methods should be selected? ii) Can fused FR methods offer any further advantages over individual methods?

To address the above-mentioned challenges, a comprehensive survey of the performance of IQA methods, especially FR and fused FR methods, is desired that gauges their performance on a large and diverse set of subject-rated IQA datasets. A number of reviews and surveys have been conducted in the field of IQA over the past decade or so. The performance of ten FR IQA methods was evaluated on the LIVE R2 database [42] in [24]. Performance evaluation criteria included the Pearson Linear Correlation Coefficient (PLCC), Root-Mean-Squared Error (RMSE), Spearman Rank-order Correlation Coefficient (SRCC), and statistical significance testing. A description of 111 FR IQA methods is given in [43], however performance evaluation was not carried out. A comprehensive review of basic computational building blocks used in the design of perceptual IQA metrics is given in [44] along with a description of six FR IQA methods. The performance of these methods is evaluated on seven IQA databases (A57 [27], CSIQ [26], IVC [30], LIVE R2 [24], MICT [29], TID2008 [25], and WIQ [28]) in terms of PLCC and SRCC. A classification, description, and evaluation of 22 FR methods is provided in [45], where PLCC and SRCC are used for performance evaluation on six datasets which include IVC [30], TID2008 [25], and four other datasets whose description can be found in [45]. In [46], the performance of 11 FR methods was evaluated on seven IQA datasets (A57 [27], CSIQ [26], IVC [30], LIVE R2 [24], MICT [29], TID2008 [25], and WIQ [28]). PLCC, RMSE, SRCC, and Kendall Rank-order Correlation Coefficient (KRCC) were used as evaluation criteria. The computational complexity of these methods was evaluated in terms of their running speed. Various aspects of subjective and objective IQA are surveyed in [47] including: description of four subjective testing methods, description of seven FR IQA methods for standard dynamic range (SDR) images, description of two FR methods for the IQA of reference and test images with different dynamic ranges, description of six IQA datasets, and performance evaluation of seven SDR FR IQA methods on three datasets (CSIQ [26], LIVE R2 [24], TID2008 [25]) in terms of PLCC, SRCC, KRCC, RMSE, and Mean Absolute Error (MAE). In addition, the performance of an FR method for the IQA of tone mapped images (TMQI [48]) is evaluated, the computation time of different FR methods is presented, and the IQA of three-dimensional images is discussed. In [49], several objective IQA methods along with seven datasets are briefly discussed, and the performance of eight FR, three RR, and eight NR methods is evaluated on the LIVE R2 database [24] in terms of PLCC, SRCC, RMSE, and MAE. In [50], the performance of 60 FR methods was evaluated on the CIDIQ database [5] which provides subjective ratings at two viewing distances. PLCC, SRCC, and KRCC were used as performance evaluation criteria. A survey of Natural Scene Statistics (NSS) and learning based non-distortion-specific (general-purpose) NR IQA methods was performed in [51], where the design of 12 NR methods was reviewed and the performance of nine such methods was evaluated on three IQA databases (LIVE R2 [24], CSIQ [26], and TID2008 [25]). PLCC, SRCC, and statistical significance testing (only on LIVE R2 database) were used as performance evaluation criteria. For comparison, four FR methods are included in the performance evaluation. The computational complexity of six NR methods was also compared. Several distortion-specific and general-purpose NR IQA approaches were reviewed in [52], along with the performance evaluation of eight NR methods on three datasets (CSIQ [26], LIVE R2 [24], TID2013 [19]) in terms of PLCC and SRCC. The computational complexity of these methods was determined in terms of their execution time. In a recent survey [53], different areas of IQA are reviewed including two-dimensional (2D) image fidelity assessment (FR, RR, NR), three-dimensional (3D) image fidelity assessment (FR, NR), image aesthetics assessment, and 3D image visual comfort assessment. In the category of 2D image fidelity assessment, the performance of 20 FR, one fused FR, five RR, and 10 NR IQA methods is evaluated on four datasets (CSIQ [26], LIVE R2 [24], TID2008 [25], TID2013 [19]) in terms of PLCC, SRCC and RMSE. A summary of these earlier IQA reviews and surveys is given in Table 2.1.

Existing IQA surveys suffer from a number of shortcomings: 1) The earlier ones [24, 43–45] do not include state-of-the-art FR methods. 2) While conducting performance evaluation, none of these surveys utilize multiply distorted IQA datasets (in some cases this is because such datasets did not exist at the time of the survey). This puts into question the assumptions made about algorithm performance while being tested on limited data (singly distorted datasets only). 3) With the exception of [50], some recent singly distorted datasets (VCLFER [54], CIDIQ [5]) are missing in these surveys. 4) Some surveys use a single dataset [24,49,50], which limits content diversity and raises concerns about the generalization of their findings. 5) None of the surveys evaluates the performance of fused FR methods with the exception of [53] which evaluates only a single FR fusion method. 6) Some surveys [51,52] are specific to the evaluation of NR methods. 7) With the exception of [24, 51], statistical significance testing is missing in these surveys. Since IQA datasets can only be regarded as small and sparse random samples from the enormous space of all possible natural images and their distorted versions, the lack of such testing puts into

Table 2.1: Summary of IQA performance evaluation surveys.

Survey	Year	Number of			Statistical Significance Testing
		Methods Evaluated	Databases Used		
			SDB[a]	MDB[b]	
Sheikh et al. [24]	2006	10 FR	1	0	Yes
Pedersen and Hardeberg [43]	2009	Description of 111 FR Methods			No
Lin and Kuo [44]	2011	6 FR	7	0	No
Pedersen and Hardeberg [45]	2012	22 FR	6	0	No
Zhang et al. [46]	2012	11 FR	7	0	No
Mohammadi et al. [47]	2014	7 FR	3	0	No
He et al. [49]	2014	19 (8 FR, 3 RR, 8 NR)	1	0	No
Pedersen [50]	2015	60 FR	1	0	No
Manap and Shao [51]	2015	13 (9 NR, 4 FR)	3	0	Yes
Xu et al. [52]	2017	8 NR	3	0	No
Niu et al. [53]	2019	36 (20 FR, 10 NR, 5 RR, 1 Fused FR)	4	0	No
This work	2019	64[c] (43 FR, 14 NR, 7[c] Fused FR)	5	4	Yes

[a]SDB: Singly Distorted Databases (Images afflicted with one distortion at a time).

[a]MDB: Multiply Distorted Databases (Images afflicted with multiple distortions at the same time).

[c]22 versions of the seven Fused FR methods were tested, which if taken into account separately means that we evaluated 79 methods.

question the universal nature of the findings in these surveys. 8) Although the survey in [53] is quite recent, it does not evaluate the performance of IQA methods on multiply distorted datasets, does not use the singly distorted datasets VCLFER [54] and CIDIQ [5], does not perform statistical significance testing, evaluates only a single fused FR method, and does not evaluate the performance of some state-of-the-art FR and NR IQA methods. Reference [53] uses both TID2008 [25] and TID2013 [19] datasets, where the latter contains all the reference and distorted images of the former. Given these shortcomings, it is evident that existing surveys are unable to identify the top performing FR, fused FR, and NR methods in a competitive and comparative setting. They are also unable to answer the question about the choice of FR or fused FR methods as alternatives to subjective ratings.

In this chapter, we attempt to address the limitations of existing IQA surveys by carrying out a comprehensive review and performance evaluation of 64 IQA methods, of which 43 are FR and seven are fused FR methods. We also include 14 NR methods in our study to provide a more thorough snapshot of the field. We tested 22 versions of the seven fused FR methods, and thus collectively a total of 79 IQA methods were evaluated. We test on nine subject-rated datasets, of which five are singly distorted and four are multiply distorted datasets. This ensures that the methods under evaluation are tested on as wide a range of reference and distorted content as possible. Apart from the usual correlation coefficient based comparison criteria, we also compare IQA methods through statistical significance testing in order to make statistically sound conclusions. To the best of our knowledge, this is the largest evaluation study carried out in IQA literature and thus is the first major contribution of this thesis. In addition to FR and NR IQA that are surveyed and evaluated in this chapter, there are other types of IQA problems such as reduced-reference (RR) IQA [11,12], and IQA of reference/test images across different spatial resolutions [55], frame rates [56,57], dynamic ranges [48], exposure levels [58], focus points [59], color/gray tones [60], and viewing devices [61], that are beyond the major focus of the current work.

The rest of this chapter is organized as follows. A review of IQA datasets and methods included in this study is provided in Sections 2.2 and 2.3, respectively. The performance of FR and fused FR IQA methods is thoroughly evaluated in Section 2.4 while that of NR methods is evaluated in Section 2.5. Section 2.6 concludes this chapter.

2.2 Review of IQA Databases

Over the last 15 years, a significant number of IQA databases with human rated image quality ratings have come out. Although recommendations have been made about the conduct of subjective testing and content selection [70–72], a *gold standard* remains elusive and the optimal method for subjective testing is still an open problem. As is evident from Table 2.2 and the following sections, IQA datasets use a variety of subjective testing methodologies, viewing distances, and ratings per image. Their benchmark quality ratings have different ranges and are either in the form of Difference Mean Opinion Score (DMOS) or Mean Opinion Scores (MOS). Reference image content is usually selected in an ad hoc manner and different distortions are simulated by degrading the reference content at differ-

Table 2.2: Summary of IQA databases used in this work.

Database	Year	No. of Images		Distortion List (No. of Images)	Distortions	Subjective	Subjective	Score	Ratings	Viewing
		Ref.	Dist.		per Image	Test Method	Data Type	Range	per Image	Distance
LIVE R2 [24, 42]	2006	29	779	1. White Gaussian Noise (145) 2. Gaussian Blur (145) 3. JPEG Compression (175) 4. JPEG2000 Compression (169) 5. Fast Fading Rayleigh Channel (145)	1	Single Stimulus	DMOS	-2.64 to 111.77	≈ 23	2 - 2.5 Screen Heights
TID2013 [19, 62]	2013	25	3000	1. Additive Gaussian Noise (125) 2. Additive Noise is more intensive in color components (125) 3. Spatially Correlated Noise (125) 4. Masked Noise (125) 5. High Frequency Noise (125) 6. Impulse Noise (125) 7. Quantization Noise (125) 8. Gaussian Blur (125) 9. Image Denoising (125) 10. JPEG Compression (125) 11. JPEG2000 Compression (125) 12. JPEG Transmission Errors (125) 13. JPEG2000 Transmission Errors (125) 14. Non Eccentricity Pattern Noise (125) 15. Local Block-wise Distortions of different intensity (125) 16. Mean Shift (Intensity Shift) (125) 17. Contrast Change (125) 18. Change of Color Saturation (125) 19. Multiplicative Gaussian Noise (125) 20. Comfort Noise (125) 21. Lossy Compression of Noisy Images (125) 22. Image Color Quantization with Dither (125) 23. Chromatic Aberrations (125) 24. Sparse Sampling and Reconstruction (125)	1 to 2	Pair-wise Comparison	MOS	0.24 to 7.21	≈ 30	Varying
CSIQ [26, 63]	2010	30	866	1. Additive White Gaussian Noise (150) 2. Gaussian Blur (150) 3. JPEG Compression (150) 4. JPEG2000 Compression (150) 5. Additive Pink Gaussian Noise (150) 6. Global Contrast Decrements (116)	1	Simultaneous Comparison	DMOS	0 to 1	≈ 6	70 cm
VCLFER [54, 64]	2012	23	552	1. Additive White Gaussian Noise (138) 2. Gaussian Blur (138) 3. JPEG Compression (138) 4. JPEG2000 Compression (138)	1	Single Stimulus	MOS	1.57 to 96.52	16 to 36	INP*
CIDIQ [5, 65]	2014	23	690	1. Poisson Noise (115) 2. Gaussian Blur (115) 3. JPEG Compression (115) 4. JPEG2000 Compression (115) 5. SGCK Gamut Mapping (115) 6. ∆E Gamut Mapping (115)	1	Double Stimulus	MOS	1.18 to 7.65 1 to 7.76	17	50 cm 100 cm
LIVE MD [31, 66]	2012	15	405	1. Gaussian Noise (45) 2. Gaussian Blur (45) 3. JPEG Compression (45) 4. Gaussian Blur + JPEG compression (135) 5. Gaussian Blur + Gaussian Noise (135)	1 1 1 2 2	Single Stimulus	DMOS	0.61 to 84.67	≈ 19	4 Screen Heights
MDID2013 [32]	2014	12	324	1. Gaussian Blur followed by JPEG compression followed by White Gaussian Noise (324)	3	Single Stimulus	DMOS	0.32 to 0.55	25	4 Image Heights
MDID [33, 67]	2017	20	1600	May include (Gaussian blur and/or contrast change) followed by (JPEG or JPEG2000 compression) followed by (Gaussian noise)	1 to 4	Pair Comparison Sorting	MOS	0.08 to 7.92	33 to 35	2 Screen Heights
MDIVL [34, 68, 69]	2017	10	750	1. Gaussian Blur followed by JPEG Compression (350) 2. Gaussian Noise followed by JPEG Compression (400)	2	Single Stimulus	MOS	1.41 to 97.97	≈ 12	INP*

*INP: Information Not Provided by authors.

ent distortion intensity levels which are themselves picked in an ad hoc manner. While the target is to have distorted images such that the quality spectrum is uniformly represented, this is often not the case (as discussed later). A majority of IQA datasets consider the simplified case of images undergoing a single distortion which is in contradiction to practical scenarios where content typically undergoes multiple distortions. Given the arbitrary nature of such benchmark data, it is unsurprising that at times the performance of IQA methods varies widely across different datasets. Thus, it is vital to test the performance of IQA methods on as many publicly available datasets as possible [73] in order to reliably test their robustness.

To mitigate dataset specific impacts on the performance evaluation of IQA methods, in this work we choose a large number of databases to carry out such an assessment. We use four database selection criteria, specifically we use databases that contain: 1) Natural images, 2) Color images, 3) Both reference and distorted content to enable evaluation of FR IQA methods, and 4) Standard Dynamic Range (SDR) images, that is, images with a bit depth of 8 bits per pixel per color channel. Following these criteria, we have selected nine databases which simulate distortions at various intensity levels. Five of these datasets can be classified as singly distorted databases while four fall under the multiply distorted category. Table 2.2 presents a summary of these databases while they are briefly introduced in the next two sub-sections. This is followed by a description of some other IQA databases and the reasons for not including them in our current work. We close this section by a discussion on the range of reference and distorted content in the datasets used in this work for algorithm testing.

2.2.1 Single Distortion Databases

These datasets are also referred to as *singly distorted* databases. While they contain a wide range of distortions, each distorted image is afflicted with only one kind of distortion. Until recently, a majority of IQA datasets fell under this category.

The LIVE Release 2 (LIVE R2) database [24, 42], developed by the Laboratory for Image and Video Engineering at UT Austin, is one of the most widely used IQA datasets. It consists of 29 reference and 779 distorted images. The database has five distortion types and up to five distortion intensity levels within each type. Images either have a resolution of 480×720 or up to 768×512. Subjective testing was carried out on 21″ CRT monitors and followed the single stimulus methodology [70] where reference images were also evaluated. After undergoing a short training session, subjects rated the quality of test images by moving a slider on a quality scale that was demarcated with five words: Bad, Poor, Fair, Good, and Excellent. A quality score in the range of [1, 100] was obtained from the slider location. Seven sessions of testing were done in order to minimize observer fatigue and scale realignment was carried out to match the quality scale of all sessions. The database provides subjective data in the form of DMOS after outlier removal, where better quality is represented by a lower DMOS. Further details about the database are provided in Table 2.2.

The Tampere Image Database 2013 (TID2013) [19, 62] builds further upon the earlier TID2008 database [25]. It consists of 25 reference images (of which 24 are natural and one is artificial) and 3,000 distorted images. The database has 24 distortion types and five distortion levels per type. All images have a resolution of 512×384. A total of 971 subjects in five different countries took part in subjective testing. Experiments were carried out either in the laboratory environment or remotely via internet, and subjects were given prior instructions about the testing process. A tristimulus methodology [19] was adopted to conduct the subjective tests where subjects observe a pair of distorted images in the presence of their reference image and select the better of the two. Tests were conducted mostly on 19″ LCD or CRT monitors. Each distorted image was part of nine pair-wise comparisons. The winning image in each pair received one point and a final score for an image was obtained by summing the winning points. After outlier removal, MOS was obtained for the database, where higher MOS represents better quality. Although we are classifying TID2013 under the single distortion category, it should be noted that some of its distortion types are multiply distorted in nature (for example, lossy compression of noisy images). See Table 2.2 for more details.

The Computational and Subjective Image Quality (CSIQ) database [26, 63] consists of 30 reference and 866 distorted images. It has six distortion types and four to five levels of distortion per type. All images have a resolution of 512 × 512. Subjective tests were carried out by placing four 24″ LCD monitors side-by-side such that their viewing distance from the subject was equal. All the distorted images derived from the same reference were simultaneously displayed on the monitor array and each subject horizontally ordered images based on their perceived quality [63]. Cross-image ratings were obtained in order to carry out realignment of the quality scale between different content. After outlier removal DMOS was obtained, where a lower DMOS value represents better quality. Further details about the database are provided in Table 2.2.

The Video Communications Laboratory @ FER (VCLFER) database [54, 64] is composed of 23 reference and 552 distorted images. It has four distortion types and six distortion levels per type. Images in VCLFER either have a resolution of up to 771 × 512 or up to 512 × 771. Subjective testing was conducted by following the single stimulus methodology [70] and by employing a numeric scale with 100 grades. After removing outliers, the results for each subject were rescaled in the range of [0, 100], and MOS for the overall database was computed. A higher MOS value is indicative of better visual quality. See Table 2.2 for more details.

The Colourlab Image Database: Image Quality (CIDIQ) [5, 65] consists of 23 reference and 690 distorted images. It has six distortion types and five distortion levels per type. All images in CIDIQ have a resolution of 800 × 800. Subjective testing was carried out in accordance with the recommendations of CIE [74] and ITU [70]. A double stimulus methodology was followed where two images were displayed simultaneously, and category judgment was used to record responses from subjects. The rating scale had nine categories where the odd numbered categories from 1 to 9 were respectively labeled as Bad, Poor, Fair, Good, and Excellent quality. The actual subjective test was preceded by a training sequence. The CIDIQ database is unique in that it carried out subjective testing at two viewing distances, that of 50 cm and 100 cm. Therefore, it provides two sets of MOS, one for each viewing distance. A higher MOS value represents better visual quality. Further details about the database are provided in Table 2.2.

2.2.2 Multiple Distortion Databases

These datasets are also referred to as *multiply distorted* databases and contain images such that an individual distorted image may have undergone multiple (two or more) distortions, thereby better mimicking practical content distribution scenarios.

The LIVE Multiply Distorted (LIVE MD) database [31, 66] is the first IQA dataset that has been specifically designed for images with multiple simultaneous distortions. The database has 15 reference and 405 distorted images of which 135 are singly distorted while 270 are multiply distorted. LIVE MD has three distortion types (Gaussian blur, JPEG compression, and white Gaussian noise) and three distortion levels per type. Apart from containing singly distorted images belonging to each of the three distortion types, the database has two multiple distortion combinations of 1) Gaussian blur followed by JPEG compression and 2) Gaussian blur followed by white Gaussian noise contamination. All images in the database have a resolution of 1280 × 720. Subjective testing was conducted by following the single stimulus [70] with hidden reference methodology. After going through a training session, subjects rated the quality of test images by moving a slider on a continuous scale from 0 to 100 which was also labeled with the words, Bad, Poor, Fair, Good, and Excellent. The test was divided into two parts based on the multiple distortion combinations and each part had two sessions of 30 minutes each. The database provides subjective scores in the form of DMOS, where a lower value is indicative of better visual quality. Further details about the database are provided in Table 2.2.

The Multiply Distorted Image Database 2013 (MDID2013) [32] is composed of 12 reference and 324 multiply distorted images. The database uses the same distortion parameters as the LIVE MD database [31]. MDID2013 uses three distortion types (Gaussian blur, JPEG compression, and white Gaussian noise) and three distortion levels per type. It contains just one multiple distortion combination, where a reference image first undergoes Gaussian blurring which is followed by JPEG compression followed by white noise contamination. Images in MDID2013 have a resolution of up to 1280 × 720. The single stimulus methodology [70] was followed to conduct the subjective test and ratings were obtained on a continuous quality scale from 0 to 1. After outlier removal, DMOS for the database was computed, where a lower value signifies better visual quality. See Table 2.2 for more details.

The Multiply Distorted Image Database (MDID) [33, 67] (different from MDID2013) contains 20 reference and 1,600 multiply distorted images. The database uses five types of distortions: Gaussian noise, Gaussian blur, contrast change, JPEG, and JPEG2000 compression. Four intensity levels are set for each distortion type. Distortions are introduced in three steps in the following order: 1) To simulate image acquisition, Gaussian blur and/or contrast change are added first in either order. 2) Image transmission is simulated by compressing the image from the first step, either by using JPEG or JPEG2000 compression (one compression technique only). 3) Finally, display imperfections are simulated by adding Gaussian noise to the image from the second step. In each of these steps, distortion intensity levels, including the no-distortion case, are picked at random. However, it is ensured that the following three rules are obeyed: 1) At least one distortion is introduced, 2) Only one compression technique (JPEG or JPEG2000) is used, and 3) Repetition of distortions is avoided. Thus, each distorted image may be afflicted with one to four distortions. MDID creates 80 distorted images for each reference image and provides details about the distortion process for each image. All images in MDID have a resolution of 512×384. The pair comparison sorting methodology [33] is used to conduct subjective testing, where two images are simultaneously displayed along with their reference and subjects are required to rate the quality of one distorted image with respect to the other by using one of three possible rating options: better, worse, or equal quality. Testing was carried out on a 19$''$ LCD monitor and was preceded by a training session. Following outlier removal and data normalization, MOS for the database is computed, where a higher value is indicative of better visual quality. Further details about the database are provided in Table 2.2.

The Multiple Distorted IVL database (MDIVL) [34,68,69] consists of 10 reference and 750 multiply distorted images. The database is divided into two parts based on two multiple distortion combinations: 1) Blur-JPEG, where each reference image undergoes seven levels of Gaussian blur and then each blurred image undergoes five levels of JPEG compression, and 2) Noise-JPEG, where each reference image undergoes ten levels of Gaussian noise and then each noisy image undergoes four levels of JPEG compression. All images in the database have a resolution of 886×591. Subjective testing followed the single stimulus methodology [70]. Subjects recorded their ratings on a continuous quality scale from 0 (Worst quality) to 100 (Best quality). To minimize fatigue effect, subjective testing was conducted in several sessions where each session had around 100 images and did not exceed 30 minutes. MOS was computed for the database after outlier removal, where a higher value indicates better quality. See Table 2.2 for more details.

2.2.3 Other IQA Databases

Apart from the nine datasets mentioned in Sections 2.2.1 and 2.2.2, a number of other datasets have been mentioned in Tables 2.3, 2.4 and 2.5 that follow in the subsequent sections. Information about these and some other datasets follows.

The A57 database [27, 75] contains three reference and 54 distorted images. It consists of grayscale images with a resolution of 512 × 512. The dataset has six distortion types which include: 1) Gaussian white noise, 2) Gaussian blur, 3) Baseline JPEG compression, 4) Baseline JPEG2000 compression, 5) JPEG2000 compression with dynamic contrast based quantization, and 6) Flat allocation (equal distortion contrast at all scales). Each distortion was applied at three distortion intensity levels. The MICT-Toyama database [29] contains 14 reference and 168 distorted images. It consists of color images with a resolution of 768 × 512. The dataset contains two distortion types: 1) JPEG compression and 2) JPEG2000 compression, and six distortion levels per type. The single stimulus methodology was used to acquire subjective ratings, using a five category discrete quality scale and through the participation of 16 subjects, on a 17″ CRT display at a viewing distance of four times the picture height. The IVC database [30] contains ten reference and 185 distorted images. The dataset consists of color images with a resolution of 512 × 512. It has four distortion types which include: 1) JPEG compression, 2) JPEG2000 compression, 3) Local adaptive resolution (LAR) coding, and 4) Blurring. The subjective ratings for IVC were obtained by following the double stimulus methodology with five rating categories. 15 observers participated in the test and viewed the content at a distance of six times the screen height. The TID2008 database [25, 76] is an earlier version of the TID2013 database [19], and contains 25 reference and 1,700 distorted images. It has color images with a resolution of 512 × 384. The dataset contains 17 distortion types and four distortion levels per type. For a list of distortions contained in TID2008, refer to the first 17 distortions listed in Table 2.2 for the TID2013 database. Subjective testing for TID2008 was carried out by using the same methodology as was later used for TID2013 (described in Section 2.2.1). The Wireless Imaging Quality (WIQ) database [28] contains seven reference and 80 distorted images. It consists of grayscale images with a resolution of 512 × 512. The dataset simulates a wireless link distortion model by passing JPEG encoded images through an uncorrelated Rayleigh flat fading channel in the presence of additive white Gaussian noise. Two subjective tests were performed at different locations, on 17″ CRT monitors at a viewing distance of four times the picture height. The double stimulus continuous quality scale (DSCQS) [70] methodology was followed to conduct the tests. The Waterloo Exploration database [21] is a very large dataset that is composed of 4,744 reference and 94,880 distorted images. It has color images of various resolutions. The dataset contains four distortion types: 1) White Gaussian noise, 2) Gaussian blur, 3) JPEG compression, and 4) JPEG2000 compression. Each distortion is applied at five fixed intensity levels. Since the database consists of such a large number of images, subjective testing is not possible. Instead, three alternative testing criteria are proposed in [21] for the performance evaluation of objective IQA models. These include: 1) the pristine/distorted image discriminability test (D-test), 2) the listwise ranking consistency test (L-test), and 3) the pairwise preference consistency test (P-test).

The above-mentioned datasets have not been used in the current work for the following reasons: The A57 and WIQ datasets are composed of grayscale images which does not fulfill one of our database selection conditions, that a dataset should be composed of color images. This condition is required to provide a uniform comparison basis, as some of the objective IQA methods that we test are designed to take the color aspect into account. Besides, these datasets are composed of only a small amount of source and distorted content. The MICT-Toyama dataset has not been selected as 11 of its 14 reference images are found in LIVE R2 dataset while the cropped versions of all its reference images are found in the TID2013 reference image set. Both LIVE R2 and TID2013 datasets contain the two distortion types found in MICT-Toyama. Since we are including LIVE R2 and TID2013 in our analysis, we believe that including MICT-Toyama would be redundant. The TID2008 dataset has not been included since all of its reference content, distortion types and levels are found in its enhanced version TID2013. The IVC dataset contains a small number of test images per distortion type and three of its four distortion types (Blur, JPEG and JPEG2000 compression) are effectively covered in the five single distortion databases that we have selected for testing. Although the Waterloo Exploration database is one of the largest available IQA datasets, we have not used it because of the unavailability of subjective ratings.

The databases discussed above, and in Sections 2.2.1 and 2.2.2, belong to the category of *simulated distortion* databases, where a number of pristine reference images are first obtained and then artificially degraded with different types and levels of distortions in a controlled manner. By contrast, *authentic distortion* databases constitute another category of IQA datasets, where distortions are captured directly in real-world environments. It is difficult to categorize images into different distortion types and intensity levels in such datasets. The following four databases fall in the authentic distortion category. The Blurred Image Database (BID) [77] consists of 585 images, taken by human users, that represent realistic blur distortions. Images are classified into five blur classes which include unblurred images, out-of-focus blur, simple motion blur, complex motion blur, and other kinds of blur. Subjective testing was carried out by using a single stimulus methodology on a continuous quality scale marked with labels (Excellent, Good, Fair, Poor, and Bad). The Camera Image Database 2013 (CID2013) [78] consists of 480 images captured by 79 different cameras of varying quality. Different types of cameras were used to capture images, including mobile phone cameras, compact cameras and SLR cameras. The database is divided into six smaller datasets each of which is composed of six different scenes that have been captured by 12-14 different cameras. A dynamic reference method [78] was proposed and used to conduct the subjective test. The subjects first saw a slideshow of the test images to get an overall idea of quality variation, and then saw each image in a single stimulus manner where they could give quality ratings on a continuous scale. Besides MOS, subjective evaluations for the attributes of sharpness, graininess, brightness, and color saturation are also provided. The LIVE in the Wild Image Quality Challenge (LIVE WC) database [79] is composed of 1,162 images taken by a diverse set of mobile device cameras. The images in this dataset depict a wide variety of real-world scenes. The subjective study was performed online by using the Amazon Mechanical Turk [80], which is a crowdsourcing platform. The single stimulus methodology was employed where subjects recorded their quality ratings on a continuous scale that was divided into five parts with appropriate labels (Excellent, Good, Fair, Poor, and Bad). Besides the subjective test to provide MOS, a separate experiment was conducted to obtain subjective opinion about the distortion category that a test image may belong to. Distortion categories included blurry, grainy, overexposed, underexposed, and no apparent distortion. A majority voting policy was adopted to arrive at a distortion category for an image. A recent database called KonIQ-10K [81] consists of 10,073 images and is by far the largest among the authentic distortion databases. The source of the KonIQ-10K images is the very large-scale YFCC100M multimedia database [82] which has 100 million Flickr based media objects (images and videos). Initially 10 million images were randomly picked from the YFCC100M database from which 10,073 authentically distorted images were sampled through the use of content and quality based indicators. The subjective study was carried out online through a crowdsourcing platform [83]. A five-point absolute category rating (ACR) scale was used to obtain subject ratings, where a rating of 1 indicated bad while that of 5 indicated excellent quality. The database provides subjective ratings in terms of MOS. In this work, we have not used authentic distortion datasets because they lack the presence of reference images, which renders them unusable for the evaluation of FR IQA methods. Nevertheless, these datasets are a valuable resource and should be used in studies that are exclusive to NR IQA methods.

A number of datasets composed of content other than natural images have been constructed. The Screen Image Quality Assessment Database (SIQAD) [84] consists of 20 reference and 980 distorted screen content images. It follows the single stimulus methodology to obtain subjective scores on an 11 point numerical scale. The Document Image Quality dataset [85] selected 25 documents from publicly available document datasets and used a smart phone camera at varying distances to capture 175 document images. The dataset provides Optical Character Recognition (OCR) accuracy as a measure of quality that has to be predicted by objective methods. The Newspaper dataset [86] is composed of 521 grayscale text zone images derived from a collection of newspaper images. As ground truth, the dataset provides OCR accuracy results. Since our focus is on natural images, we have not utilized these datasets in this work.

A valuable compilation of various image and video quality databases can be found at [87].

2.2.4 Content Analysis

The space of all possible natural images is enormous. Ideally, an IQA database should properly reflect the statistical distribution of natural image content, or contain diverse content type for a wide coverage. In practical IQA databases, however, the large natural image space is often represented by just a few source or reference images. From Table 2.2, it can be seen that subject-rated datasets usually have 10 to 30 reference images. Limitations on the amount of source content are encountered due to the constraints of subjective testing. For example, even with just 25 reference images, the TID2013 database [19] has 3,000 distorted images, which leads to significant challenges in obtaining human ratings. The limited source content that a dataset has, should thus be as diverse as possible in order to sample different parts of the space of all possible natural images. This is also an important reason for selecting as many subject-rated IQA databases as possible while testing a new algorithm, so that its performance can be gauged on as wide a set of source content as possible.

Usually the variety in reference content is described in subjective terms, such as the presence of people, human faces, landscapes, animals, closeup or wide-angle shots, buildings, indoor or outdoor shots, and so on. However, a few quantitative descriptors have also been used to describe such content. In [88], image spatial information (SI) and colorfulness (CF) have been used to represent the dimensions of space and color respectively, and the SI versus CF space has been proposed as a 2D space to represent the diversity of source content. In this work, we use the SI versus CF space to examine the range of source content in the nine IQA datasets under consideration.

SI is used to determine edge energy in an image [88]. Different SI measures have been found to have high correlation with compression based image complexity measures [89]. A standard deviation based SI measure (SI_{std}) was recommended in [71] while a root mean square based measure (SI_{rms}) was used in [88]. However in [89], SI_{std}, SI_{rms}, and a mean based SI measure (SI_{mean}) were compared and it was found that SI_{mean} has the highest correlation with compression based image complexity measures. Therefore, we will use SI_{mean} for further analysis in this work. To obtain SI_{mean}, a color image is first converted to grayscale and then filtered with horizontal $\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$ and vertical $\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ Sobel filters, leading to images s_h and s_v respectively. The pixel-level edge magnitude is then defined as [88, 89]:

$$s_{mag} = \sqrt{s_h^2 + s_v^2} \tag{2.1}$$

And SI$_{mean}$ is obtained as [89]:

$$\mathrm{SI}_{mean} = \frac{1}{N} \sum s_{mag} \tag{2.2}$$

where N is the number of pixels in the image.

CF is an indicator of the variety and intensity of colors in an image [88]. A computationally efficient CF measure was proposed in [90] which correlates well with subjective measurements of colorfulness. Assuming an image in the sRGB color space, it is first transformed to an opponent color space as follows [90]:

$$rg = R - G \tag{2.3}$$

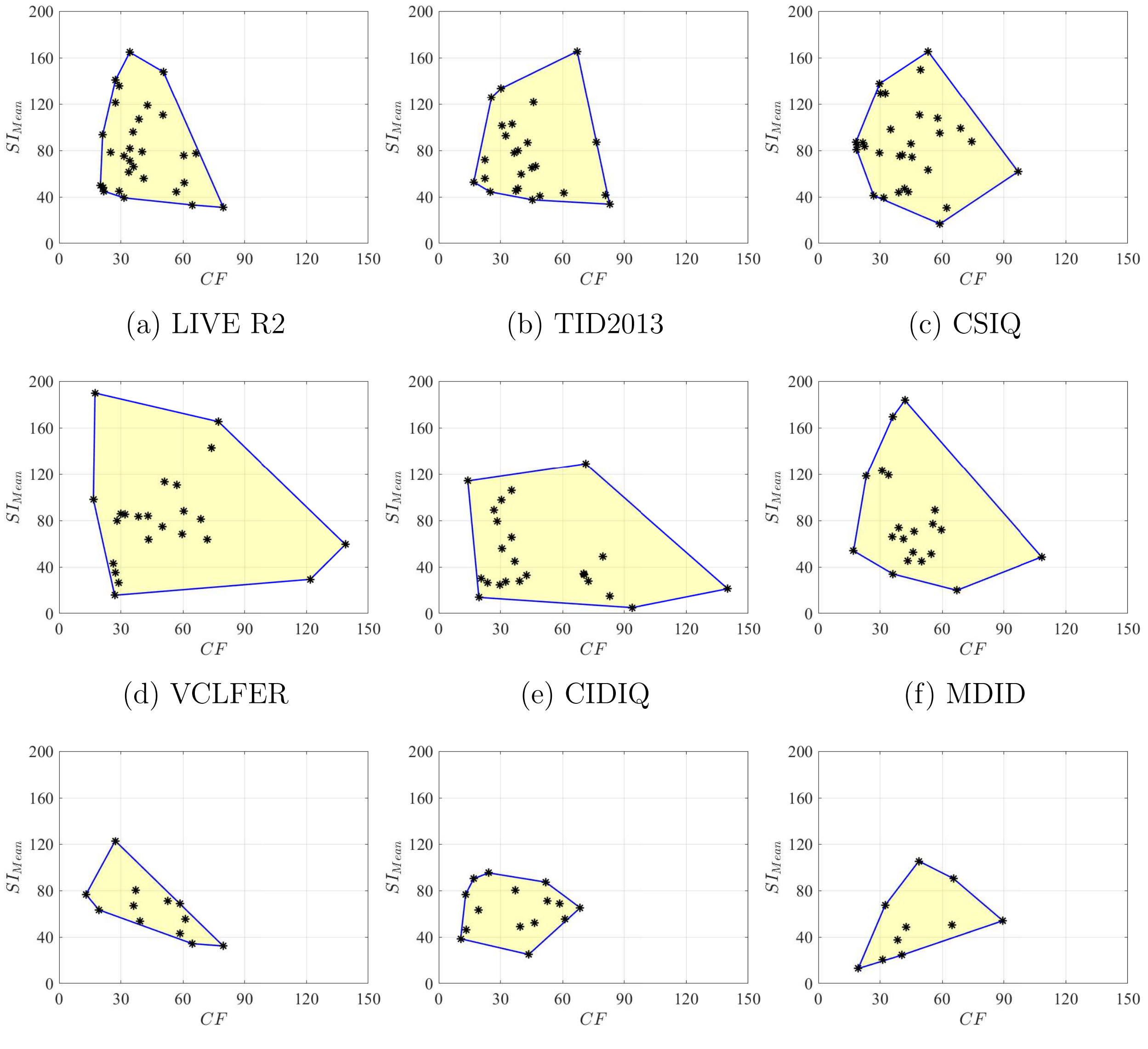
$$yb = \frac{1}{2}(R + G) - B \tag{2.4}$$

Then CF is defined as:

$$\mathrm{CF} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3 \cdot \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{2.5}$$

where σ_{rg} and σ_{yb} are the standard deviations, while μ_{rg} and μ_{yb} are the mean values, in the rg and yb directions respectively.

We computed the SI and CF values of all reference images in the nine IQA databases by using the definitions given in (2.2) and (2.5) respectively. The SI versus CF plots for these databases are given in Fig. 2.1 where the blue outer boundary marks the convex hull in each case and the area inside is marked yellow. For convenience, we have used the same scale for each axis in all the plots of Fig. 2.1. It is evident that the source content in these datasets occupies different regions in the SI versus CF space. While the VCLFER [54] and CIDIQ [5] datasets seem to cover the most area in this space, a majority



(a) LIVE R2

(b) TID2013

(c) CSIQ

(d) VCLFER

(e) CIDIQ

(f) MDID

(g) MDID2013

(h) LIVE MD

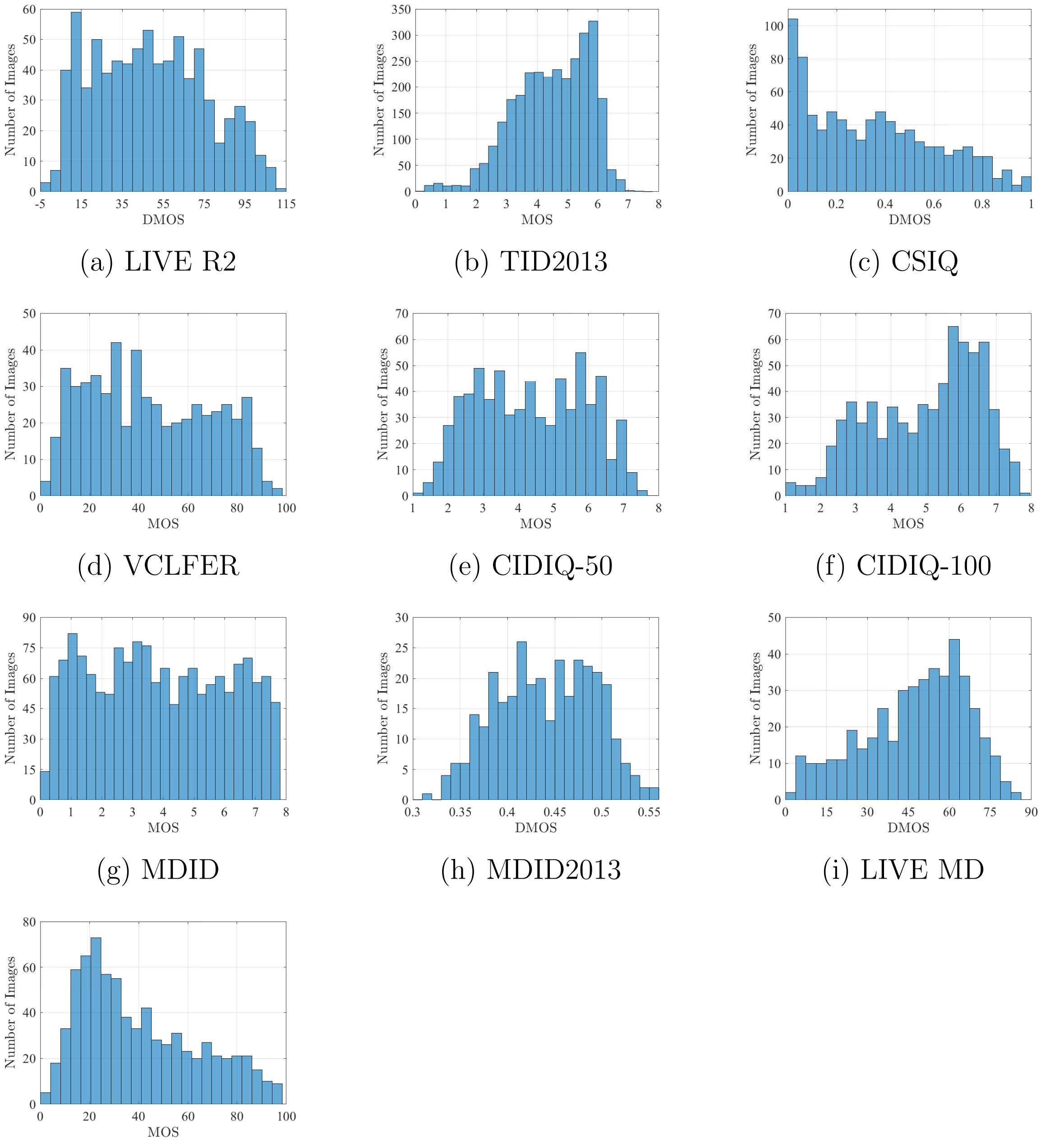
(i) MDIVL

Figure 2.1: Spatial Information (SI_{Mean}) versus Colorfulness (CF) plots of the reference images belonging to the nine databases being used for method performance evaluation in this work. The blue lines represent the convex hull in each case.

of their images are clustered in smaller regions. On the other hand, the content in the LIVE R2 [24] and CSIQ [26] datasets is more uniformly distributed inside their respective convex hulls. Among the multiply distorted datasets, MDID [33] appears to have a wider coverage region while the other datasets in this category seem to have a limited range of source content. Apart from such subjective analysis of the SI versus CF coverage of datasets, efforts have been made to quantify this coverage as well. A 2D criteria called the relative total coverage (RTC) was defined in [88] as the square root of the area of the convex hull of all points in the normalized SI versus CF space. One drawback of using RTC as a coverage metric is that it does not take into account empty spaces within the convex hull. Thus, a single image that is located further away from the rest of the content in the SI versus CF space can lead to elevated RTC values giving a false sense of better coverage. To address this issue, another metric called total effective coverage (TEF) was proposed in [91] which builds upon the RTC concept. TEF introduces a fill rate factor to weigh the RTC value obtained for a dataset. A circle of certain radius r is considered around each image point in the SI versus CF space, within which a *presence* parameter p is considered as 1. The fill rate factor is then determined as a ratio of the area inside the convex hull where $p = 1$ to the area of the entire convex hull. By using a hypothetical database, it is demonstrated in [91] that TEF is a more effective coverage metric than RTC. Apart from the MDID2013 dataset, the RTC and TEF analysis for the eight other datasets can be found in [91] (it should be noted that the root mean square definition of SI is used in [91]).

2.2.5 Distortion Analysis

In addition to wide content coverage, another important property of an ideal IQA database is diversity in terms of distortion types and levels. For a complete list of the types of distortions included in the nine IQA databases under consideration refer to Table 2.2, where this information is provided along with the number of images in each distortion type. While creating distorted content, the goal should be to simulate varying degrees of distortions such that the perceptual quality scale is uniformly sampled. This will ensure that objective IQA methods are tested across the quality spectrum. To accomplish this, IQA databases include different intensity levels for each distortion type. This information



(a) LIVE R2

(b) TID2013

(c) CSIQ

(d) VCLFER

(e) CIDIQ-50

(f) CIDIQ-100

(g) MDID

(h) MDID2013

(i) LIVE MD

(j) MDIVL

Figure 2.2: Histograms of MOS/DMOS of the nine IQA databases being used for method performance evaluation in this work. Note: The MOS of CIDIQ database has been obtained at two viewing distances of 50 cm and 100 cm [5].

is provided in Sections 2.2.1 and 2.2.2 for the datasets under consideration. To ascertain the range of distortions in each database, the histograms of their subjective ratings (MOS or DMOS) are plotted in Fig. 2.2. A higher MOS value represents better visual quality while the opposite is true for DMOS where lower values signify better visual quality. The distribution of distorted content across the quality spectrum can be regarded as relatively uniform in MDID database [33] and mildly uniform in LIVE R2 [24], VCLFER [54], and CIDIQ (at viewing distance of 50 cm) [5] databases. On the other hand, TID2013 [19] and CSIQ [26] databases contain a relatively larger amount of better quality content while LIVE MD [31] and MDIVL [34] databases contain relatively more low quality content. It has been shown that objective IQA methods find it more difficult to evaluate better quality images as compared to low quality ones [19]. Thus, a dataset with a higher proportion of low quality content may not be as challenging as one with more better quality content. The impact of viewing distance on perceptual quality can be observed while comparing the MOS histogram of the CIDIQ database at a viewing distance of 50 cm (Fig. 2.2 e) with the one obtained at 100 cm (Fig. 2.2 f). While the distorted content remains the same in both cases, the presence of more higher quality ratings in the latter case demonstrates the challenge that objective IQA methods need to overcome and also highlights the importance of IQA databases which provide ratings at different viewing distances.

The non-uniform distribution of distorted content in most databases can be attributed to the way in which distortions are simulated. In all datasets being considered here, fixed parameters for each distortion type are used to simulate different levels of distortions across the dataset. While convenient, such an approach does not take into account the nature of source content and the masking effect that it can have upon different distortions. For example, the same compression ratio may lead to very different results when applied to images with different spatial information levels and the same amount of noise may appear quite different when applied to images that differ in texture characteristics. Thus, a reasonable alternative method is to simulate distortions in a content adaptive manner, that is, content specific distortion parameters should be found for each constituent reference image that roughly correspond to predefined perceptual quality levels (we use such a method later in this thesis). Nevertheless, in the current context, the variation of distorted content across the quality spectrum for different datasets provides one more reason to use as many

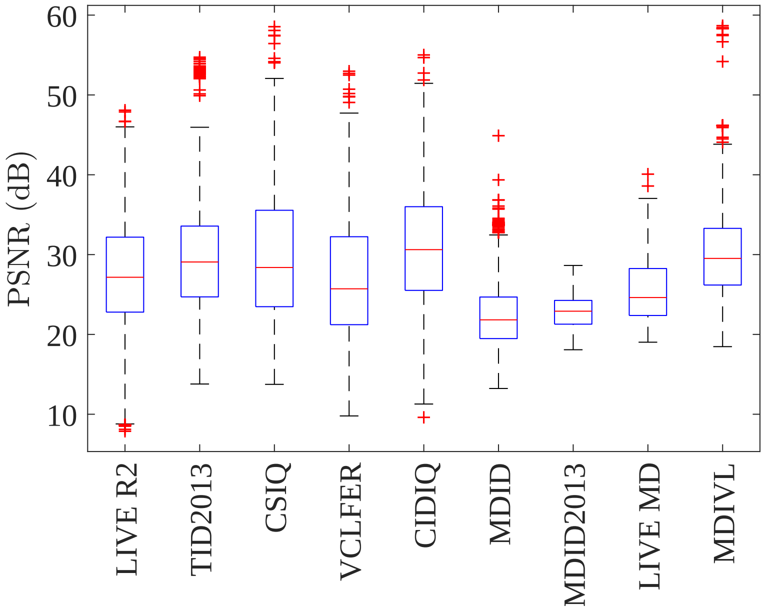


Figure 2.3: PSNR box plots for all databases. The top and bottom edges of the blue boxes represent the 75th and 25th percentiles, respectively, while the red line represents the median (50th percentile). The top and bottom black lines (whiskers) represent the extreme data points while the outliers are represented by red + symbols.

databases as possible in the performance analysis of objective IQA methods.

While the histograms in Fig. 2.2 allow for observing the distribution of distorted content within each database, it is difficult to compare one dataset with another because they use different quality scales and subjective testing methods. To provide a unified, albeit weak [88], basis for comparing different datasets with each other, we compute the peak signal-to-noise ratio (PSNR) of all distorted images in each dataset and provide the corresponding boxplots in Fig. 2.3, where the range of distortions in different datasets can be compared. It can be observed that single distortion databases offer a wider range of distortion intensities while this range is quite limited in multiple distortion datasets. However, this comparison is weak because: 1) PSNR is not a perceptual metric [9], and 2) Even if the individual distortion intensities are wide-ranging in multiple distortion datasets, the interaction of one distortion with another may diminish the effect of the overall distortion, for example, JPEG compression of noisy images may have a denoising effect. The opposite is also true, and thus more research is needed to understand how multiple distortions interact with each other and with image content (we explore this area later in this thesis).

2.3 Review of IQA Algorithms

Our focus in this work is to evaluate representative FR and NR IQA methods, designed for 2D natural images. We will also evaluate fusion based methods where the aim is to achieve better performance by combining results from multiple FR methods. We will provide a brief description of the design philosophies of the methods under consideration. As mentioned earlier, we have not evaluated the performance of RR and other types of IQA methods [92] in this work.

2.3.1 Full-Reference Image Quality Assessment

Full-Reference (FR) IQA methods evaluate the quality of a distorted image with respect to the corresponding original (reference) image that is assumed to be distortion-free and of pristine quality [11]. In this work we evaluate the performance of 43 FR IQA methods which are listed in Table 2.3 along with information about whether a method operates on color or grayscale images, year of publication, and the number and names of the IQA databases that it was tested on. Although this list is not exhaustive, it is representative of different IQA design philosophies. The FR IQA methods being considered are reviewed next and are classified based on their design philosophies.

Error Based Methods

Historically, the *mean squared error* (MSE) and the related *peak signal-to-noise ratio* (PSNR) have been used as the "standard" quality measures [11]. Let $\mathbf{X} = \{x_i|i = 1, 2, ..., N\}$ and $\mathbf{Y} = \{y_i|i = 1, 2, ..., N\}$ represent the reference and distorted images respectively, where x_i and y_i represent the intensities of the i-th samples in the images $\mathbf{X}$

Table 2.3: Information about 43 FR IQA methods under performance evaluation.

FR Method	Color/ Gray	Year	No. of Test Databases	Single Distortion Test Databases Used	Multiple Distortion Test Databases Used
AD_DWT [93]	Gray	2013	3	IVC LIVE R2 TID2008	None
ADM [94]	Gray	2011	5	CSIQ IVC LIVE R2 MICT TID2008	None
CID_MS [95] CID_SS [95]	Color	2013	2	TID2008 Images from six Gamut Mapping Datasets (See references [11], [41]-[44] of [95])	None
DSS [16]	Gray	2015	3	CSIQ LIVE R2 TID2008	None
DVICOM [96] DVICOM_F [96]	Gray	2018	3	CSIQ LIVE R2 TID2008	None
DWT_VIF [97]	Gray	2010	1	LIVE R2	None
ESSIM [98]	Gray	2013	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
FSIM [14]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
FSIMc [14]	Color	2011	5	CSIQ IVC LIVE R2 MICT TID2008	None
GMSD [99]	Gray	2014	3	CSIQ LIVE R2 TID2008	None
GSIM [100]	Gray	2012	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
IFC [101]	Gray	2005	1	LIVE R2	None
IW_PSNR [13]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
IWSSIM [13]	Gray	2011	6	A57 CSIQ IVC LIVE R2 MICT TID2008	None
MAD [26]	Gray	2010	4	CSIQ LIVE R2 MICT TID2008	None
MCSD [102]	Gray	2016	6	CSIQ IVC LIVE R2 MICT TID2008 TID2013	None
MSSSIM [4]	Gray	2003	1	Earlier Version of LIVE R2	None
NQM [103]	Gray	2000	—	Barbara, Boats, Lena, Mandrill, Peppers images	None
PSNR	Gray	—	—	Legacy Method	
PSNR_DWT [93]	Gray	2013	3	IVC LIVE R2 TID2008	None
PSNR_HAc [104]	Color	2011	1	TID2008	None
PSNR_HA [104]	Gray	2011	1	TID2008	None
PSNR_HMAc [104]	Color	2011	1	TID2008	None
PSNR_HMA [104]	Gray	2011	1	TID2008	None
PSNR_HVS [105]	Gray	2006	—	Barbara, Lena images	None
PSNR_HVSM [106]	Gray	2007	—	Test set composed of 19 images	None
QASD [107]	Color	2016	5	CSIQ IVC LIVE R2 TID2008 TID2013	None
RFSIM [108]	Gray	2010	1	TID2008	None
SFF [109]	Color	2013	5	CSIQ IVC LIVE R2 MICT TID2008	None
SNR	Gray	—	—	Legacy Method	
SRSIM [110]	Gray	2012	3	CSIQ LIVE R2 TID2008	None
SSIM [111]	Gray	2004	1	Earlier Version of LIVE R2	None
SSIM_DWT [93]	Gray	2013	3	IVC LIVE R2 TID2008	None
UQI [112]	Gray	2002	—	Lena image	None
VIF [113]	Gray	2006	1	LIVE R2	None
VIF_DWT [93]	Gray	2013	3	IVC LIVE R2 TID2008	None
VIF_P [113,114]	Gray	2005	—	Faster version of VIF, not tested in original paper [113]	None
VSI [15]	Color	2014	4	CSIQ LIVE R2 TID2008 TID2013	None
VSNR [75]	Gray	2007	1	LIVE R2	None
WSNR [103]	Gray	2000	—	Barbara, Boats, Lena, Mandrill, Peppers images	None
WSSI [115]	Gray	2009	1	LIVE R2	None

and $\mathbf{Y}$, respectively, and N is the number of image samples (pixels). MSE and PSNR are defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{2.6}$$

$$\text{PSNR} = 10\ log_{10}\ \frac{L^2}{\text{MSE}} \tag{2.7}$$

where L is the dynamic range of image pixel intensities. For gray-scale images with a bit depth of 8 bits/pixel, $L = 2^8 - 1 = 255$. The PSNR is similar to the Signal-to-Noise Ratio (SNR) which is defined as:

$$\text{SNR} = 10\ log_{10}\ \frac{\frac{1}{N}\sum_{i=1}^{N} x_i^2}{\text{MSE}} \tag{2.8}$$

The MSE has certain advantages [9] such as ease of use, clear physical meaning since it is the energy of the error signal and thus satisfying the Parseval's theorem, and ability to be used for algorithm optimization leading to closed-form solutions, etc. However, it has been repeatedly shown that MSE and PSNR have poor correlation with perceptual image quality, i.e., relative to subjective quality assessment by humans. This is because MSE-type of measures make the following underlying assumptions about perceptual image (and video) quality [9]: 1) It is independent of any spatial and temporal relationships between samples, 2) It is independent of the relationships between the image (and video) signals and error signals, 3) It is determined by the magnitude of the error signal only but ignoring the signs of errors, and 4) All signal samples are of equal importance. Unfortunately, not even one of these assumptions hold in the context of perceptual image (and video) quality assessment [9, 11]. It was also shown in [9] that images along the equal-MSE hypersphere have drastically different perceptual quality. Thus, the advantages of using signal-to-noise ratio based methods are negated by their shortcomings in the context of perceptual quality assessment.

To address the shortcomings of PSNR and SNR, several efforts have been made to modify these methods in order to make them perceptually better suited for IQA. In [103] the contrast sensitivity function (CSF), which is used to approximate the behavior of the human visual system (HVS), was used to weigh the signal and noise powers, leading to a linear quality measure called Weighted Signal-to-Noise Ratio (WSNR). The Noise Quality Measure (NQM) was also presented in [103] and uses a nonlinear quasi-local processing model of the HVS to accomplish quality assessment. An HVS based version of PSNR, called PSNR-HVS, was proposed in [105] which uses the CSF. PSNR-HVS was modified by incorporating a model that takes into account the between-coefficient contrast masking of discrete cosine transform (DCT) basis functions leading to a new method called PSNR-HVSM [106]. PSNR-HVS and PSNR-HVSM were further modified by incorporating human perception of contrast and mean brightness distortions, leading to modified methods called PSNR-HA and PSNR-HMA respectively [104]. To deal with color images, PSNR-HA and PSNR-HMA were applied separately to each component of YCbCr transformed images and the results were combined into a quality score, leading to PSNR-HAc and PSNR-HMAc respectively [104]. The Visual Signal-to-Noise Ratio (VSNR) [75] is another HVS based method, which uses wavelet based models of visual masking and visual summation to first ascertain if the distortions are beyond contrast thresholds of detection, in which case they are deemed visible. For suprathreshold distortions, low-level and mid-level visual properties of perceived contrast and global precedence respectively, are modeled as Euclidean distances in the distortion-contrast space of a multiscale wavelet decomposition. VSNR is then calculated as the ratio of the RMS contrast of the pristine reference image to the weighted sum of the two Euclidean distances. An information content weighted version of PSNR, called IW-PSNR is proposed in [13], where the underlying premise is that some regions of visual content are perceptually more important than others, either due to the visual attention property of the HVS or due to the influence of distortions [116, 117]. IW-PSNR uses information theoretic principles to compute information content weights which are used in the pooling stage of quality score generation. In [93] a Haar wavelet based discrete wavelet transform (DWT) framework is developed to compute image quality methods in the DWT domain. Image quality methods are separately applied to the approximation subbands and edge-maps obtained from detail subbands, leading to approximation and edge quality scores which are linearly combined to yield the final quality scores. Of the four developed methods in [93], two are error-based methods and include PSNR-DWT and absolute difference based AD-DWT.

Structural Similarity Based Methods

It can be seen from the previous section that HVS characteristics have been used to modify error based methods such as the MSE and PSNR. This is essentially a *bottom-up* approach to IQA design since the functionality of different HVS components is being simulated. By contrast, the *top-down* approach to IQA design does not try to model the functionality of individual HVS components. Instead, it tries to mimic the functionality of HVS as a whole [11]. The last two decades have seen the advent of a number of successful IQA methods that follow the *top-down* approach, some of which will be briefly explained in this and subsequent sub-sections.

One of the most well-known FR methods following the *top-down* approach is the Structural Similarity (SSIM) index [111], which is a modified version of the Universal image Quality Index (UQI) [112], and is based on the assumption that the HVS is adapted for extracting structural information from visual content. SSIM operates in the spatial domain and performs three types of comparisons between the reference and distorted images: luminance, contrast and structure. Luminance comparison is a function of mean intensity of the images being compared, while contrast comparison is a function of standard deviations. Structural comparison is done through correlation between the image patches being compared after mean subtraction and variance normalization. All comparisons are done locally by a sliding window and the three SSIM components are combined, leading to local quality scores, which together lead to a quality map. The overall quality score for the entire distorted image with respect to the reference image is obtained by taking the mean of all the local quality scores. The SSIM index is a single-scale approach, that is, it can take into account only one set of viewing conditions. To account for the variations in viewing conditions, a multi-scale version of SSIM called Multi-scale Structural Similarity (MSSSIM) was developed in [4] and uses 5 scales. Images at different scales are obtained by downsampling the images at the previous scale by a factor of 2. The contrast and structural comparisons are performed at all scales, while the luminance comparison takes place only at the final scale. The quality scores obtained at each scale are combined through a weighted product, where the weights assigned to different scales are obtained through an image synthesis calibration experiment that involved subjective testing. To generate final quality scores, both SSIM and MSSSIM use mean pooling, which assigns equal importance to all areas of visual content. As discussed earlier, some regions of visual content are perceptually more important, either because of the visual attention property of the HVS or due to the influence of distortions [116, 117]. In [13], a modified version of MSSSIM, called Information content Weighted Structural Similarity (IWSSIM) was presented. IWSSIM operates at 5 scales and uses information theoretic principles to compute information content weights that are used in the pooling stage. A wavelet domain implementation of SSIM, called Wavelet Structural Similarity Index (WSSI) was proposed in [115] which uses the Haar wavelet for image decomposition. In WSSI, edge-maps are obtained from detail subbands followed by the generation of approximation and edge structural similarity maps. A contrast map is used to pool together the different wavelet domain structural similarity maps leading to approximation and edge similarity scores which are then linearly combined into the final WSSI quality score. Another SSIM based wavelet domain method called the SSIM-DWT was developed in [93] and uses the same design philosophy as WSSI.

Besides SSIM and methods that are directly based on it, several other FR IQA methods have been proposed that utilize the SSIM design philosophy. The Riesz-transform based Feature Similarity (RFSIM) index was proposed in [108]. RFSIM uses first and second order Riesz Transform coefficients as features and compares them only at key locations identified by an edge-based feature mask which is obtained by using the Canny edge detection operator without thinning. The final RFSIM quality score is obtained as a product of similarity scores of individual feature maps. The Feature Similarity (FSIM) index was proposed in [14] and uses phase congruency as the primary feature to evaluate image similarity. Since phase congruency is contrast invariant, the gradient magnitude is used as a secondary feature in FSIM to capture contrast information. The phase congruency and gradient magnitude maps of the reference and distorted images are compared leading to phase congruency and gradient magnitude quality maps which are then combined into a single quality map for the luminance channel of the images through a weighted product. The final FSIM quality score is obtained by pooling this quality map by using a weighting function that is derived from the phase congruency maps of the images being compared.

A color version of FSIM, called FSIMc, has also been proposed in [14]. RGB color versions of the images being compared are first converted to the YIQ color space [118]. Phase congruence and gradient magnitude based comparisons are performed on the luminance channel Y, as in FSIM, leading to the luminance similarity map. Additionally, the I and Q chromatic channels are compared leading to I and Q similarity maps whose product leads to a chrominance similarity map. The luminance and chrominance similarity maps are pooled into the final FSIMc score by using the phase congruence based weighting function. The Spectral Residual based Similarity (SRSIM) index is proposed in [110] and uses the Spectral Residual based Visual Saliency (SRVS) model [119] to perform two functions: 1) SRVS maps act as features to ascertain local quality and 2) A weighting function is derived from the SRVS map to highlight the importance of visual regions when pooling to obtain the final quality score. To account for the lack of contrast sensitivity of SRVS, SRSIM uses gradient magnitude maps of the images being compared as supplementary features. Following a similar design approach as SRSIM, the Visual Saliency-based Index (VSI) is proposed in [15] which is able to handle color images. VSI uses the visual saliency model called Saliency Detection by combining Simple Priors (SDSP) [120] to generate visual saliency maps, which are used as features in local quality estimation and also act as a weighting function during the pooling stage for final quality score generation. It was shown in [15] that visual saliency maps are insensitive to change of contrast and color saturation, which thus requires VSI to include additional features. This is accomplished by first transforming the RGB color images into an opponent color space. Next, gradient magnitude is used as a feature to generate gradient similarity maps in order to make VSI contrast sensitive, while chrominance similarity maps are generated through the two chromatic channels to make VSI color saturation sensitive. An IQA method based on Gradient Similarity (GSIM) is proposed in [100], where changes in contrast and structure in images being compared are measured through gradient comparison. It also takes into account masking effects, visibility threshold and luminance distortions. GSIM combines the measurement of luminance distortion and contrast-structure distortion in an adaptive manner to give a final quality score, where more weight is given to the latter. An IQA method based on Edge Strength Similarity (ESSIM) is proposed in [98], where it is assumed that the edge-strength of each pixel fully represents the semantic information of images. Based on the characteristics of the edge in images, ESSIM defines edge-strength to take both anisotropic regularity and irregularity into account. Another FR IQA method based on gradient similarity called Gradient Magnitude Similarity Deviation (GMSD) is proposed in [99]. While GMSD compares the gradient magnitude maps of the reference and distorted images to compute a local quality map, it uses standard deviation as the pooling strategy to generate the final quality score from the local quality map. The underlying premise is that the global variation of local image quality is an indicator of overall image quality. Following the design philosophy of GMSD, an FR IQA method called Multiscale Contrast Similarity Deviation (MCSD) is proposed in [102]. First, the pristine reference and distorted images are downsampled by a factor of 2 and a contrast similarity map for the images being compared is computed by using their respective contrast maps. Next, standard deviation is used as a pooling strategy to generate a contrast similarity deviation (CSD) quality score from the contrast similarity map. To incorporate the effect of viewing distance, this process is repeated at two further scales by downsampling by a factor of 2 each time and computing the CSD at each scale. The product of the three CSD scores gives the final MCSD quality score. A discrete cosine transform (DCT) domain FR IQA method called the DCT Subbands Similarity (DSS) is proposed in [16]. DSS measures the amount of local change of respective subband coefficients by comparing the local variance and generates a quality score for each subband. A final DSS quality score is obtained by combining the individual subband scores such that more weight is given to subbands corresponding to lower spatial frequencies in accordance with the characteristics of the HVS. The Color-Image-Difference measure (CID) [95] is an FR IQA method for color images. CID uses an image-appearance model to normalize the images being compared and transforms them to a working color space. It then extracts features from both the reference and distorted images, which are compared for similarity. Feature comparisons include lightness, chroma, hue, contrast and structure comparisons. A multiscale approach similar to MSSSIM [4] is used for contrast and structure comparisons. Lightness comparison is made on the smallest scale. A factorial combination model is finally used to combine the scores from different feature comparisons into a single CID quality score.

Natural Scene Statistics based Methods

IQA methods belonging to this paradigm regard natural images as entities with certain statistical properties which can be defined in terms of representative models and that are effected due to distortions [12]. Statistical models of the reference and distorted images are compared using principles of information theory, thereby providing an opportunity for quality assessment. Early works apply this idea to RR IQA [121, 122], where the original reference image is not fully available, but certain statistical features (in this case natural scene statistics features) are extracted and compared with those extracted from the test image to yield a quality evaluation. The idea was later extended for FR IQA.

A well-known FR IQA method following this design approach is the Information Fidelity Criterion (IFC) that was proposed in [101]. IFC treats IQA as an information fidelity problem where the reference image from the natural image source is being communicated to a receiver who is a human observer, through a channel which is the distortion process. Here, the reference and distorted images are the input and output of the channel respectively. IFC uses a Natural Scene Statistics (NSS) [123] based Gaussian Scale Mixtures (GSM) model [124] in the wavelet domain to represent the source where the steerable pyramid decomposition [125] with six orientations is used. The distortion model is obtained by attenuating the source model and adding Gaussian noise to it. The task of image fidelity measurement is then accomplished by determining the mutual information between respective wavelet subbands of the reference and distorted images represented through the source and distortion models respectively. The final IFC fidelity or quality score is obtained by summing the mutual information for all subbands. Using the IFC as a base, the FR IQA method called Visual Information Fidelity (VIF) was proposed in [113]. Like IFC, the VIF uses a NSS [123] based GSM model [124] in the wavelet domain to model the source and uses the same steerable pyramid decomposition [125]. VIF also uses a similar distortion model as the IFC. However, VIF introduces an HVS model in the wavelet domain to incorporate the uncertainty that is introduced by the HVS channel as it processes the visual signal. VIF models the HVS channel through a stationary, zero mean, additive white Gaussian noise model. VIF then defines two types of information: 1) The reference image information represents the information in the reference image and is defined as the mutual information between the input and output of the HVS channel without the distortion channel. 2) The test image information is the information in the distorted image and is defined as the mutual information between the input of the distortion channel and the output of the HVS channel, where these two channels are in series (distortion channel followed by the HVS channel). VIF is then defined as the ratio of the test image information to the reference image information (for all subbands). The designers of VIF [113] provide a pixel domain version of VIF, called VIFP which is computationally simpler. Although the implementation details of VIFP have not been provided in [113], some information and its implementation code can be found at [114]. While VIF [113] uses a vector GSM implementation, VIFP [114] uses a scalar GSM implementation and is multi-scale in nature. A low-complexity wavelet-domain version of VIF, called the DWT-VIF has been proposed in [97]. To reduce the computational complexity, DWT-VIF adopts a one-level decomposition using the Haar wavelet instead of the over-complete steerable pyramid decomposition as in VIF. This allowed the use of a scalar GSM model in DWT-VIF instead of the vector GSM model that was required in VIF. DWT-VIF computes quality scores separately between approximation subbands and edge maps extracted from the detail subbands of the reference and distorted images being compared. A linear combination of the approximation and edge similarity scores gives the final DWT-VIF quality score. The designers of DWT-VIF [97] provide a similar method called VIF-DWT in [93].

Since natural images are known to possess sparse structures, sparsity based approaches to IQA can also be placed under the NSS category. A sparse coding based FR IQA method for color images called Sparse Feature Fidelity (SFF) is proposed in [109]. SFF computes the fidelity of the distorted image with respect to the reference image by using two subtasks, feature similarity and luminance correlation. A universal feature detector is trained once on a set of natural images using independent component analysis (ICA) and then used to transform a given image into a sparse coefficient vector. The reference and distorted images are first split into corresponding patches and only those patches are selected for further processing which display suprathreshold distortions. Next, the feature detector is applied to the selected reference and distorted image patches to extract sparse feature vectors. The feature vectors of the reference image are used to determine a visual threshold to identify visually important patches. This process of patch and feature vector selection is done to incorporate the HVS properties of visual attention and visual thresholding [116, 117]. Once features have been selected from the reference and distorted images, similarity between them is determined. Separately, correlation between the mean values of selected image patches from the reference and distorted images is used to represent luminance correlation. Finally, the feature similarity and luminance correlation values are linearly combined to yield the final SFF quality score. Another sparsity based FR IQA method for color images called sparse representation based image Quality index with Adaptive Sub-Dictionaries (QASD) has recently been proposed in [107]. QASD utilizes a universal overcomplete dictionary, which is trained by using natural images, to extract sparse features which are the primary features being used for quality assessment. First, QASD utilizes the universal overcomplete dictionary to extract sparse coefficients from blocks of the reference image. Next, it adaptively forms sub-dictionaries for respective image blocks by using only the basis vectors obtained in the sparse representation of the reference image. The sparse representation of the distorted image blocks is then obtained only by using the respective sub-dictionaries. This ensures that the same set of basis vectors are used in the sparse representation of both the reference and distorted images, therefore ensuring meaningful comparison for IQA. Using the sparse representations, feature maps are generated for the reference and distorted images which are then compared for similarity. It is mentioned in [107] that weak distortions have limited influence on sparse representations, therefore, supplementary features are employed to capture the effect of such distortions. Three supplementary features are used which include image gradient, color and luminance. The RGB color image is first converted to the YCbCr color space, which is followed by image gradient similarity computation in the Y channel and color similarity computation in the chroma channels. Luminance similarity is determined as in SFF [109]. The sparse feature maps are used to generate a weighting map, which is used in the weighted pooling of the sparse feature similarity map, gradient similarity map, and chroma similarity map. The final QASD quality score is obtained as a weighted product of the various similarity maps.

Mixed Strategy based Methods

Some other design philosophies have also been used for the task of FR IQA, which use an overlap of different strategies. The Most Apparent Distortion (MAD) [26] method, assumes that the HVS adopts two different strategies to determine image quality: 1) For high quality images with only near-threshold distortions, MAD uses a detection based strategy. A spatial domain local visual mask, based on the CSF, luminance and contrast masking, is used to find regions in which the near-threshold distortions are considered as visible. Image quality of the distorted image with respect to the reference is then estimated in the identified regions through the mean squared error. 2) For low quality images with clearly suprathreshold distortions, MAD uses an appearance based strategy. A log-Gabor filter bank is used to decompose the reference and distorted images into coefficients, with greater weight given to coarser scales. Image quality is determined as the absolute difference between low level statistics including the mean, variance, skewness and kurtosis, of the weighted coefficients. Based on the amount of distortion, the detection and appearance quality scores are then combined through a weighted geometric mean to give the final MAD score.

An FR IQA method (ADM) was proposed in [94] which uses a wavelet domain decoupling algorithm for impairment separation and then evaluates detail losses and additive impairments. It simulates the HVS by incorporating the CSF and contrast masking characteristics of the HVS. Detail loss, defined as the loss of useful visual information, is computed after the decoupling process as the ratio of the Minkowski sum of the restored image to that of the original image. Additive impairment, defined as redundant visual information due to the influence of distortions, is computed as the Minkowski sum of the additive impairment image obtained after the decoupling process. The detail loss and additive impairment quality scores are then adaptively combined such that more weight is given to the detail loss based score for low quality images.

The Detail Virtual Cognitive Model (DVICOM) [96] combines two separate metrics that measure the perceptual impact of detail losses and spurious details. Using the images being compared and Least Squares decomposition, DVICOM breaks down the gradient field of the distorted image into two components, a prediction of the gradient field of the original image and an unpredictable gradient residual. Detail loss is then determined by the attenuation of the predicted gradient, measured through the loss of positional information. The gradient residual is used to measure the spurious detail component as the ratio of the original gradient energy and the residual gradient energy. These two components are considered as coordinates of a 2D space and mapping is done to a DMOS estimate by using a parametric function that has been trained on experimental data. In addition to the standard version of DVICOM, a computationally faster version has also been provided by its inventors, which we refer to as DVICOM_F.

2.3.2 FR Fusion based Image Quality Assessment

It is evident from Sections 2.4 and 2.5 that state-of-the-art FR IQA methods achieve good correlation with human perception of quality (where the weighted average SRCC of top performing FR methods is around 0.86 on nine subject-rated databases), while there is significant room for improvement in the performance of general-purpose NR IQA methods (where the top performing NR method has a weighted average SRCC of around 0.61 on the same set of data). However, it has been observed in the past [73] and we shall demonstrate later in this thesis as well that the performance of state-of-the-art FR methods fluctuates across different IQA databases that have different sets of distortions. The question is: How to achieve objective IQA that has stable, robust, and perceptually well-correlated performance across different distortion types? Researchers have tried to answer this question by combining or fusing the results from different FR IQA methods together, in the hope that the deficiencies of one method will be covered by another method in the combination set. Such FR fusion methods can be classified into three categories: 1) Empirical fusion methods, 2) Learning based fusion methods, and 3) Rank aggregation based fusion methods. In this work we evaluate the performance of seven FR fusion based methods which are listed in Table 2.4 along with information about whether they operate on grayscale or color images, year of publication, and number and names of the IQA databases that they were tested on. A brief description of these methods and their categories follows.

Empirical Fusion

In this rather simple approach, the results from two or more FR IQA methods are combined through a weighted product procedure. The weights assigned to different FR methods are obtained by optimizing on some subject-rated database.

Table 2.4: Information about seven FR Fusion based methods under performance evaluation.

FR Fusion Method	Color/ Gray	Year	No. of Test Databases	Single Distortion Test Databases Used							Multiple Distortion Test Databases Used
CISI [126]	Color	2012	7	A57	CSIQ	IVC	LIVE R2	MICT	TID2008	WIQ	None
CM3 [127]	Gray	2014	1	None							LIVE MD
CM4 [127]	Gray	2014	1	None							LIVE MD
CNNM [128]	Color	2015	1	TID2013							None
HFSIMc [129]	Color	2012	7	A57	CSIQ	IVC	LIVE R2	MICT	TID2008	WIQ	None
MMF [130]	Gray/ Color	2013	6	A57	CSIQ	IVC	LIVE R2	MICT	TID2008		None
RAS [41]	Gray/ Color	2014	3	CSIQ	LIVE R2	TID2008					None

The Hybrid Feature Similarity (HFSIMc) index [129] combines results from two feature similarity based FR methods, FSIMc [14] and RFSIM [108], in the following manner:

$$\text{HFSIMc} = (\text{RFSIM})^a \cdot (\text{FSIMc})^b \tag{2.9}$$

where the exponent values of $a = 0.4$ and $b = 3.5$ have been optimized on the TID2008 database [25].

The Combined Image Similarity Index (CISI) [126] combines results from three FR methods, FSIMc [14], MSSSIM [4] and VIF [113], as follows:

$$\text{CISI} = (\text{MSSSIM})^a \cdot (\text{VIF})^b \cdot (\text{FSIMc})^c \tag{2.10}$$

where the exponent values of $a = 0.5$, $b = 0.3$, and $c = 5$, have been optimized on the TID2008 database [25].

Two combined metrics designed for multiply distorted images are proposed in [127]. They are called CM3 and CM4, and are respectively defined as:

$$\text{CM3} = (\text{IFC})^{0.34} \cdot (\text{NQM})^{2.4} \cdot (\text{VSNR})^{-0.3} \tag{2.11}$$

$$\text{CM4} = (\text{IFC})^{0.2} \cdot (\text{NQM})^{2.9} \cdot (\text{VSNR})^{-0.54} \cdot (\text{VIF})^{0.5} \tag{2.12}$$

where IFC [101], NQM [103], VSNR [75], and VIF [113] are FR methods, and the exponent values have been optimized on the LIVE MD database [31].

Although some other FR fusion based methods that follow the weighted product approach have been proposed, such as the CQM [131] and the EHIS [132], we will use the above-mentioned four methods as representatives of this category.

Learning based Fusion

A general-purpose learning based FR fusion approach called Multi-Method Fusion (MMF) was first proposed in [133] and then further refined in [130]. Given an annotated training dataset, MMF selects a subset of FR IQA methods from a larger pool, and then uses support vector regression (SVR) to learn a model that is a non-linear combination of the methods being fused. Defining similar distortion types as a *context*, two kinds of fusion methods are constructed: 1) Context-Free (CF) MMF is independent of distortion type where regression is done at the level of the entire training set. 2) Context-Dependent (CD) MMF takes distortion type into account and performs regression within each group of similar distortions. In the published version of MMF [130], the pool of FR IQA methods is composed of 10 methods which include: MSSSIM [4], SSIM [111], VIF [113], VSNR [75], NQM [103], PSNR-HVS [105], IFC [101], PSNR, FSIM [14], and MAD [26]. However, it is noted that any other FR method pool can be used for MMF construction. To ensure a level playing field, scores from different FR methods are linearly rescaled to the range of $[0, 1]$, as per the recommendations in [134], before learning a combination model through SVR. For CD-MMF, Support Vector Machine (SVM) is used to learn a classification algorithm to automatically determine the context of a given image. To accomplish this, the distortions in known IQA databases are divided into five groups based on similarity among distortions and five spatial domain features are used to learn the classification algorithm. With the context determined, FR method fusion is carried out through an SVR based model which may involve a different set of FR methods for each context. To determine the best possible set of FR methods to be fused, for both CF-MMF and CD-MMF exhaustive search becomes infeasible if the FR method pool is large. Two algorithms are proposed in [130] for FR method selection: 1) Sequential Forward Method Selection (SFMS) uses PLCC as the objective function and starts with a single FR method that has the highest PLCC with respect to the training subjective data. It then combines this method with every other FR method in the pool one at a time and trains the MMF model, where the method that gives the highest PLCC is selected as the second FR method. This process is repeated sequentially until all the FR methods in the pool have been exhausted. The number of FR methods being combined is then selected based on computational complexity requirements. 2) Biggest Index Ranking Difference (BIRD) selects FR methods that are most dissimilar to each other in order to have an FR set that works well for a wide variety of distortions. The number of FR methods to be fused for a particular training dataset is determined based on a formula that balances performance and complexity. For example, the fusion count is estimated to be six for the TID2008 database [25] while using the SFMS algorithm, and the following methods are selected: FSIM [14], VIF [113], IFC [101], MAD [26], PSNR-HVS [105], and MSSSIM [4]. This combination will be used later in this work while evaluating the performance of MMF where we have restricted ourselves to CF-MMF. We will also use three other pools for FR method selection, details of which are provided in Section 2.4.3.

A neural networks based general-purpose supervised FR fusion based approach called Combined Neural Network Metric (CNNM) was proposed in [128]. As input, CNNM takes the scores from six FR IQA methods without any pre-processing and gives a combined quality score at its output. In order to select FR methods for fusion, 27 different methods were analyzed on the 24 different types of distortions in the TID2013 database [19]. Based on results from this analysis and the evaluation done in [135], six FR methods were chosen such that they reliably cover the distortions in TID2013 between them. The selected FR methods include VIF [113], PSNR-HVS [105], PSNR-HMAc [104], FSIMc [14], SFF [109], and SRSIM [110]. A 4-layer cascade-forward backprop neural network with 10, 10, and 20, neurons in hidden layers was used with training being done on the TID2013 database [19] using MATLAB. The TID2013 database has 3,000 images, of which 1,500 were used for training while the remainder were used for later analysis. During training itself, MATLAB used 500 of the 1,500 images for training, while 500 were used for validation and 500 for testing.

Rank Aggregation based Fusion

The FR fusion methods discussed above require training with respect to subject-rated databases. The empirical fusion approaches need such databases to optimize exponent values while the learning based fusion approaches need them to learn the combination model. These approaches often suffer from overfitting problems, as will be demonstrated in Section 2.4. On the other hand, a training-free fusion approach could potentially alleviate these issues.

A recently proposed framework called Blind Learning of Image quality using Synthetic Scores (BLISS) [41] replaces human opinion scores with synthetic quality scores that act as ground truth data. Such synthetic quality scores are generated by using a training-free FR fusion method which involves two steps: 1) Generation of consensus ranking through unsupervised rank aggregation, and 2) Score adjustment of a base FR method based on the consensus ranking. Since different FR measures have different score ranges, their outcomes cannot be combined by averaging their values. Instead rank aggregation is used as an alternative. Given a set of test images and their associated scores assigned by a number of FR methods, a consensus ranking is first obtained by using the unsupervised rank aggregation method called Reciprocal Rank Fusion (RRF) [23], which was first developed for combining document rankings from multiple information retrieval systems. The RRF score of an image I_i is defined as [23, 41]:

$$RRF_{score}(I_i) = \sum_{j=1}^{J} \frac{1}{k + r_j(i)} \tag{2.13}$$

where J is the number of FR methods being combined, $r_j(i)$ is the rank given by the j-th FR method to the image I_i, and $k = 60$ is a constant that counters the impact of high rankings by outliers. The value of the constant k was determined through a pilot investigation in [23].

It is mentioned in [41] that RRF values cannot be directly used as quality scores since they indicate the quality of an image relative to other images in the dataset. Instead, a quality measure is obtained by adjusting the scores of a base FR method with respect to the consensus ranking obtained through RRF. While generating the final synthetic quality scores, the mean squared error between the combined scores and the base FR scores is minimized and a penalty is applied when there is an inconsistency with respect to the consensus ranking. The entire process of FR method fusion and synthetic score generation is training-free. In this work we will call the FR fusion approach proposed in [41] as RRF based Adjusted Scores (RAS). In [41], five FR methods are used in fusion which include GMSD [99], VIF [113], FSIM [14], FSIMc [14], and IWSSIM [13]. Two combinations are adopted, where the first fuses all five FR methods while the second one excludes VIF. We shall respectively call them as RAS_B1 and RAS_B2 in this work and will evaluate their performance in addition to several other RAS fusion combinations in Section 2.4.

2.3.3 No-Reference Image Quality Assessment

No-Reference (NR) IQA methods evaluate the quality of a distorted image in the absence of any reference information [11], and thus they are also referred to as *blind* IQA (BIQA) methods. By its very nature, BIQA is a difficult task and early efforts were made towards the design of NR IQA methods for specific distortions, such as for blur [136], JPEG compression [137], JPEG2000 compression [138]. However, with advances in domain-knowledge, technology and with the availability of subject-rated IQA databases, several general-purpose NR methods have been designed in the last decade that work with a number of distortions. Contemporary NR IQA methods are usually classified into two categories [3]: 1) Opinion-Aware (OA) methods which are trained on distorted images whose quality has been rated by human subjects, and 2) Opinion-Unaware (OU) methods (also referred to as Opinion-Free) which do not train on human-rated distorted images. In this work we evaluate the performance of 14 NR IQA methods (8 OA and 6 OU) which are listed in Table 2.5 along with information about whether they operate on grayscale or color images, year of publication, and number and names of the IQA databases that they were tested on. Although this is not an exhaustive list, we selected NR methods for a good representation of various BIQA design philosophies in addition to computational time constraints. A brief description of the NR IQA methods being evaluated in this work is given next.

Table 2.5: Information about 14 NR IQA methods under performance evaluation.

NR Method	Color/ Gray	Year	No. of Test Databases	Single Distortion Test Databases Used	Multiple Distortion Test Databases Used
BIQI [139]	Gray	2010	1	LIVE R2	None
BRISQUE [140]	Gray	2012	2	LIVE R2 TID2008	None
CORNIA [141]	Gray	2012	2	LIVE R2 TID2008	None
dipIQ [36]	Gray	2017	4[b]	CSIQ LIVE R2 TID2013	None
GWHGLBP [142]	Gray	2016	2	None	LIVE MD MDID2013
HOSA [143]	Gray	2016	10[a]	CSIQ LIVE R2 MICT TID2013	LIVE MD
ILNIQE [144]	Color	2015	4	CSIQ LIVE R2 TID2013	LIVE MD
LPSI [145]	Gray	2015	2	LIVE R2 TID2008	None
MEON [146]	Color	2018	4[b]	CSIQ LIVE R2 TID2013	None
NIQE [3]	Gray	2013	1	LIVE R2	None
NRSL [147]	Gray	2016	7[c]	CSIQ LIVE R2 TID2013	LIVE MD
QAC [35]	Gray	2013	3	CSIQ LIVE R2 TID2008	None
SISBLIM [32]	Color	2014	7	CSIQ IVC LIVE R2 MICT TID2008	LIVE MD MDID2013
WaDIQaM-NR [148]	Color	2018	4[d]	CSIQ LIVE R2 TID2013	None

[a] HOSA was also tested on two authentic distortion databases: CID2013 [78], LIVE WC [79]; one database of screen content images: SIQAD [84]; and on two document image databases: Newspaper database [86], Document Image Quality database [85].

[b] dipIQ and MEON are also tested on the Waterloo Exploration database [21] which is a single distortion database that does not have subject rated quality scores.

[c] NRSL was also tested on three authentic distortion databases: BID [77], CID2013 [78], LIVE WC [79].

[d] WaDIQaM-NR was also tested on one authentic distortion database: LIVE WC [79].

Opinion-Aware NR Methods

OA NR methods can be further classified into two categories based on whether handcrafted or learned features are used.

In the handcrafted features based approach, features that correlate well with image quality, such as NSS based statistical parameters representing the empirical distributions of image coefficients in either the spatial or some transform domain, are extracted from the distorted images. Next, these feature vectors and associated image subjective ratings are used to train a model by using machine learning techniques such as SVR [149]. In the testing phase, the OA NR method extracts features from the test image and uses the learned quality model to map them to a quality score. The Blind Image Quality Index (BIQI) [139] is a pioneering general-purpose NR IQA method based on the premise that different distortions affect the natural scene statistics (NSS) of images in a specific manner. BIQI uses the Daubechies 9/7 wavelet basis [150] to decompose an image into three-scales and three-orientations. The Generalized Gaussian Distribution (GGD) is then used to represent the coefficients of each subband. GGD parameters are estimated using the approach proposed in [151] and form a feature vector to represent the image. BIQI then follows a two-step process to determine image quality. First, the feature vector is used to determine the presence of various distortions. Since BIQI is trained on the LIVE R2 database [24, 42], its published version uses a distortion set of JPEG compression, JPEG2000 compression, white noise, Gaussian blur, and fast fading, since these distortions are present in LIVE R2. In the training phase, BIQI uses SVM [149] to learn the classification model which assigns probability scores to various distortions based on their perceived magnitude. In the second step, the same feature vector is used for quality score assignment for each distortion category. In the training phase SVR [149] is used to learn a regression based model. The final BIQI quality score is then determined as a probability-weighted sum of the quality scores for various distortions. The Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [140] is an NSS based NR method that operates in the spatial domain. BRISQUE operates on locally normalized luminance values which are termed as Mean Subtracted Contrast Normalized (MSCN) coefficients. A benefit of this normalization process is that it leads to relatively decorrelated neighboring coefficients as compared to non-normalized pixel values. NSS features are extracted from the models of the MSCN coefficients and their pairwise products. The GGD is used to fit the empirical MSCN distributions, where the procedure proposed in [151] is used to estimate GGD parameters which form one set of features. The relationships between neighboring pixels are modeled through the pairwise products of neighboring MSCN coefficients along four orientations. The Asymmetric Generalized Gaussian Distribution (AGGD) [152] is used to fit the empirical distributions of these pairwise products and the estimated fitting parameters lead to another set of features. To incorporate multiscale operation, BRISQUE extracts features at two scales. It is shown in [140] that distortions affect these NSS features such that they occupy different regions in the GGD and AGGD parameter spaces, thereby providing an opportunity to learn quality models. BRISQUE uses SVR [153] to learn a model to map features to a quality score and uses the LIVE R2 database [24, 42] for training. The degradation of structural features has been used in the design of OA NR methods, such as the Gradient-Weighted Histogram of Local Binary Pattern calculated on the Gradient map (GWHGLBP) [142] which has been designed for multiply distorted images. First, the gradient map of a distorted image is obtained through the Prewitt filter. Structural information is extracted from the gradient map by applying the Local Binary Pattern (LBP) operator [154] leading to GLBP codes. It is claimed that these codes are affected in unique ways by different distortions, making them effective features for IQA. Contrast information is incorporated with structural information by accumulating the gradient magnitude of pixels that have the same GLBP pattern, thereby leading to a histogram of gradient-weighted GLBP codes which forms the feature space. Feature extraction is done at two scales and SVR is used to learn a mapping from this feature space to quality scores. In this work, we have used the version of GWHGLBP that has been trained on the LIVE MD database [31]. The No-Reference quality assessment using statistical Structural and Luminance features (NRSL) [147] is an OA NR method that uses both structural and luminance based features. NRSL begins by performing local contrast normalization as a means to reduce redundancy in a manner similar to [140]. The LBP operator [154] is locally applied to the contrast normalized image to obtain the LBP code of each pixel. These codes are then used to build a structural histogram. Separately a luminance histogram is built from the absolute magnitudes of the contrast normalized image. The structural and luminance histograms represent the feature space of NRSL, and feature extraction is done at three scales. SVR is then used to learn a mapping from the feature space to quality scores. In this work, we have used the version of NRSL that has been trained on the LIVE R2 database [24, 42].

The handcrafted features based approach is designed around features that have been selected based on domain knowledge. An alternative approach is to automatically learn features which are then used in the training process along with subjective ratings to design OA NR models. A pioneering method following this approach is called Codebook Representation for No-reference Image quality Assessment (CORNIA) [141], which uses unsupervised feature learning. CORNIA extracts a number of local descriptors by randomly sampling patches from an image, which are normalized and whitened before being used as local features. K-means clustering is performed on local features belonging to unlabeled training images to construct a visual codebook which is also normalized. Soft-assignment coding is performed on local descriptors by using the visual codebook which leads to a coefficient matrix that is converted to a fixed-length feature vector through max-pooling. In the publicly released version of CORNIA, the CSIQ database [26] is used for codebook construction and SVR with a linear kernel is used to learn a mapping from the feature vector to quality scores, where the LIVE R2 database [24, 42] has been used for model training. Compared to CORNIA, which uses low order statistics and a large codebook composed of 10,000 codewords, a recent OA NR method called High Order Statistics Aggregation (HOSA) [143] also utilizes higher order statistics and a much smaller codebook composed of only 100 codewords. HOSA extracts local features in a manner similar to CORNIA [141] and also uses K-means clustering for codebook construction. However, in addition to calculating the mean of each cluster, higher order statistics including covariance and coskewness of each cluster are also calculated. A quality aware representation of an image is obtained through soft weighted differences of image statistics, including high order statistics. SVR with a linear kernel is used to learn a mapping from the feature space to quality scores. In the publicly available version of HOSA, the codebook is constructed by using the CSIQ database [26], while the LIVE R2 database [24, 42] is used for model training.

Recently, deep neural networks (DNN) based approaches (mostly convolutional neural network (CNN) based), have been used to learn features and quality models. An end-to-end optimized DNN based approach is proposed in [148] that is capable of performing both FR and NR quality assessment, and built upon an earlier version [155]. The CNN used in [148] is based on the VGG network [156] and has ten convolutional layers, five pooling layers for feature extraction, and two fully connected layers for regression. Since CNNs require large training data and quality annotated IQA datasets are quite small, the size of the training set is augmented by randomly sampling multiple patches from each training image, which are assigned the same quality label as the parent image. The network takes image patches of size 32 × 32 pixels as input. Rectified Linear Unit (ReLU) [157] is used as the activation function. To perform IQA, an image is divided into 32 × 32 sized patches and local quality scores are pooled into a global image quality score either by simple or weighted average. The latter functionality aims to pool local quality scores based on the principles of visual saliency and is incorporated by adding a second branch that runs parallel to the quality regression branch of the network. This additional branch gives patchwise weights that are then used in pooling. For our tests we have selected the weighted average version of the NR approach proposed in [148] which is called Weighted Average Deep Image Quality Measure for NR IQA (WaDIQaM-NR) that is trained on the

LIVE R2 database [24, 42]. The Multi-task End-to-end Optimized deep neural Network (MEON) [146] is another recent DNN based approach. MEON breaks the IQA task into two subtasks that are performed by respective sub-networks: 1) Distortion identification, and 2) Quality score prediction. Instead of using ReLU [157], MEON uses the bio-inspired generalized divisive normalization (GDN) transform [122] as the activation function which allows for a reduction of model parameters. The two sub-networks in MEON share the early layers, specifically four stages are shared where each stage consists of a convolutional, GDN, and maxpooling layers. Thereafter, sub-network 1 which is responsible for distortion identification and has two fully connected layers with a GDN layer in between, produces a probability vector to identify the likelihood of each distortion. Sub-network 2 which itself has two dedicated fully connected layers with a GDN layer in between, is responsible for quality prediction and produces a score vector containing quality scores corresponding to each distortion. The probability vector from sub-network 1 is fed into sub-network 2 where it is combined with the score vector to give a final quality score in terms of a scalar value, thereby giving the network a causal structure. Due to its multi-task nature, MEON is able to break the training phase into two steps. The loss function of subtask 1 is minimized in the initial pre-training step. Since training for distortion type identification does not require subject-rated data, MEON is able to train the shared layers and sub-network 1 on a large amount of data in the pre-training step. In the second training step, the entire network is joint optimized in an end-to-end manner by using a subject-rated database. In its publicly available version, MEON used the LIVE R2 database [24, 42] for performing joint optimization. Although a number of other deep learning based approaches have been proposed recently [38, 39, 158–168], in this work we have evaluated the performance of WaDIQaM-NR [148, 155] and MEON [146] as only their author-trained models are publicly available.

Opinion-Unaware NR Methods

OU NR methods may be training-free or they may require some form of training that does not involve subject-rated images.

The Natural Image Quality Evaluator (NIQE) [3] is a pioneering general-purpose OU

NR method. Like BRISQUE [140] (discussed in the previous sub-section), NIQE operates at two scales in the spatial domain by converting an image into MSCN coefficients, uses the GGD to fit the empirical distribution of these coefficients, uses the AGGD to fit the empirical distribution of pairwise coefficient products, and uses the estimated GGD and AGGD parameters as NSS features. However, unlike BRISQUE, NIQE does not use these features in conjunction with subject-rated distorted images to train a quality model. Instead, NIQE uses the features obtained from a distorted image to fit a multivariate Gaussian (MVG) model whose distance from a universally learned MVG model of pristine natural images is regarded as a measure of quality. Although some training is required to obtain the MVG model representing pristine natural images, no training is necessary with respect to quality annotated distorted images which is what makes NIQE an OU NR method. The Integrated Local Natural Image Quality Evaluator (ILNIQE) index [144] further builds upon the approach taken in NIQE. In addition to the two NSS features employed in NIQE (statistics of MSCN coefficients and their pairwise products), three additional NSS features are included. Information about structural degradations is incorporated by including image gradient features that include image gradient components through empirical fitting parameters of a GGD and gradient magnitude through the empirical fitting parameters of a Weibull distribution. To capture the selective response of neurons in the visual cortex to stimulus orientation and frequency, multi-scale multi-orientation filter responses are obtained through log-Gabor filters. NSS features are then extracted from response maps through GGD fitting and another round of gradient statistics extraction. ILNIQE also includes color based NSS features which are obtained by first taking the RGB color image to the logarithmic scale and then converting it to an opponent color space. A Gaussian model is then used to empirically fit the coefficient distributions in the opponent color space, thereby providing another set of NSS features. Like NIQE, ILNIQE determines the quality of a distorted image by measuring the distance between the MVG fit of its NSS features and the universal MVG model of pristine natural images. However, instead of using a single MVG model for the distorted image, image quality is determined at the patch level and then pooling is done to obtain a final quality score. ILNIQE also uses principal component analysis (PCA) to reduce correlation between features and for dimensional reduction.

The Quality Aware Clustering (QAC) method [35] takes an alternative approach to OU NR design. QAC partitions an image into a set of overlapping patches which are first divided into groups based on similar quality and then patches with similar local structures are clustered together. The local features are extracted through the application of a high pass filter. A set of centroids are learned for each quality group and form a codebook which is used to determine the quality of each patch. QAC has the capability to give a local quality map as well as an overall quality score. Although during its development QAC needs to divide image patches into groups based on quality, it does not use subject-rated databases to accomplish this. Instead it builds a new database starting from 10 source images from the Berkeley Segmentation database [169], and uses the FR IQA method FSIM [14] to annotate patch quality which is normalized through a percentile pooling procedure. Although QAC training does involve working with distorted images, it is still an OU NR method since it does not train against subject-rated distorted images. A quality-discriminable image pair (DIP) based recent OU NR method called DIP inferred quality (dipIQ) index [36] uses DIPs for training. First a new dataset is constructed that has 840 source and 16,800 distorted images (which include Gaussian noise, Gaussian blur, JPEG and JPEG2000 compression). A DIP generation engine is constructed which uses three FR IQA methods, GMSD [99], MSSSIM [4], and VIF [113], to annotate distorted image quality. Each candidate image pair is assigned with a non-negative T value equivalent to the smallest score difference of the FR models. A raised-cosine function is used to quantify the quality discriminability uncertainty level based on T values. 80 million DIPs are produced using this DIP generation engine. Using these DIPs with their associated uncertainty levels and CORNIA features [141] as base features for image representation, RankNet [170] which is a neural network based pairwise learning-to-rank algorithm, is employed to learn an OU NR model.

Some OU NR methods take a training-free approach. The Six-Step Blind Metric (SISBLIM) [32], which is itself an improved version of FISBLIM [171], has been developed for singly and multiply distorted images and operates by determining the individual and joint impact of different distortions. It first uses the approach in [172] to estimate the amount of noise in a distorted image and then denoises the image by using the BM3D method [173]. The estimates of blur and JPEG quality are determined from the denoised image by using the methods proposed in [136] and [137] respectively. To take into account the interaction between different distortions and the masking effect due to image content, a model based on the free energy theory [174] is used to quantify the joint effects. Finally, the SISBLIM score is obtained as a linear combination of weighted quality estimates of noise, blur, JPEG compression and joint effects. The Local Pattern Statistics Index (LPSI) [145] is another recent training-free OU NR method that utilizes the LBP operator [154]. To reduce computational complexity, LPSI uses only four neighbors of each image pixel to compute LBP codes, which leads to six distinct binary patterns. Based on analysis, LPSI picks the locally weighted statistic associated with one of these six binary patterns as a quality measure since it offers the best discriminant ability to distinguish most distortions from pristine natural images.

2.4 Performance Analysis of FR and Fused FR Methods

2.4.1 Evaluation Criteria

Prediction Accuracy

The Pearson Linear Correlation Coefficient (PLCC) is used as a measure of a method's prediction accuracy [72]. Since the scores produced by objective IQA methods are usually not linear with respect to subjective ratings, a nonlinear regression step is necessary before the computation of PLCC. We do this by adopting the five-parameter modified logistic function used in [24]:

$$P(Q) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\{\beta_2 (Q - \beta_3)\}}} \right] + \beta_4 Q + \beta_5 \tag{2.14}$$

where Q denotes the objective quality scores directly from an IQA method, P denotes the IQA scores after the regression step, and β_1, β_2, β_3, β_4, and β_5 are model parameters that are found numerically in MATLAB to maximize the correlation between subjective and objective scores. Given a database with its subjective scores denoted by S, the PLCC value of an IQA method is then calculated as:

$$PLCC(P,S) = \frac{\sum_{i=1}^{N}(P_i - \bar{P}) \cdot (S_i - \bar{S})}{\sqrt{\sum_{i=1}^{N}(P_i - \bar{P})^2 \cdot \sum_{i=1}^{N}(S_i - \bar{S})^2}} \tag{2.15}$$

where P_i and S_i are respectively the values in the vectors P and S for the image i, $\bar{P}$ and $\bar{S}$ are respectively the mean values of vectors P and S, while N is the number of images in the database.

Prediction Monotonicity

The Spearman Rank-order Correlation Coefficient (SRCC) is used as a measure of a method's prediction monotonicity [72]. SRCC is a non-parametric rank-order based correlation metric and does not require the preceding nonlinear mapping step. The SRCC value of an IQA method on a database with N images is calculated as:

$$SRCC(Q,S) = 1 - \left[\frac{6\sum_{i=1}^{N} {d_i}^2}{N(N^2-1)}\right] \tag{2.16}$$

where d_i is the difference between the i-th image's ranks in the objective (Q) and subjective (S) scores. Another rank-order based method, Kendall's Rank-order Correlation Coefficient (KRCC), is found to be highly consistent with the SRCC measure and provides minimal additional information, and thus is not included in the current work.

Statistical Significance Testing

Conclusions drawn about the performance of IQA methods based on PLCC and SRCC values can only be considered *universal* if testing is done on the entire population of concerned data, which in this case is the space of all possible natural images and their distorted versions. Since this is not possible and subject-rated IQA databases can only be regarded as sparse random samples from this enormous population, hypothesis testing is performed to ascertain whether the drawn inferences on a given sample size are statistically significant at a particular confidence level. The term *statistical significance* signifies whether the difference in the performance of one IQA method with respect to another, on a set of sample points, is purely due to chance or due to some genuine underlying effect [175]. Generalizations about the difference in method performance can only be made in the latter case at the stated confidence level.

In the field of IQA, statistical significance testing is usually carried out on model prediction residuals. Given the objective scores of different IQA methods to be compared, they are converted to prediction residuals by first mapping them to the MOS/DMOS range of the database being used for testing by using the nonlinear mapping procedure explained for PLCC calculation earlier in this section, and then subtracting the actual subjective scores from these *predicted* subjective scores. In this work we use the one-sided (left-tailed) two-sample F-test [175] to statistically compare the performance of any two given IQA methods at the 5% significance level (95% confidence). The *null* hypothesis is that the data in the two residual vectors comes from normal distributions with the same variance, making them statistically indistinguishable. The *alternative* hypothesis is that the data in the residual vectors comes from normal distributions with different variances, making them statistically distinguishable. The test statistic is the ratio of the variances of the two residual vectors. Given the number of residuals and the confidence level, a critical threshold is determined. If the value of the test statistic is smaller than the critical threshold, then this indicates a failure to reject the null hypothesis. By performing the one-sided test twice with the order of the methods swapped, we were able to determine if their performance is statistically indistinguishable or whether one method performed better than the other. In the statistical significance testing tables that follow, a “1”, “-”, or “0” mean that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column, respectively. Since the tests assume the Gaussianity of prediction residuals, we use a simple kurtosis based check for Gaussianity as in [24]. If the kurtosis of prediction residuals of an IQA method is between 2 and 4, then they are accepted for the Gaussianity assumption.

2.4.2 Performance of FR Methods on Individual Databases

We tested the 43 FR methods discussed in Section 2.3.1 and given in Table 2.3, on each of the nine subject-rated IQA databases mentioned in Table 2.2, of which five are single distortion datasets discussed in Section 2.2.1 and four are multiple distortion datasets discussed in Section 2.2.2. Since the single distortion database CIDIQ [5] contains subjective scores at two viewing distances, testing was done separately for each case and results are mentioned under the headings of CIDIQ50 and CIDIQ100 for the viewing distances of 50 cm and 100 cm, respectively. For each database, testing was done on the entire dataset, that is, all distortions were considered. The test results are given in Table 2.6 in terms of PLCC and in Table 2.7 in terms of SRCC.

2.4.3 Selection of FR Methods for Fusion in Fused FR Methods

We described seven fusion based FR methods in Section 2.3.2 and listed them in Table 2.4. The four methods belonging to the *empirical fusion* category (HFSIMc [129], CISI [126], CM3 [127], and CM4 [127]) combine specific FR methods and hence do not need to select methods from a large pool. The authors of the *learning based fusion* method CNNM [128] provide a pre-trained model that combines six FR methods, and we use the same selection and order for CNNM.

Although the authors of the *rank aggregation based fusion* method RAS [41] (discussed in Section 2.3.2) provide a selection of methods to be fused, we believe that a more extensive search needs to be done to select FR methods for fusion, especially if the resulting scores are to be used as alternative ground truth for annotating large datasets. We begin by identifying a pool of FR methods to be combined in RAS [41]. Since RAS [41] not only combines FR methods but then adjusts the score of a base FR method with respect to the consensus ranking, an exhaustive search would require testing all possible combinations for each FR method being used as the base method. Given that we are considering 43 FR methods (Table 2.3), this would require testing more than 189 trillion combinations, which is computationally infeasible. To reduce the computational load, we make three sets of 15 FR methods each based on time constraints and thus test 245,760 combinations in

Table 2.6: Test results of 43 FR methods on nine subject-rated IQA databases in terms of PLCC. All distortions in each dataset were considered.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
AD_DWT	0.9384	0.3624	0.8163	0.8692	0.4100	0.5379	0.6010	0.7159	0.8501	0.8506
ADM	0.9360	0.8355	0.9285	0.9182	0.7791	0.8196	0.8349	0.6428	0.9062	0.9060
CID_MS	0.9159	0.8362	0.8732	0.9375	0.8364	0.8171	0.8414	0.6183	0.8917	0.8961
CID_SS	0.9279	0.8038	0.9079	0.9357	0.8534	0.7806	0.8617	0.6258	0.8822	0.8750
DSS	0.9618	0.8530	0.9612	0.9259	0.7715	0.8267	0.8733	0.8168	0.9023	0.8973
DVICOM	0.9734	0.8194	0.9179	0.9144	0.8035	0.8018	0.8919	0.8161	0.8873	0.8773
DVICOM_F	0.9735	0.8194	0.9191	0.9170	0.8037	0.8001	0.8916	0.8097	0.8858	0.8797
DWT_VIF	0.9658	0.7406	0.9009	0.8901	0.6952	0.5516	0.8941	0.7212	0.8716	0.8393
ESSIM	0.9566	0.8645	0.9224	0.9094	0.7953	0.8255	0.8451	0.6953	0.8861	0.9081
FSIM	0.9597	0.8589	0.9120	0.9185	0.7410	0.8265	0.8969	0.6474	0.8933	0.9037
FSIMc	0.9613	0.8769	0.9191	0.9329	0.7583	0.8410	0.8998	0.6412	0.8965	0.9039
GMSD	0.9603	0.8590	0.9541	0.9176	0.7387	0.7585	0.8776	0.8309	0.8808	0.8685
GSIM	0.9512	0.8464	0.8964	0.9155	0.7700	0.8342	0.8352	0.6647	0.8808	0.9072
IFC	0.9268	0.1737	0.8366	0.8614	0.5479	0.1724	0.9162	0.6279	0.9058	0.7990
IW_PSNR	0.9329	0.5984	0.8024	0.9212	0.6273	0.7200	0.6951	0.7649	0.8284	0.8771
IWSSIM	0.9522	0.8319	0.9144	0.9191	0.8476	0.8698	0.8983	0.8513	0.9109	0.9056
MAD	0.9675	0.8464	0.9500	0.9053	0.7809	0.8411	0.7552	0.7471	0.8948	0.8985
MCSD	0.9675	0.8648	0.9560	0.9217	0.7532	0.7727	0.8637	0.8275	0.8847	0.8787
MSSSIM	0.9489	0.8329	0.8991	0.9232	0.8180	0.8039	0.8419	0.7273	0.8747	0.8805
NQM	0.9129	0.6794	0.7200	0.9429	0.4879	0.6712	0.6170	0.3946	0.9086	0.7931
PSNR	0.8723	0.6775	0.7512	0.8321	0.6232	0.6814	0.6091	0.5564	0.7398	0.6806
PSNR_DWT	0.9301	0.6921	0.7631	0.8902	0.5792	0.6722	0.6393	0.5725	0.8630	0.8186
PSNR_HAc	0.9164	0.8418	0.9017	0.8759	0.7408	0.7624	0.7436	0.6768	0.7851	0.7322
PSNR_HA	0.9130	0.8511	0.8592	0.8697	0.6913	0.7292	0.7269	0.6825	0.8004	0.8093
PSNR_HMAc	0.9295	0.8329	0.8672	0.8977	0.7314	0.7896	0.7655	0.7255	0.8090	0.7560
PSNR_HMA	0.9249	0.8275	0.8342	0.8951	0.6831	0.7459	0.7437	0.7296	0.8192	0.8512
PSNR_HVS	0.9134	0.7031	0.7808	0.8843	0.6346	0.7073	0.6764	0.6813	0.7996	0.8085
PSNR_HVSM	0.9251	0.6709	0.7725	0.8841	0.6303	0.7042	0.6814	0.7281	0.8182	0.8506
QASD	0.9574	0.8897	0.9481	0.9253	0.7257	0.8116	0.8063	0.6312	0.8966	0.8827
RFSIM	0.9386	0.8329	0.9164	0.8904	0.6943	0.7621	0.7084	0.4738	0.8713	0.8200
SFF	0.9632	0.8706	0.9643	0.7761	0.7834	0.7721	0.8590	0.7952	0.8893	0.8904
SNR	0.8616	0.6498	0.7414	0.8228	0.6374	0.6888	0.6474	0.4264	0.7283	0.6414
SRSIM	0.9555	0.8664	0.9244	0.9022	0.7066	0.8147	0.8685	0.6401	0.8883	0.8928
SSIM	0.9449	0.7895	0.8612	0.9144	0.7674	0.8230	0.8457	0.5249	0.8915	0.8623
SSIM_DWT	0.9559	0.7799	0.9050	0.8955	0.8405	0.7821	0.8810	0.7624	0.8913	0.8594
UQI	0.8984	0.6427	0.8294	0.7981	0.6078	0.4980	0.8277	0.5318	0.8540	0.7723
VIF	0.9604	0.7720	0.9278	0.8938	0.7267	0.6415	0.9367	0.8376	0.9030	0.8736
VIF_DWT	0.9657	0.7657	0.9123	0.8969	0.7259	0.5845	0.9031	0.7531	0.8839	0.8653
VIF_P	0.9596	0.7529	0.9044	0.8921	0.7073	0.5629	0.8827	0.7589	0.8712	0.8126
VSI	0.9482	0.9000	0.9279	0.9320	0.7226	0.8240	0.8703	0.5512	0.8789	0.8749
VSNR	0.9236	0.7138	0.7355	0.8794	0.6261	0.7424	0.6805	0.3775	0.8309	0.8037
WSNR	0.9144	0.6031	0.7337	0.8468	0.5752	0.6766	0.5889	0.6853	0.8185	0.7942
WSSI	0.9549	0.7698	0.9001	0.9072	0.8406	0.7691	0.8785	0.7543	0.8843	0.8551

Table 2.7: Test results of 43 FR methods on nine subject-rated IQA databases in terms of SRCC. All distortions in each dataset were considered.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
AD_DWT	0.9412	0.5967	0.8029	0.8628	0.5522	0.6244	0.6027	0.7750	0.8040	0.7810
ADM	0.9460	0.7874	0.9333	0.9138	0.7794	0.8185	0.8186	0.6248	0.8815	0.8490
CID_MS	0.9103	0.8314	0.8789	0.9366	0.8350	0.8062	0.8330	0.6168	0.8608	0.8778
CID_SS	0.9270	0.7879	0.9116	0.9304	0.8528	0.7789	0.8535	0.6236	0.8408	0.8208
DSS	0.9616	0.7921	0.9555	0.9272	0.7755	0.8246	0.8658	0.8078	0.8714	0.8759
DVICOM	0.9750	0.7598	0.9181	0.9155	0.8034	0.7903	0.8840	0.8168	0.8672	0.8374
DVICOM_F	0.9748	0.7606	0.9226	0.9181	0.8028	0.7909	0.8837	0.8104	0.8642	0.8411
DWT_VIF	0.9671	0.6093	0.8909	0.8833	0.6909	0.5434	0.8836	0.7229	0.8269	0.7921
ESSIM	0.9597	0.8035	0.9325	0.9075	0.7968	0.8253	0.8250	0.6966	0.8517	0.8682
FSIM	0.9634	0.8015	0.9242	0.9178	0.7438	0.8149	0.8872	0.5817	0.8635	0.8585
FSIMc	0.9645	0.8510	0.9309	0.9323	0.7608	0.8285	0.8904	0.5806	0.8666	0.8613
GMSD	0.9603	0.8044	0.9570	0.9177	0.7427	0.7675	0.8613	0.8283	0.8448	0.8210
GSIM	0.9561	0.7946	0.9107	0.9121	0.7709	0.8299	0.8137	0.6637	0.8454	0.8485
IFC	0.9259	0.5389	0.7671	0.8570	0.4929	0.3427	0.9119	0.6861	0.8839	0.7807
IW_PSNR	0.9328	0.6913	0.8310	0.9166	0.6013	0.7137	0.6719	0.7816	0.7572	0.8178
IWSSIM	0.9567	0.7779	0.9212	0.9163	0.8484	0.8564	0.8911	0.8551	0.8836	0.8588
MAD	0.9669	0.7807	0.9466	0.9061	0.7815	0.8391	0.7249	0.7507	0.8646	0.8643
MCSD	0.9668	0.8089	0.9592	0.9224	0.7562	0.7808	0.8451	0.8269	0.8517	0.8370
MSSSIM	0.9513	0.7859	0.9132	0.9227	0.8196	0.7988	0.8296	0.7238	0.8363	0.8274
NQM	0.9093	0.6465	0.7411	0.9436	0.4694	0.6323	0.5827	0.4016	0.8999	0.7460
PSNR	0.8756	0.6394	0.8057	0.8246	0.6254	0.6701	0.5784	0.5604	0.6771	0.6136
PSNR_DWT	0.9325	0.6426	0.8052	0.8819	0.5401	0.6419	0.6070	0.5797	0.8206	0.7385
PSNR_HAc	0.9216	0.8187	0.9261	0.8702	0.7430	0.7684	0.7240	0.6724	0.7112	0.6789
PSNR_HA	0.9192	0.7792	0.9147	0.8610	0.6875	0.7295	0.7055	0.6785	0.7146	0.7284
PSNR_HMAc	0.9338	0.8128	0.9121	0.8907	0.7278	0.7877	0.7461	0.7249	0.7403	0.7114
PSNR_HMA	0.9298	0.7568	0.8997	0.8847	0.6634	0.7388	0.7239	0.7281	0.7423	0.7625
PSNR_HVS	0.9186	0.6533	0.8294	0.8781	0.6313	0.7011	0.6490	0.6779	0.7126	0.7278
PSNR_HVSM	0.9295	0.6246	0.8221	0.8756	0.6122	0.6969	0.6559	0.7273	0.7410	0.7619
QASD	0.9629	0.8674	0.9530	0.9231	0.7307	0.8079	0.7778	0.6687	0.8766	0.8315
RFSIM	0.9434	0.7743	0.9291	0.8871	0.6795	0.7450	0.6766	0.4151	0.8330	0.7756
SFF	0.9649	0.8513	0.9627	0.7738	0.7834	0.7689	0.8396	0.8005	0.8700	0.8535
SNR	0.8650	0.6127	0.7994	0.8101	0.6358	0.6709	0.6278	0.4383	0.6135	0.5767
SRSIM	0.9620	0.8076	0.9317	0.9021	0.7087	0.7966	0.8521	0.6238	0.8666	0.8350
SSIM	0.9479	0.7417	0.8755	0.9112	0.7697	0.8094	0.8328	0.4873	0.8604	0.7966
SSIM_DWT	0.9603	0.7093	0.9111	0.8877	0.8410	0.7815	0.8690	0.7650	0.8587	0.7929
UQI	0.8941	0.5507	0.8098	0.7984	0.5937	0.4743	0.8183	0.5334	0.8149	0.7311
VIF	0.9636	0.6769	0.9194	0.8866	0.7203	0.6257	0.9306	0.8444	0.8823	0.8381
VIF_DWT	0.9681	0.6439	0.9020	0.8930	0.7224	0.5826	0.8943	0.7553	0.8479	0.8243
VIF_P	0.9618	0.6101	0.8807	0.8919	0.7029	0.5471	0.8770	0.7594	0.8367	0.7711
VSI	0.9524	0.8965	0.9422	0.9317	0.7213	0.8106	0.8569	0.5700	0.8414	0.8269
VSNR	0.9279	0.6817	0.8108	0.8741	0.6145	0.7200	0.6594	0.3923	0.7719	0.7420
WSNR	0.9158	0.5782	0.7729	0.8381	0.5600	0.6542	0.5428	0.6998	0.7611	0.7243
WSSI	0.9586	0.6937	0.9075	0.9004	0.8411	0.7705	0.8690	0.7479	0.8494	0.7866

Table 2.8: RAS exhaustive search set composition.

S. No.	Fast Set	Medium Set	Full Set
1	ADM	ADM	ADM
2	DSS	CID_MS	CID_MS
3	ESSIM	DSS	DSS
4	FSIM	DVICOM_F	DVICOM
5	FSIMc	ESSIM	ESSIM
6	GMSD	FSIMc	FSIMc
7	GSIM	GMSD	GMSD
8	IWSSIM	GSIM	IWSSIM
9	MCSD	IWSSIM	MAD
10	MSSSIM	MCSD	MCSD
11	SFF	MSSSIM	QASD
12	SRSIM	SFF	SFF
13	SSIM_DWT	SRSIM	SRSIM
14	VIF_DWT	VIF_DWT	VIF
15	VSI	VSI	VSI

each case for a total of 737,280 tests. The sets were formed subject to the following three conditions for a color test image of size 1024 × 1024: 1) The first set, called the Fast Set, only contains top performing FR methods that require less than 1.5 seconds to determine the quality of the test image. 2) The second set, called the Medium Set, contains top performing FR methods that take less than 2.7 seconds to determine the quality of the test image. 3) The final set, called the Full Set, has no time constraints. The FR methods in each of the three sets for the RAS exhaustive search are given in Table 2.8. Based on weighted average SRCC, top performing FR method combinations were selected in each set.

Instead of just computing the weighted average SRCC across all nine subject-rated databases, we compute weighted average SRCC for three categories: 1) Across all databases, 2) Across only the five single distortion databases, and 3) Across only the four multiple distortion databases. This was done to more thoroughly analyze how the performance of RAS varies for these different conditions. Within each category, all distorted images of the constituent databases were considered. These three categories were considered for each of the three sets of FR methods (Table 2.8), leading to a total of nine possibilities. The top performing combinations obtained in the exhaustive search for all these possibilities are given in Table 2.9, where each distinct combination is assigned a unique name (RAS1 to RAS7). The following observations can be made: 1) Although combinations of up to 15 FR methods were tested, the top performing combinations only include two to four FR methods in the fusion process. Thus, the notion of *the more the better* is not valid when it comes to fusion based FR methods. 2) The methods in each combination usually follow different design philosophies. While RAS4 and RAS5 differ in the base FR method, they combine the same three FR methods that include CID_MS [95] which follows a multiscale similarity based approach with emphasis on color features, SFF [109] which follows a sparsity based approach, and VSI [15] which follows a similarity based approach that incorporates visual saliency based weighted pooling. RAS2 combines a similarity based approach (VSI) with a sparsity based approach (SFF). RAS3 combines two similarity based approaches, DSS [16] (similarity in the DCT domain) and IWSSIM [13] (multiscale similarity measure that employs information content weighting in the pooling stage), with VIF_DWT [93] which follows a NSS based approach to IQA. RAS6 builds on RAS3 by adding CID_MS [95] to the combination which emphasizes on color based similarity. It is thus evident that RAS prefers combining different IQA design philosophies, such that they complement each other. The deficiencies in one design philosophy with regard to a particular distortion may be addressed by the strengths of another design approach. 3) RAS favors color based FR methods. All FR methods combined in RAS1, RAS2, RAS4, and RAS5, are color based, while RAS6 and RAS7 combine both color and grayscale based methods. Only RAS3 combines exclusively grayscale based methods.

Table 2.9: RAS exhaustive search outcome for each search set and database category.

Search Set	Database Category	Individual FR Methods included in Fusion				Base FR Method	Name Given in this Work
		Method 1	Method 2	Method 3	Method 4		
Fast	All Databases	FSIMc	SFF	VSI	–	SFF	RAS1
	Single Distortion Databases	SFF	VSI	–	–	VSI	RAS2
	Multiple Distortion Databases	DSS	IWSSIM	VIF_DWT	–	DSS	RAS3
Medium	All Databases[a]	CID_MS	SFF	VSI	–	CID_MS	RAS4
	Single Distortion Databases[b]	CID_MS	SFF	VSI	–	VSI	RAS5
	Multiple Distortion Databases	CID_MS	DSS	IWSSIM	VIF_DWT	DSS	RAS6
Full	All Databases[a]	CID_MS	SFF	VSI	–	CID_MS	RAS4
	Single Distortion Databases[b]	CID_MS	SFF	VSI	–	VSI	RAS5
	Multiple Distortion Databases	CID_MS	DSS	VIF	–	VIF	RAS7

[a] Exhaustive Search for the *All Databases* category leads to the same outcome for the Medium and Full Sets (RAS4).

[b] Exhaustive Search for the *Single Distortion Databases* category leads to the same outcome for the Medium and Full Sets (RAS5).

As stated earlier, for the *learning based fusion* method MMF [130] (discussed in Section 2.3.2) we test its context free version called CF-MMF. Since MMF follows a supervised learning based approach using SVR, different sets of FR methods can be combined. For this work, we select the version of CF-MMF recommended for the TID2008 database [25] through the SFMS strategy in [130], where it is computed that for this dataset, six FR methods should be combined. For the said version of CF-MMF [130], the six FR methods that are part of the fusion process are: FSIM [14], VIF [113], IFC [101], MAD [26], PSNR_HVS [105], and MSSSIM [4]. Since a pre-trained version of this model is not available, we follow the approach in [130] and train the model ourselves through SVR with a radial basis function (RBF). Instead of the TID2008 database [25], we use its enhanced version TID2013 [19], and the above-mentioned six FR methods to learn a fusion model. Half of the TID2013 dataset is used for training, half for validation, and grid search is employed to ascertain optimal SVR parameters. We refer the corresponding model as MMF1.

To provide a more thorough comparison of MMF [130] with RAS [41], we train three other CF-MMF models, one each for the three FR method pools identified for RAS in Table 2.8. As computed in [130] six FR methods should be combined for the TID2008 database [25], and we follow this recommendation for TID2013 [19] as well. We use the SFMS strategy [130] to identify the methods to be fused for each of the three FR method pools (Table 2.8) and built the following three CF-MMF models: 1) MMF2 developed on the Fast Set combines VSI [15], ADM [94], VIF_DWT [93], MCSD [102], IWSSIM [13], and SFF [109]. 2) MMF3 developed on the Medium Set combines VSI [15], ADM [94], VIF_DWT [93], CID_MS [95], GMSD [99], and SRSIM [110]. 3) MMF4 developed on the Full Set combines VSI [15], ADM [94], CID_MS [95], MCSD [102], GMSD [99], and IWSSIM [13]. Training for each of these MMF models was done in a manner similar to MMF1 and data scaling was applied where necessary as recommended in [130]. In each case, eight FR methods were combined but performance gain beyond the combination of six methods was negligible, and hence we combine six methods in the final models. Since the combinations in the four CF-MMF methods are not being used in the seven RAS methods discussed earlier, to provide another comparison point between RAS and MMF, we construct four additional RAS models that use the same FR method combinations as in each of the CF-MMF models. Specifically, RAS_MMF1, RAS_MMF2, RAS_MMF3, and RAS_MMF4, use the RAS technique [41] to fuse the FR methods that are selected for combination in MMF1, MMF2, MMF3, and MMF4, respectively. For these additional RAS methods, the base FR method was selected as the first one identified by the SFMS strategy. The details of all the fusion based methods whose performance is being evaluated in this work, including the various versions of RAS and MMF, are given in Table 2.10. RAS_B1 and RAS_B2 are the versions of RAS discussed in [41]. Although overall, we are evaluating the performance of seven different fused FR techniques, it can be noted from Table 2.10 that we are considering four different versions of MMF and 13 different versions of RAS. Thus, in total, 22 fused FR methods are being evaluated in this work.

Table 2.10: Fused FR methods information table.

Fused FR	Methods	Individual FR Methods included in Fusion						Notes	
Method	Fused	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6		
CISI	3	MSSSIM	VIF	FSIMc	–	–	–	–	
CM3	3	IFC	NQM	VSNR	–	–	–	–	
CM4	4	IFC	NQM	VSNR	VIF	–	–	–	
CNNM	6	FSIMc	PSNR_HMAc	PSNR_HVS	SFF	SRSIM	VIF	–	
HFSIMc	2	RFSIM	FSIMc	–	–	–	–	–	
MMF1	6	FSIM	IFC	MAD	MSSSIM	PSNR_HVS	VIF	Selection Method: SFMS	
MMF2	6	VSI	ADM	VIF_DWT	MCSD	IWSSIM	SFF		
MMF3	6	VSI	ADM	VIF_DWT	CID_MS	GMSD	SRSIM		
MMF4	6	VSI	ADM	CID_MS	MCSD	GMSD	IWSSIM		
RAS_B1	5	FSIM	FSIMc	GMSD	IWSSIM	VIF	–	Base FR	GMSD
RAS_B2	4	FSIM	FSIMc	GMSD	IWSSIM	–	–		GMSD
RAS_MMF1	6	FSIM	IFC	MAD	MSSSIM	PSNR_HVS	VIF		FSIM
RAS_MMF2	6	VSI	ADM	VIF_DWT	MCSD	IWSSIM	SFF		VSI
RAS_MMF3	6	VSI	ADM	VIF_DWT	CID_MS	GMSD	SRSIM		VSI
RAS_MMF4	6	VSI	ADM	CID_MS	MCSD	GMSD	IWSSIM		VSI
RAS1	3	FSIMc	SFF	VSI	–	–	–		SFF
RAS2	2	SFF	VSI	–	–	–	–		VSI
RAS3	3	DSS	IWSSIM	VIF_DWT	–	–	–		DSS
RAS4	3	SFF	CID_MS	VSI	–	–	–		CID_MS
RAS5	3	SFF	CID_MS	VSI	–	–	–		VSI
RAS6	4	DSS	IWSSIM	CID_MS	VIF_DWT	–	–		DSS
RAS7	3	CID_MS	DSS	VIF	–	–	–		VIF

2.4.4 Performance of Fused FR Methods on Individual Databases

We tested the performance of the 22 fused FR methods mentioned in Table 2.10 on each of the nine subject-rated databases mentioned in Table 2.2 (CIDIQ database [5] at two viewing distances). The test results are given in Table 2.11 in terms of PLCC and in Table 2.12 in terms of SRCC. Testing was done on all distortion types included in each dataset.

Table 2.11: Test results of 22 Fused FR methods on nine subject-rated IQA databases in terms of PLCC. All distortions in each dataset were considered.

Fused Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
CISI	0.9625	0.8575	0.9364	0.9289	0.8239	0.8193	0.9220	0.7095	0.9032	0.9007
CM3	0.8337	0.6058	0.6870	0.9435	0.6383	0.7238	0.6362	0.4604	0.8718	0.8062
CM4	0.8072	0.5891	0.6597	0.9432	0.6238	0.7086	0.6128	0.5622	0.8422	0.8219
CNNM	0.8892	0.9338	0.9007	0.8741	0.6439	0.6548	0.8328	0.6378	0.8249	0.7665
HFSIMc	0.9579	0.8635	0.9304	0.9211	0.7365	0.8120	0.8357	0.5242	0.8918	0.8893
MMF1	0.8561	0.9504	0.9202	0.8624	0.7326	0.7572	0.8185	0.6736	0.8523	0.8075
MMF2	0.8887	0.9512	0.9120	0.8608	0.6437	0.5736	0.8416	0.6056	0.8678	0.7964
MMF3	0.8831	0.9516	0.9274	0.8463	0.6186	0.6962	0.8692	0.7012	0.8241	0.8047
MMF4	0.8818	0.9532	0.9394	0.8681	0.6392	0.7131	0.8751	0.5785	0.7911	0.8391
RAS_B1	0.9683	0.8582	0.9539	0.9241	0.8255	0.8299	0.9177	0.7991	0.8980	0.9030
RAS_B2	0.9647	0.8701	0.9408	0.9255	0.7905	0.8350	0.9030	0.7730	0.8958	0.9041
RAS_MMF1	0.9696	0.8273	0.9622	0.9284	0.8521	0.8097	0.9059	0.7850	0.9034	0.9108
RAS_MMF2	0.9662	0.8596	0.9604	0.9200	0.8502	0.8436	0.9059	0.7643	0.8990	0.9084
RAS_MMF3	0.9638	0.8616	0.9568	0.9384	0.8569	0.8646	0.9104	0.7080	0.9015	0.9121
RAS_MMF4	0.9620	0.8815	0.9420	0.9409	0.8412	0.8719	0.9039	0.7620	0.9023	0.9187
RAS1	0.9659	0.8958	0.9567	0.9008	0.7995	0.8427	0.8958	0.7160	0.8945	0.9006
RAS2	0.9617	0.9003	0.9514	0.8930	0.7983	0.8387	0.8873	0.7100	0.8896	0.8897
RAS3	0.9701	0.8423	0.9660	0.9266	0.8583	0.8313	0.9262	0.8424	0.9111	0.9111
RAS4	0.9590	0.8912	0.9348	0.9383	0.8555	0.8739	0.8986	0.7173	0.9108	0.9167
RAS5	0.9588	0.8980	0.9406	0.9313	0.8462	0.8713	0.8999	0.7119	0.9048	0.9124
RAS6	0.9682	0.8488	0.9640	0.9408	0.8832	0.8585	0.9294	0.8181	0.9150	0.9202
RAS7	0.9687	0.8320	0.9670	0.9393	0.8788	0.8428	0.9434	0.8189	0.9144	0.9167

2.4.5 Overall Performance

Since we are evaluating the performance of IQA methods on nine different databases, a measure of overall performance is necessary. We provide this measure by computing the

Table 2.12: Test results of 22 Fused FR methods on nine subject-rated IQA databases in terms of SRCC. All distortions in each dataset were considered.

Fused Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
CISI	0.9680	0.8150	0.9425	0.9270	0.8231	0.8063	0.9135	0.6920	0.8740	0.8612
CM3	0.9207	0.7136	0.8073	0.9450	0.6452	0.7659	0.7114	0.5055	0.9206	0.7733
CM4	0.9316	0.7195	0.8247	0.9441	0.6417	0.7686	0.7661	0.6209	0.9224	0.7891
CNNM	0.8928	0.9201	0.8850	0.8763	0.6270	0.6451	0.8218	0.6720	0.8048	0.7260
HFSIMc	0.9610	0.8228	0.9423	0.9205	0.7315	0.7982	0.8202	0.5075	0.8624	0.8453
MMF1	0.8741	0.9409	0.9043	0.8594	0.7241	0.7379	0.8084	0.6799	0.8085	0.7703
MMF2	0.8907	0.9436	0.8910	0.8448	0.5720	0.5318	0.8196	0.6111	0.8533	0.7785
MMF3	0.8947	0.9455	0.9303	0.8345	0.6286	0.6517	0.8580	0.6606	0.7253	0.7265
MMF4	0.8852	0.9452	0.9438	0.8685	0.5990	0.6422	0.8596	0.6004	0.7533	0.8123
RAS_B1	0.9690	0.8034	0.9563	0.9223	0.8252	0.8312	0.9089	0.8013	0.8688	0.8595
RAS_B2	0.9653	0.8116	0.9464	0.9235	0.7917	0.8319	0.8932	0.7741	0.8654	0.8580
RAS_MMF1	0.9717	0.7355	0.9607	0.9268	0.8536	0.8133	0.8984	0.7923	0.8756	0.8720
RAS_MMF2	0.9689	0.8158	0.9642	0.9185	0.8490	0.8392	0.8952	0.7686	0.8714	0.8635
RAS_MMF3	0.9663	0.8195	0.9610	0.9383	0.8573	0.8583	0.9010	0.7132	0.8733	0.8699
RAS_MMF4	0.9642	0.8350	0.9493	0.9404	0.8430	0.8662	0.8953	0.7698	0.8730	0.8763
RAS1	0.9672	0.8756	0.9602	0.8958	0.7986	0.8375	0.8857	0.7205	0.8675	0.8593
RAS2	0.9637	0.8876	0.9591	0.8902	0.7975	0.8313	0.8759	0.7164	0.8589	0.8473
RAS3	0.9712	0.7794	0.9625	0.9261	0.8575	0.8271	0.9204	0.8455	0.8842	0.8796
RAS4	0.9590	0.8819	0.9422	0.9395	0.8562	0.8638	0.8913	0.7230	0.8836	0.8914
RAS5	0.9599	0.8864	0.9471	0.9313	0.8471	0.8632	0.8920	0.7168	0.8770	0.8822
RAS6	0.9680	0.7930	0.9603	0.9405	0.8840	0.8532	0.9250	0.8214	0.8867	0.8954
RAS7	0.9687	0.7724	0.9606	0.9388	0.8788	0.8363	0.9397	0.8250	0.8886	0.8986

weighted average PLCC and SRCC values for each IQA method over different databases (as in [13]). The weight assigned to a database depends on its size in terms of the number of distorted images. The weighted average PLCC and SRCC for an IQA method over different databases are computed as:

$$PLCC_{WA} = \frac{\sum_{i=1}^{D} n_i \cdot PLCC_i}{\sum_{i=1}^{D} n_i} \tag{2.17}$$

$$SRCC_{WA} = \frac{\sum_{i=1}^{D} n_i \cdot SRCC_i}{\sum_{i=1}^{D} n_i} \tag{2.18}$$

where $PLCC_i$ and $SRCC_i$ are respectively the PLCC and SRCC values of the IQA method for database i, n_i is the number of images in database i, and D is the number of databases being considered. Although we are using nine IQA databases in this work (five singly distorted and four multiply distorted), since the singly distorted database CIDIQ [5] provides MOS at two viewing distances, it will be regarded as two datasets. We compute weighted average PLCC and SRCC for three cases: 1) All databases ($D = 10$), 2) Only single distortion databases ($D = 6$), and 3) Only multiple distortion databases ($D = 4$). Information about the number of distorted images in each dataset is provided in Table 2.2. All distorted images in each database, regardless of distortion type, have been used for the computation of PLCC and SRCC values.

While determining the overall performance, we consider the 43 individual FR methods (Table 2.3) and the 22 fused FR methods (Table 2.10) together, in order to observe if fused FR methods offer any benefits over individual methods, and if so, then by how much. Table 2.13 depicts the overall performance of the 65 methods in terms of weighted average PLCC and SRCC, where parts 1, 2, and 3 of the table correspond to the cases of all databases, single distortion databases, and multiple distortion databases, respectively. Within each case, the methods have been sorted in the descending order with respect to the weighted average PLCC and SRCC values. Therefore, the best performing methods for each case are towards the top of the table, while methods at the bottom of the table have the worst performance for that case. The names of the fused FR methods are mentioned in bold, in order to distinguish them from the individual FR methods.

2.4.6 Statistical Significance Testing

We carried out statistical significance testing in accordance with the description given in Section 2.4.1. First, a Kurtosis based check for Gaussianity was performed on the prediction residuals of all 65 individual and fused FR methods on all the datasets. The outcome of this test is presented in Table 2.14, where a “1” means that the kurtosis of the residuals is between 2 and 4, while a “0” means that it is outside of this range. The prediction residuals are assumed to be Gaussian in the former case, while they are not in the latter. While doing this test, all distorted images within each dataset were considered. It can be seen from Table 2.14 that the kurtosis based assumption of Gaussianity of prediction

Table 2.13: Weighted Average PLCC and SRCC values of individual and fused FR methods. Fused FR Methods are highlighted in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
FR Method	PLCC	FR Method	SRCC	FR Method	PLCC	FR Method	SRCC	FR Method	PLCC	FR Method	SRCC
RAS5	0.8985	**RAS4**	0.8907	**RAS5**	0.9054	**RAS5**	0.9003	**RAS7**	0.9199	**RAS7**	0.9106
RAS6	0.8979	**RAS5**	0.8903	**RAS4**	0.9034	**RAS4**	0.8992	**RAS6**	0.9136	**RAS6**	0.9016
RAS4	0.8977	**RAS1**	0.8783	**RAS_MMF4**	0.8988	**RAS2**	0.8909	**RAS3**	0.9117	**RAS3**	0.8976
RAS_MMF4	0.8967	**RAS2**	0.8777	**RAS1**	0.8969	**RAS1**	0.8872	VIF	0.9064	VIF	0.8925
RAS7	0.8935	**RAS_MMF4**	0.8771	**RAS2**	0.8965	VSI	0.8847	**RAS_B1**	0.8990	**RAS_B1**	0.8801
RAS_MMF3	0.8912	**RAS6**	0.8761	**RAS_MMF3**	0.8925	**RAS_MMF4**	0.8783	IWSSIM	0.8970	IWSSIM	0.8785
RAS3	0.8911	**RAS_MMF3**	0.8724	**RAS6**	0.8905	**MMF1**	0.8773	**RAS_MMF1**	0.8942	**RAS_MMF1**	0.8778
RAS1	0.8908	**RAS7**	0.8710	**RAS_MMF2**	0.8879	QASD	0.8741	**RAS_MMF4**	0.8925	**RAS_MMF4**	0.8745
RAS_MMF2	0.8888	**RAS_MMF2**	0.8690	VSI	0.8855	**RAS_MMF3**	0.8735	**CISI**	0.8921	**RAS4**	0.8728
RAS_B1	0.8881	**RAS3**	0.8665	**MMF1**	0.8848	FSIMc	0.8700	**RAS_MMF2**	0.8908	**CISI**	0.8723
RAS2	0.8879	**RAS_B1**	0.8653	**RAS_B2**	0.8832	**RAS_MMF2**	0.8680	**RAS_B2**	0.8887	**RAS_MMF2**	0.8710
RAS_B2	0.8850	**CISI**	0.8634	QASD	0.8830	**MMF3**	0.8641	**RAS_MMF3**	0.8885	**RAS_MMF3**	0.8701
CISI	0.8831	VSI	0.8631	**RAS_B1**	0.8830	**RAS6**	0.8640	**RAS4**	0.8859	**RAS5**	0.8693
IWSSIM	0.8787	FSIMc	0.8628	**RAS3**	0.8813	**MMF4**	0.8634	**RAS5**	0.8841	**RAS_B2**	0.8684
RAS_MMF1	0.8786	**RAS_B2**	0.8607	**RAS7**	0.8810	**CISI**	0.8592	DVICOM	0.8799	DVICOM	0.8634
FSIMc	0.8785	IWSSIM	0.8559	FSIMc	0.8809	**RAS_B1**	0.8583	DVICOM_F	0.8794	DVICOM_F	0.8631
DSS	0.8757	SFF	0.8527	**CISI**	0.8788	SFF	0.8572	**RAS1**	0.8781	DSS	0.8630
VSI	0.8707	DSS	0.8520	**MMF4**	0.8777	**RAS_B2**	0.8570	DSS	0.8774	**RAS1**	0.8596
MCSD	0.8705	QASD	0.8482	ESSIM	0.8754	CID_MS	0.8536	VIF_DWT	0.8757	VIF_DWT	0.8564
FSIM	0.8687	**MMF1**	0.8479	DSS	0.8749	**RAS7**	0.8523	FSIMc	0.8735	IFC	0.8530
ESSIM	0.8674	MCSD	0.8464	MCSD	0.8724	**RAS3**	0.8517	FSIM	0.8721	**RAS2**	0.8500
GMSD	0.8671	**RAS_MMF1**	0.8452	MAD	0.8719	**HFSIMc**	0.8509	GMSD	0.8710	FSIMc	0.8479
SFF	0.8658	**MMF4**	0.8449	**RAS_MMF1**	0.8712	ESSIM	0.8493	**RAS2**	0.8698	GMSD	0.8458
DVICOM	0.8631	CID_MS	0.8445	IWSSIM	0.8700	**CNNM**	0.8490	MCSD	0.8666	FSIM	0.8452
DVICOM_F	0.8631	GMSD	0.8433	**MMF3**	0.8697	MCSD	0.8484	SSIM_DWT	0.8650	SFF	0.8433
QASD	0.8625	FSIM	0.8430	**HFSIMc**	0.8696	DSS	0.8467	SFF	0.8643	MCSD	0.8422
SRSIM	0.8616	ESSIM	0.8418	FSIM	0.8671	IWSSIM	0.8452	WSSI	0.8608	SSIM_DWT	0.8385
MMF4	0.8602	DVICOM_F	0.8394	SFF	0.8665	GMSD	0.8421	DWT_VIF	0.8598	DWT_VIF	0.8368
MMF1	0.8593	**MMF3**	0.8392	SRSIM	0.8654	FSIM	0.8419	IFC	0.8567	WSSI	0.8338
MMF3	0.8569	DVICOM	0.8387	GMSD	0.8653	MAD	0.8413	SRSIM	0.8535	VIF_P	0.8336
GSIM	0.8553	SRSIM	0.8347	GSIM	0.8619	GSIM	0.8401	VIF_P	0.8514	SRSIM	0.8264
HFSIMc	0.8550	**HFSIMc**	0.8345	**CNNM**	0.8595	**MMF2**	0.8399	ESSIM	0.8506	ESSIM	0.8259
MSSSIM	0.8537	CID_SS	0.8325	**MMF2**	0.8592	MSSSIM	0.8386	MSSSIM	0.8440	CID_MS	0.8253
ADM	0.8536	MSSSIM	0.8323	ADM	0.8590	SRSIM	0.8386	CID_SS	0.8434	CID_SS	0.8200
MAD	0.8516	ADM	0.8308	MSSSIM	0.8583	CID_SS	0.8385	ADM	0.8423	MSSSIM	0.8191
CID_MS	0.8511	GSIM	0.8307	CID_MS	0.8570	ADM	0.8384	GSIM	0.8414	VSI	0.8177
CID_SS	0.8452	**CNNM**	0.8270	DVICOM_F	0.8553	PSNR_HAc	0.8361	VSI	0.8395	ADM	0.8149
SSIM_DWT	0.8436	**MMF2**	0.8248	DVICOM	0.8551	PSNR_HMAc	0.8352	CID_MS	0.8386	GSIM	0.8111
MMF2	0.8434	MAD	0.8220	CID_SS	0.8460	**RAS_MMF1**	0.8297	**MMF3**	0.8298	**MMF4**	0.8060
CNNM	0.8389	SSIM_DWT	0.8137	PSNR_HAc	0.8425	DVICOM_F	0.8281	**HFSIMc**	0.8243	**HFSIMc**	0.7999
VIF	0.8388	WSSI	0.8069	RFSIM	0.8393	DVICOM	0.8270	**MMF4**	0.8236	QASD	0.7936
WSSI	0.8384	SSIM	0.8029	PSNR_HMAc	0.8391	RFSIM	0.8112	SSIM	0.8230	**MMF2**	0.7930
SSIM	0.8271	PSNR_HMAc	0.8028	SSIM_DWT	0.8335	SSIM	0.8080	QASD	0.8195	SSIM	0.7923
VIF_DWT	0.8220	VIF	0.8024	PSNR_HA	0.8315	PSNR_HA	0.8057	**MMF2**	0.8100	**MMF3**	0.7868
PSNR_HMAc	0.8153	PSNR_HAc	0.7942	SSIM	0.8290	SSIM_DWT	0.8020	MAD	0.8089	**MMF1**	0.7859
PSNR_HAc	0.8094	VIF_DWT	0.7768	WSSI	0.8278	PSNR_HMA	0.7952	**MMF1**	0.8057	MAD	0.7812
PSNR_HMA	0.8080	PSNR_HMA	0.7762	PSNR_HMA	0.8219	WSSI	0.7941	**CNNM**	0.7955	**CNNM**	0.7808
PSNR_HA	0.8061	**CM4**	0.7758	VIF	0.8066	**CM4**	0.7743	UQI	0.7875	**CM4**	0.7791
VIF_P	0.8059	PSNR_HA	0.7747	VIF_DWT	0.7965	**CM3**	0.7682	PSNR_HMA	0.7789	UQI	0.7673
RFSIM	0.8055	RFSIM	0.7740	VIF_P	0.7843	VIF	0.7596	PSNR_HMAc	0.7653	PSNR_HMA	0.7363
DWT_VIF	0.8032	**CM3**	0.7575	DWT_VIF	0.7763	IW_PSNR	0.7501	IW_PSNR	0.7652	**CM3**	0.7350
PSNR_HVS	0.7402	DWT_VIF	0.7531	VSNR	0.7492	VSNR	0.7410	PSNR_HA	0.7527	PSNR_HMAc	0.7347
PSNR_HVSM	0.7364	VIF_P	0.7526	PSNR_HVS	0.7467	VIF_DWT	0.7390	PSNR_HVSM	0.7466	IW_PSNR	0.7306
VSNR	0.7335	IW_PSNR	0.7438	PSNR_DWT	0.7323	PSNR_HVS	0.7295	PSNR_HAc	0.7399	PSNR_HA	0.7095
IW_PSNR	0.7263	VSNR	0.7174	PSNR_HVSM	0.7315	VIF_P	0.7142	RFSIM	0.7343	PSNR_HAc	0.7060
PSNR_DWT	0.7244	PSNR_HVS	0.7136	PSNR	0.7180	PSNR_HVSM	0.7141	PSNR_HVS	0.7264	PSNR_HVSM	0.7010
UQI	0.7226	PSNR_HVSM	0.7099	NQM	0.7136	DWT_VIF	0.7134	AD_DWT	0.7087	RFSIM	0.6958
NQM	0.7022	PSNR_DWT	0.6944	IW_PSNR	0.7078	PSNR_DWT	0.7076	PSNR_DWT	0.7076	AD_DWT	0.6924
PSNR	0.6927	IFC	0.6924	SNR	0.7043	PSNR	0.7066	VSNR	0.7003	PSNR_HVS	0.6801
CM3	0.6893	AD_DWT	0.6875	UQI	0.6918	NQM	0.6949	**CM3**	0.6927	VSNR	0.6677
WSNR	0.6820	UQI	0.6829	**CM3**	0.6876	SNR	0.6923	**CM4**	0.6908	PSNR_DWT	0.6665
SNR	0.6819	NQM	0.6801	WSNR	0.6824	AD_DWT	0.6852	WSNR	0.6813	NQM	0.6488
CM4	0.6768	PSNR	0.6720	**CM4**	0.6701	WSNR	0.6717	NQM	0.6782	WSNR	0.6341
AD_DWT	0.6054	SNR	0.6606	AD_DWT	0.5563	UQI	0.6428	PSNR	0.6396	PSNR	0.5992
IFC	0.5789	WSNR	0.6596	IFC	0.4470	IFC	0.6161	SNR	0.6347	SNR	0.5938

residuals holds in most cases (around 82% cases). Next, the prediction residuals of all methods were compared by making all possible pairs of individual and fused FR methods, and carrying out hypothesis testing through the one-sided (left-tailed) two-sample F-test at 95% confidence (see Section 2.4.1).

Table 2.15 provides the outcome of statistical significance testing for 16 of the 22 fused FR methods. These methods include all four methods belonging to the *empirical fusion* category (HFSIMc [129], CISI [126], CM3 [127], and CM4 [127]). We include both methods of the *learning based fusion* category (CNNM [128] and MMF [130, 133]). As discussed in Section 2.4.3, we tested four versions of MMF. Here, we include the top three MMF versions that have the highest weighted average PLCC for the *All Databases* case (see Table 2.13). These versions are MMF1, MMF3, and MMF4. Of the 13 versions of RAS [41], which belongs to the *rank aggregation based fusion* category, we selected the following eight versions: Among the seven RAS versions found through the exhaustive search procedure in Section 2.4.3 and listed in Table 2.9, the top four RAS versions that have the highest weighted average PLCC for the *All Databases* case (see Table 2.13) were selected. These versions include RAS4, RAS5, RAS6, and RAS7. The three RAS versions corresponding to the MMF versions included above were also selected (RAS_MMF1, RAS_MMF3, and RAS_MMF4). Finally, RAS_B1, which is one of the original RAS versions in [41] is included as well.

Table 2.16 provides the outcome of statistical significance testing for 14 of the 43 individual FR methods. These methods were selected by analyzing the weighted average PLCC of the *All Databases* case in Table 2.13 and picking the top performing methods such that: A) The overall top four methods are selected which include IWSSIM [13], FSIMc [14], DSS [16], and VSI [15], all of which are *structural similarity based* approaches. B) There is representation from each of the four categories of individual FR methods discussed in Section 2.3.1. PSNR is selected from the *error based methods* category. In addition to the four top performing methods (IWSSIM, FSIMc, DSS, and VSI), three additional methods, CID_MS [95], ESSIM [98], and GMSD [99] are selected from the *structural similarity based methods* category. VIF [113], SFF [109], and QASD [107] represent the *NSS based methods* category. Finally, ADM [94], MAD [26], and DVICOM_F [96], represent the *mixed strategy based methods* category. To help statistically compare individual FR methods with fused

Table 2.14: Kurtosis based check for Gaussianity of prediction residuals of individual/fused FR Methods. Fused FR Methods are highlighted in bold.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
RAS5	1	1	0	1	1	1	1	1	1	1
RAS6	1	1	1	1	1	1	1	1	1	1
RAS4	1	1	1	1	1	1	1	1	1	1
RAS_MMF4	0	0	0	1	1	1	1	1	1	1
RAS7	1	1	1	1	1	1	0	1	1	1
RAS_MMF3	1	0	0	1	1	1	0	1	1	1
RAS3	1	1	1	1	1	1	1	1	1	1
RAS1	0	1	0	1	1	1	1	1	1	1
RAS_MMF2	1	0	0	1	1	1	1	1	1	1
RAS_B1	0	1	1	1	1	1	1	1	1	1
RAS2	1	1	0	1	1	1	1	1	1	1
RAS_B2	0	0	1	1	1	1	1	1	1	1
CISI	1	1	0	1	1	1	1	1	1	1
IWSSIM	0	0	1	1	1	1	1	1	1	1
RAS_MMF1	1	0	1	1	0	1	1	1	1	1
FSIMc	0	0	0	1	1	1	1	1	1	1
DSS	0	0	1	0	1	0	1	1	1	0
VSI	1	0	0	1	1	1	1	1	1	1
MCSD	0	0	0	1	1	0	1	1	1	0
FSIM	0	0	1	1	1	1	1	1	1	1
ESSIM	0	0	0	1	1	1	1	1	1	1
GMSD	0	0	0	1	1	0	1	1	1	0
SFF	1	1	0	1	1	1	0	1	1	1
DVICOM	1	0	1	1	0	1	1	1	1	1
DVICOM_F	1	0	1	1	0	1	1	1	1	1
QASD	1	0	0	1	1	1	1	1	1	1
SRSIM	1	0	1	1	1	1	1	1	1	1
MMF4	1	0	1	1	1	1	0	1	1	1
MMF1	0	0	0	1	1	1	1	1	1	0
MMF3	0	0	1	1	1	1	1	1	1	1
GSIM	1	0	0	1	1	1	1	1	1	1
HFSIMc	1	0	0	1	1	1	1	1	1	1
MSSSIM	1	0	1	1	1	1	1	1	1	1
ADM	0	0	1	1	1	1	1	1	1	1
MAD	0	0	1	1	1	1	1	1	1	1
CID_MS	1	1	1	1	1	1	1	1	1	1
CID_SS	1	1	1	1	1	1	1	1	1	1
SSIM_DWT	1	1	1	1	0	1	0	1	1	1
MMF2	0	0	0	1	1	1	1	1	1	1
CNNM	0	0	0	1	1	0	0	1	1	1
VIF	1	1	1	1	1	1	1	1	1	1
WSSI	1	1	1	1	0	1	0	1	1	1
SSIM	1	1	0	1	1	1	1	1	1	1
VIF_DWT	1	1	1	1	1	1	1	1	1	1
PSNR_HMAc	1	0	1	1	1	1	1	1	1	1
PSNR_HAc	1	0	1	1	1	1	1	1	1	1
PSNR_HMA	0	0	1	0	1	1	1	1	1	1
PSNR_HA	1	0	1	1	1	1	1	1	1	1
VIF_P	1	1	1	1	1	1	0	1	1	1
RFSIM	1	1	0	1	1	1	1	1	1	1
DWT_VIF	1	1	1	1	1	1	0	1	1	1
PSNR_HVS	1	0	1	0	1	1	1	1	1	1
PSNR_HVSM	0	0	1	1	1	1	1	1	1	1
VSNR	1	0	0	1	1	1	1	1	1	1
IW_PSNR	0	0	1	1	1	1	1	1	1	1
PSNR_DWT	0	0	1	1	1	1	1	1	1	1
UQI	1	1	1	1	1	1	1	1	1	1
NQM	0	0	1	1	1	1	1	1	1	0
PSNR	1	0	1	1	1	1	1	1	1	1
CM3	1	0	1	1	1	1	1	1	0	1
WSNR	0	0	1	1	1	1	1	1	1	1
SNR	1	0	1	1	1	1	1	1	1	1
CM4	1	0	1	1	1	1	1	1	1	1
AD_DWT	0	1	0	1	1	1	1	1	1	1
IFC	1	1	1	1	1	1	1	1	1	1

Table 2.15: Statistical significance testing results of **fused** FR methods based on prediction residuals. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence).

	RAS5	RAS6	RAS4	RAS_MMF4	RAS7	RAS_MMF3	RAS_B1	CISI	RAS_MMF1	MMF4	MMF1	MMF3	HFSIMc	CNNM	CM3	CM4
RAS5	__________	01000_00__	_1________	_1_0______	010_0100__	010___0___	010__100__	_1___10___	010__1_0__	10_1111111	1011111_11	1011111_11	_11_1111_1	1011111111	1110111111	1110111111
RAS6	10111_11__	__________	101_1_11__	101_1_11__	_1____0___	101_1_11__	__11111_11	1_111111_1	_1_1111___	1011111111	1011111111	1011111111	1011111111	1011111111	111_111111	111_111111
RAS4	_0________	010_0_00__	__________	_10_______	010_0100__	010___0___	01011100_1	_1__110__1	0101_1_0__	10_1111111	1011111_11	10_1111_11	_1_1111111	1011111111	111_111111	111_111111
RAS_MMF4	_0_1______	010_0_00__	_01_______	__________	010_0100__	_10_______	0101_10__1	_1_1_10__1	0101_1____	10_1111111	1011111111	1011111111	_1111111_1	1011111111	111_111111	111_111111
RAS7	101_1011__	_0____1___	101_1011__	101_1011__	__________	101_1011__	_0111_1_11	10111111_1	__11111___	1011111111	1011111111	1011111111	1011111111	1011111111	111_111111	111_111111
RAS_MMF3	101___1___	010_0_00__	101___1___	_01_______	010_0100__	__________	0__111_0__	__1_110___	0101_1_0__	1011111111	1011111_11	1011111_11	1_111111_1	1011111_11	111_111111	111_111111
RAS_B1	101__011__	__00000_00	10100011_0	1010_01__0	_1000_0_00	1__000_1__	__________	1_1____1__	_10_0_1___	1011111111	1011111111	1011111111	1_1_1_11_1	1011111111	1110111111	1110111111
CISI	_0___01___	0_000000_0	_0__001__0	_0_0_01__0	01000000_0	__0_001___	0_0____0__	__________	010_0_10__	10_1111111	1011111_11	1011111_11	____1_11__	1011111_11	1110111111	1110111111
RAS_MMF1	101__0_1__	_0_0000___	1010_0_1__	1010_0____	__00000___	1010_0_1__	_01_1_0___	101_1_01__	__________	1011111111	1011111111	1011111111	101_1_11_1	1011111111	1110111111	1110111111
MMF4	01_0000000	0100000000	01_0000000	01_0000000	0100000000	0100000000	0100000000	01_0000000	0100000000	__________	1_1_001001	__11___0_1	0110001_00	_11__11_01	1110__1_11	1110__1_0_
MMF1	0100000_00	0100000000	0100000_00	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	0_0_110110	__________	0___110_1_	0100_00100	011_11__11	11101_111_	11101111__
MMF3	0100000_00	0100000000	01_0000_00	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	__00___1_0	1___001_0_	__________	01_0001100	_110__1__1	1110__111_	1110__11__
HFSIMc	_00_0000_0	0100000000	_0_0000000	_0000000_0	0100000000	0_000000_0	0_0_0_00_0	____0_00__	010_0_00_0	1001110_11	1011_11011	10_1110011	__________	101111_011	1110111_11	1110111_11
CNNM	0100000000	0100000000	0100000000	0100000000	0100000000	0100000_00	0100000000	0100000_00	0100000000	_00__00_10	100_00__00	_001__0__0	010000_100	__________	1110_01110	1110_01__0
CM3	0001000000	000_000000	000_000000	000_000000	000_000000	000_000000	0001000000	0001000000	0001000000	0001__0_00	00010_000_	0001__000_	0001000_00	0001_10001	__________	1_______0_
CM4	0001000000	000_000000	000_000000	000_000000	000_000000	000_000000	0001000000	0001000000	0001000000	0001__0_1_	00010000__	0001__00__	0001000_00	0001_10__1	0_______1_	__________

Table 2.16: Statistical significance testing results of **individual** FR methods based on prediction residuals. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). Fused FR Methods RAS6 and MMF1 are included for comparison and are highlighted in bold.

	RAS6	IWSSIM	FSIMc	DSS	VSI	ESSIM	GMSD	SFF	DVICOM_F	QASD	**MMF1**	ADM	MAD	CID_MS	VIF	PSNR
RAS6	__________	11111_10_1	101_1_1111	1__1111__1	101_111111	1011111111	1011111_11	10_1111_11	0111111_11	1011111111	1011111111	11111111_1	__111_1111	111_111111	1111110__1	1111111111
IWSSIM	00000_01_0	__________	00_011_1__	000_1111__	_000111111	_0__11111_	000_111_11	0001111111	01__11_111	_00_1111_1	10_1111111	1_0_1111__	000111111_	1_10_1111_	0101110__1	1111111111
FSIMc	010_0_0000	11_100_0__	__________	_10___10__	10____1__1	_1_10_1___	_101_110_1	__01_110_1	01_101_0_1	_00__11__1	10_1_11_11	1101__1___	0101__10__	111_011___	_1_1_100_1	1111111_11
DSS	0__0000__0	111_0000__	_01___01__	__________	101_1__111	1011__11__	__1__1__11	_0_1_11___	011_010__1	_01_1_11_1	1011111111	111___11__	0_11__11__	11100_11__	_111110__1	1111111111
VSI	010_000000	_111000000	01____0__0	010_0__000	__________	01_10_10_0	0101_1_0__	010101_0_0	01_10_00__	010___1___	10_1_11011	11_10_1_00	01010_10_0	111_0_1__0	01_1_1000_	1111111_11
ESSIM	0100000000	_1__00000_	_0_01_0___	0100__00__	10_01_01_1	__________	__0_1100_1	0_01_100_1	01____00_1	_0001_1__1	10_1111_11	11______0_	010___1___	11100_____	_1_11100_1	1111111111
GMSD	0100000_00	111_000_00	_010_001_0	__0__0__00	1010_0_1__	__1_0011_0	__________	_0010_1__0	011_000___	_01__011__	1011__1111	111_001100	01__0011_0	11100011_0	_111_10_0_	1111111111
SFF	01_0000_00	1110000000	__10_001_0	_1_0_00___	101010_1_1	1_10_011_0	_1101_0__1	__________	0110__0___	10101011__	10101_1111	1110_01100	0110_011__	11100011__	_1101100_1	1110111111
DVICOM_F	1000000_00	10__00_000	10_010_1_0	100_101__0	10_01_11__	10____11_0	100_111___	1001__1___	__________	100_1_11__	10_1111111	100___1100	100__011_0	10100_11_0	11_1110___	1111111111
QASD	0100000000	_11_0000_0	_11__00__0	_10_0_00_0	101___0___	_1110_0__0	_10__100__	01010100__	011_0_00__	__________	1011_1__11	111_0_0__0	01_10010_0	11100_0___	_111_100__	1111111_11
MMF1	0100000000	01_0000000	01_0_00_00	0100000000	01_0_00100	01_0000_00	0100__0000	01010_0000	01_0000000	0100_0__00	__________	01_0000_00	0100001000	0110000_00	01_0_10000	_111111111
ADM	00000000_0	0_1_0000__	0010__0___	000___00__	00_01_0_11	00______1_	000_110011	0001_10011	011___0011	000_1_1__1	10_1111_11	__________	000___10__	1_100_____	01_11100_1	1111111_11
MAD	__000_0000	111000000_	1010__01__	1_00__00__	10101_01_1	101___0___	10__1100_1	1001_100__	011__100_1	10_01101_1	1011110111	111___01__	__________	1_100101__	111_1100_1	1111111111
CID_MS	000_000000	0_01_0000_	000_100___	00011_00__	000_1_0__1	00011_____	00011100_1	00011100__	01011_00_1	00011_1___	1001111_11	0_011_____	0_011010__	__________	01011100_1	1111111_11
VIF	0000001__0	1010001__0	_0_0_011_0	_000001__0	10_0_0111_	_0_00011_0	_000_01_1_	_0010011_0	00_0001___	_000_011__	10_1_01111	10_00011_0	000_0011_0	10100011_0	__________	11111_1111
PSNR	0000000000	0000000000	0000000_00	0000000000	0000000_00	0000000000	0000000000	0001000000	0000000000	0000000_00	_000000000	0000000_00	0000000000	0000000_00	00000_0000	__________

ones, two fused FR methods are also included in Table 2.16. These include MMF1 and RAS6, as representatives of learning based and rank aggregation based fusion, respectively.

Each entry in Tables 2.15 and 2.16 is a codeword composed of ten symbols. Each symbol represents the outcome of statistical significance testing for one IQA database. The location of the symbol in the codeword represents specific IQA databases in the following order, from left to right: LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ at viewing distance of 50 cm (CIDIQ50), CIDIQ at viewing distance of 100 cm (CIDIQ100), MDID, MDID2013, LIVE MD, and MDIVL. Each symbol can take one of three possibilities, a “1”, “_”, or “0” meaning that the method in the row is statistically (with 95% confidence) better, indistinguishable, or worse than the method in the column respectively, for a particular database. The order of methods in Tables 2.15 and 2.16 is based on their order in the weighted average PLCC portion of the *All Databases* case in Table 2.13.

2.4.7 Computational Complexity

The computational complexity of all 43 FR IQA methods under test was evaluated in terms of their execution time to determine the quality of a 1024 × 1024 color image on a Lenovo laptop computer with a 2.4GHz Intel Core i7-4700MQ processor, 12GB of RAM, Samsung 850 EVO Solid State Drive, and Windows 10 Home operating system. The execution times of all FR methods are given in Table 2.17, where methods have been sorted in ascending order with respect to execution time. Since PSNR is the fastest method, we also provide the execution time relative to PSNR for convenience in comparison.

2.4.8 Analysis and Discussion

Based on the results obtained in the previous sub-sections and in particular on Table 2.13 (Overall performance) and Tables 2.15 and 2.16 (Statistical Significance Testing), the following observations can be made.

Table 2.17: Execution time of individual FR methods for a test image. Methods are sorted in ascending order with respect to the execution time.

FR Method	Execution Time (Seconds)	Execution Time (Relative to PSNR)
PSNR	0.0044 s	1.00
SNR	0.0107 s	2.43
GMSD	0.0293 s	6.66
SRSIM	0.0309 s	7.02
SSIM	0.0462 s	10.50
MCSD	0.0475 s	10.80
GSIM	0.0510 s	11.59
RFSIM	0.0578 s	13.14
ESSIM	0.1230 s	27.95
UQI	0.1652 s	37.55
MSSSIM	0.1958 s	44.50
WSNR	0.2002 s	45.50
SFF	0.2173 s	49.39
DSS	0.2196 s	49.91
WSSI	0.2402 s	54.59
VIF_DWT	0.2527 s	57.43
VIF_P	0.2546 s	57.86
VSI	0.2727 s	61.98
DWT_VIF	0.2851 s	64.80
FSIM	0.3210 s	72.95
FSIMc	0.3210 s	72.95
VSNR	0.3861 s	87.75
SSIM_DWT	0.4473 s	101.66
ADM	0.5897 s	134.02
PSNR_DWT	0.6578 s	149.50
AD_DWT	0.7790 s	177.05
NQM	0.9878 s	224.50
IW_PSNR	1.4777 s	335.84
IWSSIM	1.5670 s	356.14
PSNR_HVS	2.2685 s	515.57
PSNR_HVSM	2.2685 s	515.57
DVICOM_F	2.3753 s	539.84
CID_SS	2.7699 s	629.52
QASD	2.9208 s	663.82
CID_MS	3.1687 s	720.16
PSNR_HA	3.2619 s	741.34
PSNR_HMA	3.2619 s	741.34
VIF	4.5277 s	1029.02
IFC	4.5797 s	1040.84
MAD	5.4482 s	1238.23
DVICOM	6.9084 s	1570.09
PSNR_HAc	9.7606 s	2218.32
PSNR_HMAc	9.7606 s	2218.32

Individual FR Methods

Considering the top ten methods in each category and for each evaluation criterion in Table 2.13, it can be seen that most top performing methods, especially for the *all databases* and *single distortion databases* categories, belong to the structural similarity based class of FR methods. These methods include IWSSIM [13], FSIMc/FSIM [14], DSS [16], VSI [15], GMSD [99], MCSD [102], ESSIM [98], and CID_MS [95]. For these categories, the sparsity based NSS methods QASD [107] and SFF [109], and the mixed strategy based methods DVICOM [96] and MAD [26] also do well. For the *multiple distortion databases* category, the NSS methods VIF [113] and VIF_DWT [93], and the mixed strategy based method DVICOM/DVICOM_F [96], do well in addition to the structural similarity based approaches. It can be observed from Table 2.13 that error based FR methods do not offer competitive performance against other IQA design philosophies. From Table 2.13 it can be seen that overall: 1) For the *all databases* case, IWSSIM [13] is the top performing method in terms of weighted average PLCC, while VSI [15] is the top performer in terms of weighted average SRCC, 2) For the *single distortion databases* case, VSI [15] is the top performing method both in terms of weighted average PLCC and SRCC, and 3) For the *multiple distortion databases* case, VIF [113] is the top performer both in terms of weighted average PLCC and SRCC.

While using weighted average PLCC and SRCC is one way to determine overall performance, it has the drawback of favoring larger databases. Thus, in our case, the TID2013 database [19] is given the largest weight since it has the most images, while the MDID2013 database [32] is given the smallest weight. This is done even though both these databases contain entirely different distortion processes, where TID2013 contains images afflicted with a single distortion, while MDID2013 has images that have undergone three kinds of distortions. It is thus unfair to develop an opinion solely on the basis of weighted average PLCC and SRCC. Another way to compare methods and to determine which one is performing better than others, is to observe the statistical significance testing tables. From Table 2.16, we can observe that IWSSIM [13] is statistically better than most other methods on most of the databases. This shows that IWSSIM is a robust method that does well across different kinds of distortion types. The FR methods DSS [16] and FSIMc [14] follow IWSSIM in performance and do quite well when statistically compared to other FR methods.

Fused FR Methods

For all three cases of *all databases*, *single distortion databases*, and *multiple distortion databases*, it is clear from Table 2.13 that the rank aggregation based FR fusion technique RAS [41] significantly outperforms all other FR fusion techniques, in terms of both weighted average PLCC and SRCC. The same conclusion can be comprehensively drawn from Table 2.15, where it can be seen that RAS based methods are statistically superior than all other fusion based methods for the vast majority of datasets. Among the 13 RAS methods, it can be observed from Table 2.15 that statistically, RAS6 is overall the top performer, followed closely by RAS7. It can also be observed that RAS methods selected through the exhaustive search procedure described in Section 2.4.3, especially those belonging to the *medium* and *full* sets (Table 2.8) such as RAS6 and RAS7 respectively, perform better than the FR methods combination described in the original RAS work [41], thereby highlighting the importance of finding the set of FR methods to be fused through a more structured approach.

The two learning based fusion approaches, MMF [130] and CNNM [128] do not appear to be competitive when compared to the rank aggregation based approach, as can be seen from Table 2.13. It can be observed from Table 2.15 that the different MMF approaches (MMF1, MMF3, and MMF4) and CNNM perform better than the different RAS methods only on the TID2013 database [19]. However, as described in Section 2.3.2, the MMF methods and CNNM, are all trained on this very database, and hence comparing these methods with other approaches on TID2013 is unreliable and unfair. On all other datasets, the MMF methods and CNNM are statistically outperformed by the RAS methods, which shows that learning based fusion approaches suffer from model overfitting issues. Since the four RAS methods RAS_MMF1, RAS_MMF2, RAS_MMF3, and RAS_MMF4 combine the same set of individual FR methods as the four MMF methods MMF1, MMF2, MMF3, and MMF4, respectively (see Table 2.10), the two FR fusion approaches can be directly compared. Since the TID2013 database was used to train the four MMF methods and it contributes the largest weight to the weighted average PLCC and SRCC computation, we avoid using these evaluation criteria. Instead we statistically compare these methods by using Table 2.15, where it can be seen that the MMF based methods are outperformed by their RAS counterparts on all datasets except TID2013. This again highlights the superiority of the rank aggregation based fusion, which does not involve any training, and hence does not suffer from model overfitting issues. By contrast, the learning based fusion approaches, even when they use one of the largest subject-rated dataset for training, suffer from overfitting issues because the number of distorted images per distortion type are quite small even in the TID2013 database [19] (only 125 images per distortion type).

The empirical fusion based methods CM3 and CM4 [127], described in Section 2.3.2 and given in Equations 2.11 and 2.12 respectively, perform inadequately, even for multiply distorted content for which they are designed, as is evident from Part 3 of Table 2.13. This is because of the choice of FR methods that are being fused in CM3 and CM4, especially IFC [101], NQM [103], and VSNR [75], and the way in which exponent values are obtained on a single database (LIVE MD [31]). It can be observed from Tables 2.6 and 2.7 that while IFC, NQM, and VSNR, perform quite well on the LIVE MD database, their performance is lacking on other IQA datasets. This is further substantiated from Table 2.13 in terms of weighted average PLCC and SRCC. However, since the exponent values in Equations 2.11 and 2.12 are only optimized on LIVE MD database, CM3 and CM4 are highly database dependent. Thus, they perform well only on a few datasets (VCLFER [54] and LIVE MD [31]), while their performance on other datasets is inferior as can be observed in Tables 2.11 and 2.12. This highlights the pitfalls of: 1) the empirical fusion based approach which is rather ad hoc, 2) the selection of FR methods to be fused on the basis of a single dataset, and 3) the use of a single dataset for parameter tuning. It can be observed from Tables 2.13 and 2.15 that the empirical fusion based methods HFSIMc [129] and CISI [126], described in Section 2.3.2 and given in Equations 2.9 and 2.10 respectively, perform better than CM3 and CM4. CISI also performs better than HFSIMc. Both these methods, especially CISI, perform statistically better than the learning based fusion methods (MMF and CNNM) as can be observed from Table 2.15. This performance gain is because HFSIMc and CISI, especially the latter, fuse FR methods that perform well across most databases individually as well. However, even these empirical fusion methods cannot outperform rank aggregation based fusion methods.

Individual and Fused FR Methods

When individual and fused FR methods are considered together, the following observations can be made: 1) The rank aggregation based fusion methods (RAS) [41] outperform the best individual FR methods, as can be seen from Table 2.13. This is also evident from Table 2.16 where it is clear that RAS6 performs statistically better than the top performing FR methods on a majority of databases. Although statistical significance testing results for other RAS methods in comparison with individual FR methods have not been provided due to space constraints, they are also found to be statistically superior. 2) The learning based fusion methods, MMF [130] and CNNM [128], are outperformed by the best individual FR methods on datasets that are not involved in training these fusion methods. This can be seen from Table 2.13 in terms of weighted average PLCC and SRCC, and also from Table 2.16 for MMF1 in terms of statistical significance testing (statistical analysis for MMF2, MMF3, MMF4, and CNNM yielded similar conclusions). 3). Of the four empirical fusion methods, CM3 [127], CM4 [127], and HFSIMc [129], are outperformed by the best individual FR methods as can be observed from Table 2.13 in terms of weighted average PLCC and SRCC. The only exception is the empirical fusion method CISI [126], which performs at par with or better than top performing individual FR methods.

It can therefore be concluded that learning based fusion (MMF and CNNM) and empirical fusion techniques (CM3, CM4, HFSIMc), do not generalize very well when tested across a wide variety of IQA datasets, thereby revealing that they suffer from model overfitting and training database dependency issues. Such drawbacks make them less robust to handle unseen data, where they are outperformed by the best individual FR methods. On the other hand, the rank aggregation based fusion methods (RAS), perform better than other fusion techniques, but more importantly, they outperform the best individual FR methods across the entire range of IQA datasets used. Since these methods are completely training-free, they do not suffer from model overfitting and database dependence issues, making them truly robust. While it can be seen from Tables 2.6 and 2.7 that the performance of even the top performing FR methods varies, sometimes widely, across different IQA datasets, Tables 2.11 and 2.12 show that such performance variation across different datasets is less pronounced for RAS based methods. It can be concluded that by aggregating the ranks generated from various top performing FR IQA methods, the deficiencies of some methods in the combination are compensated by the strengths of other constituents. These characteristics of rank aggregation based fusion methods make them ideal candidates to annotate large-scale IQA datasets in place of subject ratings. While opinions provided by humans will continue to be the ultimate benchmark when it comes to annotating IQA databases, as we discussed earlier, it is quite impossible to obtain human opinions in adequate numbers for very large-scale datasets. Here, rank aggregation based fusion methods can be used to annotate such large datasets in place of human opinion scores instead of choosing one or the other individual FR method.

2.5 Performance Analysis of NR Methods

To analyze the performance of NR IQA methods, we use the same evaluation criteria as described in Section 2.4.1, and compute the evaluation metrics for two types of data. First, like the performance analysis of FR and fused FR methods in Section 2.4, all images within a database are considered, that is, all distortion types are taken into account while calculating PLCC, SRCC, and performing statistical significance testing. This will be referred to as the *all distortions* category. Second, evaluation metrics are calculated for a subset of distortion types in each dataset, which we shall refer to as the *subset distortions* category. For single distortion databases (LIVE R2 [24], TID2013 [19], CSIQ [26], VCLFER [54], and CIDIQ [5]), we constitute a subset of images belonging to four common distortion types: 1) Noise, 2) Gaussian Blur, 3) JPEG compression, and 4) JPEG2000 compression. It should be noted that the noisy images in the CIDIQ database [5] are afflicted with Poisson noise, while they are afflicted with additive white Gaussian noise in the other four single distortion datasets. However, for the purposes of the subset performance analysis, we do not make a distinction between the two. For multiply distorted databases, we constitute subsets of images by separately calculating evaluation metrics for individual distortion combinations (where possible). This means that we separately consider the Blur-JPEG and Blur-Noise combinations in the LIVE MD database [31], and the Blur-JPEG and

Noise-JPEG combinations in the MDIVL database [34]. Since the MDID2013 database [32] contains only one distortion combination, while the MDID database [33] has many possible distortion combinations due to the random choice of distortions at different stages, the images in these two datasets cannot be split into subsets, and hence the entire datasets will be considered for the subset case as well. The rationale for conducting performance analysis for a subset of distortion types, especially for single distortion databases, stems from the fact that most training-based OA NR models are trained for the above-mentioned common distortion types that are found in almost all single distortion datasets. Thus, these subsets of distortions provide a more fair ground for comparison of these methods. However, we also consider the case of all distortions in each database and do not retrain these NR models on individual databases but use the original versions from the authors, in order to more rigorously test NR methods, as the ultimate goal of NR or *blind* IQA methods is to be robust to *unseen* data. The gap in performance for these two cases should highlight future research directions as well.

2.5.1 Performance of NR Methods

We tested the 14 NR methods discussed in Section 2.3.3 and given in Table 2.5, on each of the nine subject-rated IQA databases mentioned in Table 2.2. Testing was done separately for the two viewing distances in the CIDIQ database [5], where labels of CIDIQ50 and CIDIQ100 correspond to the viewing distances of 50 cm and 100 cm, respectively. For all databases, the test results for the *all distortions* case are given in Table 2.18 in terms of PLCC and in Table 2.19 in terms of SRCC. The test results for the *subset distortions* case are given in Tables 2.20 and 2.21 in terms of PLCC and SRCC respectively. While considering Tables 2.18, 2.19, 2.20, and 2.21, it should be noted that the OA NR methods BIQI [139], BRISQUE [140], NRSL [147], CORNIA [141], HOSA [143], WaDIQaM-NR [148], and MEON [146] are trained on the LIVE R2 database [24,42], and GWHGLBP [142] is trained on the LIVE MD database [31]. Thus, comparing these OA NR methods with other approaches on these respective databases is unreliable and unfair.

The overall performance of the 14 NR methods was determined by using the same approach as in Section 2.4.5. The weighted average PLCC and SRCC were computed for

Table 2.18: PLCC of 14 NR methods on nine subject-rated IQA databases. All distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
BIQI	0.9224	0.4489	0.6796	0.6106	0.3542	0.2563	0.6372	0.0169	0.7389	0.6215
BRISQUE	0.9671	0.4747	0.7006	0.8209	0.2924	0.3257	0.4558	0.1403	0.6045	0.6516
CORNIA	0.9665	0.5715	0.7325	0.8366	0.4496	0.1991	0.7907	0.6935	0.8679	0.8277
dipIQ	0.9348	0.4774	0.7787	0.8942	0.5208	0.2498	0.6738	0.4355	0.7669	0.7627
GWHGLBP	0.8079	0.4973	0.7002	0.6427	0.3653	0.2978	0.7035	0.7430	0.9663	0.5737
HOSA	0.9991	0.5481	0.7240	0.8496	0.4969	0.3761	0.6521	0.2513	0.6768	0.7167
ILNIQE	0.7061	0.5090	0.8024	0.7289	0.2768	0.3003	0.7053	0.5146	0.8923	0.6303
LPSI	0.8280	0.4892	0.7216	0.6020	0.4037	0.3981	0.4336	0.0999	0.5464	0.5715
MEON	0.9389	0.4946	0.7804	0.9221	0.4306	0.3854	0.5168	0.2430	0.2339	0.5722
NIQE	0.9052	0.4001	0.7170	0.8040	0.3703	0.2708	0.6728	0.5571	0.8387	0.5688
NRSL	0.9815	0.5345	0.7413	0.8905	0.4672	0.3069	0.6502	0.3088	0.4829	0.6794
QAC	0.8625	0.4371	0.7067	0.7615	0.3573	0.2856	0.6043	0.4240	0.4145	0.5713
SISBLIM	0.8077	0.3961	0.6945	0.7574	0.4782	0.4532	0.6700	0.8123	0.8948	0.5724
WaDIQaM-NR	0.9341	0.4707	0.7372	0.7862	0.4133	0.3481	0.4215	0.1371	0.2897	0.5213

Table 2.19: SRCC of 14 NR methods on nine subject-rated IQA databases. All distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
BIQI	0.9198	0.3935	0.6186	0.6170	0.3433	0.2353	0.6276	0.0077	0.5556	0.5711
BRISQUE	0.9654	0.3672	0.5563	0.8130	0.3640	0.2496	0.4035	0.2209	0.5018	0.6647
CORNIA	0.9681	0.4288	0.6534	0.8354	0.3727	0.2071	0.7918	0.7055	0.8340	0.8336
dipIQ	0.9378	0.4377	0.5266	0.8957	0.4135	0.2100	0.6612	0.4153	0.6678	0.7131
GWHGLBP	0.7410	0.3844	0.5773	0.6243	0.3337	0.2412	0.7032	0.7555	0.9698	0.5841
HOSA	0.9990	0.4705	0.5925	0.8574	0.4494	0.3248	0.6412	0.2993	0.6393	0.7399
ILNIQE	0.8975	0.4939	0.8144	0.7391	0.2997	0.3127	0.6900	0.5148	0.8778	0.6238
LPSI	0.8181	0.3949	0.5303	0.5865	0.2060	0.1411	0.0306	0.0168	0.2717	0.5736
MEON	0.9409	0.3750	0.7248	0.9215	0.4101	0.2497	0.4861	0.2980	0.1917	0.5466
NIQE	0.9073	0.3132	0.6271	0.8126	0.3458	0.2212	0.6523	0.5451	0.7738	0.5713
NRSL	0.9796	0.4277	0.6750	0.8930	0.4249	0.2894	0.6458	0.4088	0.4145	0.6047
QAC	0.8683	0.3722	0.4900	0.7686	0.3196	0.1944	0.3239	0.2272	0.3579	0.5524
SISBLIM	0.7741	0.3177	0.6603	0.7622	0.4435	0.4098	0.6554	0.8089	0.8770	0.5375
WaDIQaM-NR	0.9417	0.4393	0.6388	0.7524	0.3588	0.2235	0.4040	0.1316	0.2379	0.5614

Table 2.20: PLCC of 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG
BIQI	0.9534	0.7565	0.7968	0.6106	0.4797	0.4900	0.6372	0.0169	0.7743	0.6634	0.7398	0.6035
BRISQUE	0.9760	0.8399	0.9196	0.8209	0.5257	0.3906	0.4558	0.1403	0.8663	0.4596	0.8249	0.6511
CORNIA	0.9715	0.8824	0.9135	0.8366	0.5900	0.5477	0.7907	0.6935	0.8774	0.8723	0.9419	0.7900
dipIQ	0.9559	0.8879	0.9479	0.8942	0.7433	0.6472	0.6738	0.4355	0.8235	0.7897	0.8311	0.7882
GWHGLBP	0.8088	0.7675	0.7839	0.6427	0.5196	0.5345	0.7035	0.7430	0.9677	0.9684	0.7745	0.4943
HOSA	0.9992	0.8858	0.9360	0.8496	0.6504	0.6283	0.6521	0.2513	0.8968	0.6728	0.9005	0.7022
ILNIQE	0.7031	0.8491	0.8143	0.7289	0.3127	0.3892	0.7053	0.5146	0.9048	0.8968	0.8293	0.5759
LPSI	0.8440	0.8114	0.8657	0.6020	0.5509	0.6289	0.4336	0.0999	0.8820	0.1182	0.7959	0.5075
MEON	0.9907	0.8940	0.9334	0.9221	0.6495	0.6379	0.5168	0.2430	0.0995	0.3881	0.3875	0.7405
NIQE	0.9162	0.8091	0.8876	0.8040	0.4694	0.4338	0.6728	0.5571	0.9099	0.8481	0.7996	0.4507
NRSL	0.9887	0.9108	0.9058	0.8905	0.4216	0.4500	0.6502	0.3088	0.3283	0.6263	0.6418	0.7334
QAC	0.8777	0.8051	0.8736	0.7615	0.4512	0.5068	0.6043	0.4240	0.5378	0.6722	0.6765	0.6090
SISBLIM	0.8220	0.7309	0.7967	0.7574	0.5792	0.6741	0.6700	0.8123	0.9030	0.8913	0.8056	0.4871
WaDIQaM-NR	0.9302	0.8983	0.8577	0.7862	0.4600	0.5530	0.4215	0.1371	0.6842	0.4379	0.6415	0.5231

Table 2.21: SRCC of 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each dataset were considered.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG
BIQI	0.9528	0.7763	0.7972	0.6170	0.4976	0.4849	0.6276	0.0077	0.6542	0.4902	0.6591	0.5302
BRISQUE	0.9757	0.8401	0.8992	0.8130	0.4727	0.4771	0.4035	0.2209	0.7923	0.2991	0.7385	0.6612
CORNIA	0.9732	0.8727	0.8987	0.8354	0.5740	0.5053	0.7918	0.7055	0.8278	0.8523	0.9254	0.8027
dipIQ	0.9574	0.8720	0.9290	0.8957	0.7460	0.6433	0.6612	0.4153	0.6979	0.7391	0.6512	0.7730
GWHGLBP	0.7447	0.6538	0.6728	0.6243	0.4768	0.4454	0.7032	0.7555	0.9640	0.9751	0.7584	0.4502
HOSA	0.9991	0.8681	0.9111	0.8574	0.6677	0.6236	0.6412	0.2993	0.8437	0.5357	0.8789	0.7150
ILNIQE	0.9153	0.8417	0.8802	0.7391	0.3669	0.4248	0.6900	0.5148	0.8915	0.8821	0.7915	0.5797
LPSI	0.8333	0.7046	0.7711	0.5865	0.3382	0.3949	0.0306	0.0168	0.8387	0.0012	0.7348	0.4692
MEON	0.9906	0.9012	0.9300	0.9215	0.6421	0.5830	0.4861	0.2980	0.0476	0.3257	0.3255	0.7397
NIQE	0.9168	0.7972	0.8710	0.8126	0.4703	0.4180	0.6523	0.5451	0.8713	0.7938	0.7625	0.4510
NRSL	0.9880	0.8965	0.8874	0.8930	0.5732	0.5564	0.6458	0.4088	0.2634	0.5991	0.4684	0.7125
QAC	0.8857	0.8055	0.8415	0.7686	0.4450	0.4566	0.3239	0.2272	0.3959	0.4707	0.5537	0.5282
SISBLIM	0.7835	0.7703	0.8059	0.7622	0.5565	0.6314	0.6554	0.8089	0.8746	0.8782	0.7584	0.3320
WaDIQaM-NR	0.9399	0.8646	0.8636	0.7524	0.4777	0.4691	0.4040	0.1316	0.5012	0.2502	0.6121	0.4830

three cases: 1) All databases, 2) Only single distortion databases, and 3) Only multiple distortion databases. Table 2.22 depicts the overall performance of the 14 NR methods for *all distortions* in terms of weighted average PLCC and SRCC, where parts 1, 2, and 3 of the table correspond to the cases of all databases, single distortion databases, and multiple distortion databases, respectively. Within each case, the methods have been sorted in the descending order with respect to the weighted average PLCC and SRCC values, where the best performing methods can be found towards the top of the table. Table 2.23 provides the results for *subset distortions*. In both Tables 2.22 and 2.23 we are including results for the FR methods IWSSIM [13] and PSNR for quick comparison. For a thorough comparison of the overall performance of NR methods with that of individual and fused FR methods, these tables should be compared with Table 2.13.

Table 2.22: Weighted Average PLCC and SRCC values of NR methods for *all distortions*. FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC
IWSSIM*	0.8787	**IWSSIM***	0.8559	**IWSSIM***	0.8700	**IWSSIM***	0.8452	**IWSSIM***	0.8970	**IWSSIM***	0.8785
PSNR*	0.6927	**PSNR***	0.6720	**PSNR***	0.7180	**PSNR***	0.7066	CORNIA	0.8006	CORNIA	0.7990
CORNIA	0.6713	CORNIA	0.6147	HOSA	0.6266	ILNIQE	0.5651	GWHGLBP	0.7143	GWHGLBP	0.7184
HOSA	0.6275	ILNIQE	0.6031	NRSL	0.6136	HOSA	0.5641	ILNIQE	0.6945	ILNIQE	0.6830
dipIQ	0.6181	HOSA	0.5851	CORNIA	0.6099	NRSL	0.5499	SISBLIM	0.6937	SISBLIM	0.6749
NRSL	0.6085	dipIQ	0.5620	MEON	0.6026	CORNIA	0.5272	dipIQ	0.6838	dipIQ	0.6491
GWHGLBP	0.5949	NRSL	0.5589	dipIQ	0.5869	MEON	0.5245	NIQE	0.6597	NIQE	0.6392
ILNIQE	0.5919	SISBLIM	0.5408	WaDIQaM-NR	0.5683	dipIQ	0.5207	**PSNR***	0.6396	HOSA	0.6292
SISBLIM	0.5821	GWHGLBP	0.5377	BRISQUE	0.5571	WaDIQaM-NR	0.5203	HOSA	0.6296	**PSNR***	0.5992
NIQE	0.5642	NIQE	0.5181	LPSI	0.5509	BRISQUE	0.4877	NRSL	0.5977	NRSL	0.5780
MEON	0.5570	BIQI	0.5007	ILNIQE	0.5432	BIQI	0.4824	BIQI	0.5837	BIQI	0.5394
BIQI	0.5397	MEON	0.4969	GWHGLBP	0.5382	SISBLIM	0.4770	QAC	0.5503	BRISQUE	0.4614
BRISQUE	0.5360	BRISQUE	0.4792	SISBLIM	0.5291	NIQE	0.4606	BRISQUE	0.4915	MEON	0.4387
QAC	0.5338	WaDIQaM-NR	0.4782	QAC	0.5259	QAC	0.4556	MEON	0.4610	WaDIQaM-NR	0.3896
LPSI	0.5179	QAC	0.4292	NIQE	0.5189	GWHGLBP	0.4518	LPSI	0.4483	QAC	0.3736
WaDIQaM-NR	0.5131	LPSI	0.3558	BIQI	0.5188	LPSI	0.4325	WaDIQaM-NR	0.3970	LPSI	0.1943

*FR Methods included for comparison.

Statistical significance testing was conducted in the same manner as described in Sections 2.4.1 and 2.4.6. The outcome of the kurtosis based check for Gaussianity of prediction residuals is presented in Table 2.24 where a "1" means that the kurtosis of the residuals is between 2 and 4, and they can be assumed to be Gaussian distributed, while a "0" means that the kurtosis of residuals is not between 2 and 4, and they are assumed to be non-Gaussian. Each entry in the table may be composed of more than one symbol, and

Table 2.23: Weighted Average PLCC and SRCC values of NR methods for *subset distortions*. FR Methods IWSSIM and PSNR are included for comparison and are in bold.

Part 1: All Databases				Part 2: Single Distortion Databases				Part 3: Multiple Distortion Databases			
NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC	NR Method	PLCC	NR Method	SRCC
IWSSIM*	0.9116	**IWSSIM***	0.9002	**IWSSIM***	0.9226	**IWSSIM***	0.9179	**IWSSIM***	0.9004	**IWSSIM***	0.8820
CORNIA	0.8088	CORNIA	0.8007	dipIQ	0.8584	dipIQ	0.8527	CORNIA	0.8096	CORNIA	0.8062
dipIQ	0.7805	dipIQ	0.7562	MEON	0.8535	MEON	0.8449	GWHGLBP	0.7269	GWHGLBP	0.7208
HOSA	0.7534	HOSA	0.7438	HOSA	0.8407	HOSA	0.8364	ILNIQE	0.7110	ILNIQE	0.6974
SISBLIM	0.7227	ILNIQE	0.7078	CORNIA	0.8080	NRSL	0.8171	SISBLIM	0.7093	SISBLIM	0.6733
PSNR*	0.7148	**PSNR***	0.7048	NRSL	0.7855	**PSNR***	0.8054	dipIQ	0.7005	dipIQ	0.6571
NIQE	0.7093	SISBLIM	0.7008	**PSNR***	0.7836	CORNIA	0.7954	NIQE	0.6763	NIQE	0.6537
GWHGLBP	0.7073	NRSL	0.6996	BRISQUE	0.7689	BRISQUE	0.7685	HOSA	0.6639	HOSA	0.6488
NRSL	0.6937	NIQE	0.6954	WaDIQaM-NR	0.7653	WaDIQaM-NR	0.7477	**PSNR***	0.6441	**PSNR***	0.6015
ILNIQE	0.6801	GWHGLBP	0.6672	NIQE	0.7415	NIQE	0.7360	NRSL	0.5996	NRSL	0.5790
QAC	0.6637	MEON	0.6441	SISBLIM	0.7359	SISBLIM	0.7276	QAC	0.5944	BIQI	0.5464
MEON	0.6609	BIQI	0.6272	QAC	0.7312	QAC	0.7200	BIQI	0.5918	BRISQUE	0.4756
BIQI	0.6466	BRISQUE	0.6239	LPSI	0.7284	ILNIQE	0.7179	BRISQUE	0.5193	MEON	0.4379
BRISQUE	0.6457	WaDIQaM-NR	0.5786	BIQI	0.6999	BIQI	0.7059	MEON	0.4632	WaDIQaM-NR	0.4051
WaDIQaM-NR	0.6096	QAC	0.5529	GWHGLBP	0.6882	LPSI	0.6252	LPSI	0.4586	QAC	0.3815
LPSI	0.5953	LPSI	0.4254	ILNIQE	0.6501	GWHGLBP	0.6150	WaDIQaM-NR	0.4498	LPSI	0.2203

*FR Methods included for comparison.

Table 2.24: Kurtosis based check for Gaussianity of prediction residuals of NR Methods, for *all* and *subset* distortions. FR Methods IWSSIM and PSNR are in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
IWSSIM*	01	01	11	1	10	11	1	1	111	111
PSNR*	11	00	11	1	11	11	1	1	111	111
CORNIA	01	11	11	1	11	11	1	1	111	101
HOSA	01	10	11	0	11	11	1	1	111	111
dipIQ	01	11	11	1	11	11	1	1	111	111
NRSL	00	11	10	1	11	11	1	1	111	110
GWHGLBP	11	11	11	1	11	11	1	1	110	111
ILNIQE	11	00	00	1	11	11	1	1	111	101
SISBLIM	00	11	11	1	11	11	1	1	111	101
NIQE	11	11	11	0	11	11	1	1	111	101
MEON	00	11	10	1	11	11	1	1	111	111
BIQI	00	10	11	1	11	11	1	1	111	101
BRISQUE	01	10	10	0	11	11	1	1	111	101
QAC	01	10	10	1	11	11	1	1	111	111
LPSI	11	11	11	1	11	11	1	1	111	111
WaDIQaM-NR	01	11	10	0	11	11	1	1	111	111

*FR Methods included for comparison.

Table 2.25: Statistical significance testing results of NR methods based on prediction residuals for the *all distortions* case. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, MDID, MDID2013, LIVE MD, MDIVL]. A “1” means that the method in the row, for a particular database, is statistically better than the method in the column, a “0” means that it is statistically worse, while a “_” means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

	IWSSIM	**PSNR**	CORNIA	HOSA	dipIQ	NRSL	GWHGLBP	ILNIQE	SISBLIM	NIQE	MEON	BIQI	BRISQUE	QAC	LPSI	WaDIQaM-NR
IWSSIM	__________	1111111111	0111111111	0111111111	1111111111	0111111111	1111111101	1111111111	1111111111	1111111111	111_111111	1111111111	0111111111	1111111111	1111111111	1111111111
PSNR	0000000000	__________	01__110000	01__11011_	01_0110__0	01_011011_	1111110001	1101110_0_	1111110001	01__110_01	01_0111111	011111_1_1	011_11111_	_11111__11	11_1111111	01_1111111
CORNIA	1000000000	10__001111	__________	0_____1111	1100__1111	0__0__1111	11_1__1_01	11011_1101	11_1_01001	11_1__1111	1100__1111	1111__1111	_1__1_1111	11_1__1111	11_1_01111	11_1__1111
HOSA	1000000000	10__00100_	1_____0000	__________	1100____00	1__0____1_	11_11_0001	1_011_0001	11_1___001	11_11__001	1100__1_11	11111___01	11_11_1_11	11_11_1_11	11_1__1_11	11_1__1_11
dipIQ	0000000000	10_1001__1	0011__0000	0011____11	__________	001_____11	1_111__001	1__11_0_01	1111_0_001	11111___01	___0__1_11	1_111_11_1	0_111_1111	1_111_1_11	1_111_1111	__111_1111
NRSL	1000000000	10_100100_	1__1__0000	0__1____0_	110_____00	__________	1_11__0001	1_011_000_	1111_0_001	11_1___001	1_00__1_11	1111____01	11111_1_0_	11_1__1__1	11_1__1__1	11_1__1_11
GWHGLBP	0000000010	0000001110	00_0__0_10	00_00_1110	0_000__110	0_00__1110	__________	1_00___11_	_1_0_0101_	01_0___11_	0_00__111_	0_____111_	0__0__1110	01_0__111_	______111_	0__0__111_
ILNIQE	0000000000	0010001_1_	00100_0010	0_100_1110	0__00_1_10	0_100_111_	0_11___00_	__________	011_0010__	0110__1_1_	0__0__111_	0111__111_	0_10__111_	011___1_1_	0_11__111_	0_10__1111
SISBLIM	0000000000	0000001110	00_0_10110	00_0___110	0000_1_110	0000_1_110	_0_1_1010_	100_1101__	__________	0__0_1_11_	0000__111_	0__111_11_	00_01_1110	0____1111_	_0_1__111_	000___111_
NIQE	0000000000	10__001_10	00_0__0000	00_00__110	00000___10	00_0___110	10_1___00_	1001__0_0_	1__1_0_00_	__________	0000__111_	0__1___11_	00____1110	1__1__1_1_	10_1__111_	00____111_
MEON	000_000000	10_1000000	0011__0000	0011__0_00	___1__0_00	0_11__0_00	1_11__000_	1__1__000_	1111__000_	1111__000_	__________	1_11__0_0_	0_11____00	1111__0___	1_11__1_0_	__11__1___
BIQI	0000000000	100000_0_0	0000__0000	00000___10	0_000_00_0	0000____10	1_____000_	1000__000_	1__000_00_	1__0___00_	0_00__1_1_	__________	0__0__1_1_	1__0___01_	1_0___1_1_	0_00__1_11
BRISQUE	1000000000	100_00000_	_0__0_0000	00_00_0_00	1_000_0000	00000_0_1_	1__1__0001	1_01__000_	11_10_0001	11____0001	1_00____11	1__1__0_0_	__________	1__1__0_11	1__1_____1	1__1____11
QAC	0000000000	_00000__00	00_0__0000	00_00_0_00	0_000_0_00	00_0__0__0	10_1__000_	100___0_0_	1____0000_	0__0__0_0_	0000__1___	0__1___10_	0__0__1_00	__________	10_1__110_	0_____1___
LPSI	0000000000	00_0000000	00_0_10000	00_0__0_00	0_000_0000	00_0__0__0	______000_	1_00__000_	_1_0__000_	01_0__000_	0_00__0_1_	0_1___0_0_	0__0_____0	01_0__001_	__________	0__0____1_
WaDIQaM-NR	0000000000	10_0000000	00_0__0000	00_0__0_00	__000_0000	00_0__0_00	1__1__000_	1_01__0000	111___000_	11____000_	__00__0___	1_11__0_00	0__0____00	1_____0___	1__1____0_	__________

Table 2.26: Statistical significance testing results of NR methods based on prediction residuals for the *subset distortions* case. Each entry is a codeword composed of ten symbols, where each symbol represents the test outcome for one IQA database. The symbol location within a codeword represents IQA databases in the following order: [LIVE R2, TID2013, CSIQ, VCLFER, CIDIQ50, CIDIQ100, LIVE MD BJPG, LIVE MD BN, MDIVL BJPG, MDIVL NJPG]. A "1" means that the method in the row, for a particular database, is statistically better than the method in the column, a "0" means that it is statistically worse, while a "_" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). FR Methods IWSSIM and PSNR are included for comparison and are highlighted in bold.

	IWSSIM	**PSNR**	CORNIA	HOSA	dipIQ	NRSL	GWHGLBP	ILNIQE	SISBLIM	NIQE	MEON	BIQI	BRISQUE	QAC	LPSI	WaDIQaM-NR
IWSSIM	__________	1111111111	0111111101	011111_111	_111111111	0111111111	1111110011	111111__11	111111__11	111111_111	011_111111	_111111111	0111111111	1111111111	1111111111	1111111111
PSNR	0000000000	__________	0_0___0000	0_0___010_	0_00000_00	00001_1110	11111_0001	1111110001	1111_00001	010_110001	0_00__1110	01111__1__	010_11010_	_1011_11__	11_11_0101	0__11__111
CORNIA	1000000010	1_1___1111	__________	0_0____111	1_0000111_	00_01_1111	1111__0011	1111110_11	1111_00_11	11111_0_11	0_00_01111	11111_1111	01___1_111	11111_1111	1111___111	1_111_1111
HOSA	100000_000	1_1___101_	1_1____000	__________	1_000_1010	1010111_1_	1111110011	111111_011	1111___011	111111_011	1__0__111_	1111111_11	111111111_	1111111_11	11111__111	1_111_1111
dipIQ	_000000000	1_11111_11	0_1111000_	0_111_0101	__________	001_111111	1111110011	11111100_1	11111_00_1	11111100_1	0_101_1111	_11111_111	01111101_1	1111111111	11111_01_1	1_11111111
NRSL	1000000000	11110_0001	11_10_0000	0101000_0_	110_000000	__________	1111__0001	1111__0001	1111000001	11111_0001	010000_11_	1111__0_01	1101__0101	1111__0__1	1111000101	1_11__01_1
GWHGLBP	0000001100	00000_1110	0000__1100	0000001100	0000001100	0000__1110	__________	100011110_	___0_011__	00001_11__	0000001110	0_____11_0	0000_11100	0000__1110	000__011__	0000__111_
ILNIQE	000000__00	0000001110	0000001_00	000000_100	00000011_0	0000__1110	011100001_	__________	01__00____	01001__1_1	0000001110	01_10_111_	0_000_11__	010___111_	010100_1__	0000_0111_
SISBLIM	000000__00	0000_11110	0000_11_00	0000___100	00000_11_0	0000111110	___1_100__	10__11____	__________	000011_1__	0000__1110	0__1_11110	0000_111_0	000_111110	_001___1__	000_11111_
NIQE	000000_000	101_001110	00000_1_00	000000_100	00000011_0	00000_1110	11110_00__	10110__0_0	111100_0__	__________	0000001110	01110_1110	000_0_11_0	1__10_1110	1_110011__	001_00111_
MEON	100_000000	1_11__0001	1_11_10000	0__1__000_	1_010_0000	101111_00_	1111110001	1111110001	1111__0001	1111110001	__________	1111110001	1111110_01	1111110001	11111_0_01	1_11110_01
BIQI	_000000000	10000__0__	00000_0000	0000000_00	_00000_000	0000__1_10	1_____00_1	10_01_000_	1__0_00001	10001_0001	0000001110	__________	0000__010_	1000__1_1_	100__0010_	1000__111_
BRISQUE	1000000000	101_00101_	10___0_000	000000000_	10000010_0	0010__1010	1111_00011	1_111_00__	1111_000_1	111_1_00_1	0000001_10	1111__101_	__________	1111__101_	1111_0_1_1	1011_01_11
QAC	0000000000	_0100_00__	00000_0000	0000000_00	0000000000	0000__1__0	1111__0001	101___000_	111_000001	0__01_0001	0000001110	0111__0_0_	0000__010_	__________	1__1_00101	00____01__
LPSI	0000000000	00_00_1010	0000___000	00000__000	00000_10_0	0000111010	111__100__	101011_0__	_110___0__	0_001100__	00000_1_10	011__1101_	0000_1_0_0	0__0_11010	__________	00_0__1_1_
WaDIQaM-NR	0000000000	1__00__000	0_000_0000	0_000_0000	0_00000000	0_00__10_0	1111__000_	1111_1000_	111_00000_	110_11000_	0_00001_10	0111__000_	0100_10_00	11____10__	11_1__0_0_	__________

depicts the outcome of the check for either the *all* or *subset* (SS) distortions cases. Specifically, the order of symbols within each table entry is as follows: LIVE R2 (All, SS), TID2013 (All, SS), CSIQ (All, SS), VCLFER (All), CIDIQ50 (All, SS), CIDIQ100 (All, SS), MDID (All), MDID2013 (All), LIVE_MD (All, Blur-JPEG, Blur-Noise), MDIVL (All, Blur-JPEG, Noise-JPEG). It can be observed from Table 2.24 that the kurtosis based assumption of Gaussianity holds in around 85% of cases. The prediction residuals of all NR methods were compared by carrying out hypothesis testing through the one-sided (left-tailed) two-sample F-test at 95% confidence (as in Section 2.4.6). Tables 2.25 and 2.26 provide the outcome of the statistical significance testing for the *all distortions* and *subset distortions* cases, respectively. For details of how to interpret the tables, refer to Section 2.4.6, and to the captions of Tables 2.25 and 2.26.

As in Section 2.4.7, the computational complexity of all 14 NR IQA methods under test was evaluated in terms of their execution time to determine the quality of a 1024 × 1024 color image on a Lenovo laptop computer with a 2.4GHz Intel Core i7-4700MQ processor, 12GB of RAM, Samsung 850 EVO Solid State Drive, and Windows 10 Home operating system. The execution times of all NR methods are given in Table 2.27, where methods have been sorted in ascending order with respect to execution time. As before, we provide the execution time of NR methods relative to the FR method PSNR for convenience in comparison with Table 2.17. Apart from the 14 NR methods being evaluated in this work, we have included the execution times of seven other well-known NR IQA methods in Table 2.27, which include: BLIINDS2 [176], DIIVINE [177], FRIQUEE [178,179], Jet-LBP [180], MS-LQAF [181], NFERM [182], and TCLT [183]. We have not evaluated the performance of these methods because they take an excessive amount of time to estimate the quality of an image, and are infeasible for large-scale or real-time use. It should also be noted that while WaDIQaM-NR [148] takes a lot of time to determine the quality of the test image on the CPU (10.1277 seconds), it runs considerably faster when executed on the GPU. For reference, on another machine, WaDIQaM-NR ran around 40 times faster on the GPU as compared to the CPU.

Table 2.27: Execution Time of NR methods on a test image. Methods are sorted in ascending order with respect to the execution time. FR Method PSNR is included for comparison and is highlighted in bold.

NR Method	Execution Time (Seconds)	Execution Time (Relative to PSNR)
PSNR[*]	0.0044 s	1.00
LPSI	0.0827 s	18.80
MEON	0.2348 s	53.36
QAC	0.2811 s	63.89
HOSA	0.3312 s	75.27
NRSL[a]	0.3895 s	88.52
GWHGLBP[a]	0.3945 s	89.66
NIQE	0.4558 s	103.59
BRISQUE	0.4641 s	105.48
BIQI	1.2045 s	273.75
dipIQ	2.8367 s	644.70
Jet-LBP[a,b,c]	3.1004 s	704.64
CORNIA	3.6154 s	821.68
ILNIQE	4.0060 s	910.45
SISBLIM	5.3890 s	1224.77
TCLT[c]	7.8548 s	1785.18
WaDIQaM-NR	10.1277 s	2301.75
MS-LQAF[a,c]	36.9052 s	8387.55
DIIVINE[c]	38.2215 s	8686.70
BLIINDS2[c]	94.6167 s	21503.80
FRIQUEE[c]	109.1559 s	24808.16
NFERM[c]	128.8809 s	29291.11

[*] FR Method included for comparison.

[a] Feature extraction time only.

[b] The performance of Jet-LBP was not evaluated as SVR model parameters are not available.

[c] The performance of these methods was not evaluated due to their large computation times.

2.5.2 Analysis and Discussion

It can be observed from Tables 2.22 and 2.23 that in terms of weighted average PLCC and SRCC, the NR method CORNIA [141] outperforms other NR methods, sometimes by a clear margin, for the cases of *all databases* and *multiple distortion databases* in both the *all distortions* and *subset distortions* categories. In case of *single distortion databases*, HOSA [143] does well for the *all distortions* category, while dipIQ [36] and MEON [146] do well in the *subset distortions* category. Since the OA NR methods are trained on databases that are constituents in the weighted average PLCC and SRCC computation, as described in Section 2.3.3, these results should be considered in conjunction with the statistical significance testing outcome. From Tables 2.25 and 2.26, it can be respectively observed that for both the categories of *all distortions* and *subset distortions*, the NR methods CORNIA [141], HOSA [143], and dipIQ [36], perform better than most other methods on most databases. CORNIA and HOSA are OA NR methods that first learn image features and then a quality model, while dipIQ is an OU NR method that utilizes millions of DIPs and a learning-to-rank algorithm to learn the quality model. However, HOSA itself can be regarded as a modified version of CORNIA, while dipIQ uses CORNIA features at its base. This shows that CORNIA features [141] are quite effective when it comes to blind IQA.

The following observations can be made about various NR design philosophies: 1) The OA NR methods that use handcrafted features (BIQI [139], BRISQUE [140], GWHGLBP [142], and NRSL [147]), do not show robust cross-dataset performance. While they may perform better on one class of data, such as single distortion or multiple distortion datasets, their performance degrades considerably on another class of data. This shows that such models suffer from model overfitting and database dependency issues, and also that truly general-purpose handcrafted features for perceptual IQA remain lacking. 2) OA NR techniques that utilize unsupervised feature learning, such as CORNIA [141] and HOSA [143], demonstrate relatively robust performance. For example, even though these methods are trained on singly distorted content, they perform relatively well on multiply distorted databases, which is somewhat surprising. 3) Among OA NR methods that employ deep learning, MEON [146] performs better than WaDIQaM-NR [148]. This may be because MEON uses two sub-tasks to perform IQA, where a large amount of data is used to pre-train the distortion identification aspect of the network. However, unlike CORNIA and HOSA, these methods do not perform adequately on multiply distorted content, even though they are trained on individual distortion types that make up the multiple distortion combinations. This further highlights the difficulties encountered while doing IQA for multiply distorted content and while training deep learning models on small-scale datasets. 4) dipIQ [36] performs better than most NR methods. In addition to using CORNIA features, dipIQ utilizes a novel training process which does not use human annotated data for training. Instead, it alleviates the issue of small-scale subject-rated datasets by using millions of DIPs, generated by using FR IQA methods, to train the model. This approach highlights the advantages of utilizing techniques that use large-scale datasets which employ alternative annotation techniques. 5) The performance of OU NR methods (NIQE [3], ILNIQE [144], QAC [35], SISBLIM [32], and LPSI [145]) shows considerable room for improvement, which also highlights the difficult nature of the OU NR IQA problem. 6) While many training-based NR methods are usually trained and tested on each database separately, which often leads to high PLCC and SRCC numbers, we believe that cross-dataset testing is crucial to the performance analysis of NR methods. 7) While NR methods such as CORNIA and dipIQ may be relatively better in quality prediction performance compared to other methods, they have a large execution time, as can be seen from Table 2.27. This implies that such methods are infeasible for real time usage. 8) We have included the FR IQA methods IWSSIM [13] and PSNR for comparison in the tables of this section, and it can be seen that the performance of all NR IQA methods is still a considerable distance away from top performing FR methods such as the IWSSIM, a disparity which is even more pronounced in the *all distortions case*. Even the perceptually inaccurate PSNR outperforms many NR methods, especially for the *all distortions* case. The above-mentioned observations highlight the significant room for improvement that exists in the area of NR IQA, both in terms of quality prediction performance and execution time.

2.6 Summary

In this chapter, we carried out an extensive review and performance evaluation study of the field of IQA. In all, we evaluated the performance of 43 FR, seven fused FR and 14 NR IQA methods. If the 22 different versions of the seven fused FR methods are considered separately, then this means that we evaluated 79 IQA methods. In order to ensure the diversity of test data, we used nine subject-rated IQA datasets, five of which are composed of singly distorted content, while four contain multiply distorted content. To the best of our knowledge, this is so far the largest study of its kind, and hopefully will plug the gap that previously existed with regard to the lack of such surveys in the area of image quality assessment.

In summary, this chapter has the following findings: 1) Among the individual FR methods, structural similarity based methods IWSSIM [13], FSIMc [14], DSS [16], and VSI [15], are the top performers. 2) The empirical (HFSIMc [129], CM3 [127], CM4 [127]) and learning based (MMF [130], CNNM [128]) fusion approaches are not only outperformed by rank aggregation based fusion approach RAS [41], but also by top performing individual FR methods, thereby implying that existing empirical and learning based fusion methods do not offer clear advantages over individual FR methods. 3) However, the rank aggregation based fusion approach RAS [41] not only comprehensively outperforms other fusion approaches but also top performing individual FR methods. Its training-free nature and robust cross-dataset performance make it highly promising as a means to annotate very large-scale IQA datasets in the future. 4) Among NR methods, we have found CORNIA [141], HOSA [143], and dipIQ [36], to perform better than other methods. 5) While the perceptual quality prediction performance of FR methods has matured quite well, the performance of NR methods, both in terms of perceptual quality prediction accuracy and computational complexity is still a long distance away from top performing FR methods.

This chapter not only highlights the current state-of-the-art in the field of IQA of 2D natural images, but also the challenges that IQA researchers need to address, especially in the area of BIQA. As discussed in Section 2.5.2, the top performing NR models CORNIA [141], HOSA [143], and dipIQ [36] utilize CORNIA features which are learned automatically in an unsupervised manner, thereby highlighting the strength of learned against handcrafted features. DNN based models have enjoyed a lot of success in other areas of computer vision and image processing [17], which is largely due to the availability of very large-scale annotated datasets such as ImageNet [18]. On the other hand, DNN based BIQA models, such as the ones evaluated in this chapter (WaDIQaM-NR [148] and MEON [146]) show a lot of room for improvement. These and other DNN based IQA models identified in Section 2.3.3 train on the available small-scale IQA datasets (with hundreds or a few thousands of images) and may try to increase training data size by data augmentation, but achieved only limited success. The design of very large-scale annotated IQA datasets is an open problem [184]. The real challenge is that it is impossible to perform subjective tests to annotate such very large-scale datasets, thus the use of alternative data annotation techniques is highly desirable. One important discovery of the work in this chapter is that rank aggregation based training-free FR fusion methods offer good promise of robust perceptual quality prediction performance when tested across a wide range of available subject-rated datasets. Thus, very large-scale simulated distortion datasets, with millions of images, may be developed where distortions are added in a content adaptive manner. Such datasets can then be synthetically annotated by using rank aggregation based FR fusion methods. DNN models can then be trained by utilizing such new datasets. This research direction deserves deeper investigation and will be the focus of the next chapter in this thesis.

Chapter 3

Addressing the Data Challenge

Visual content has come to play a central role in our lives and new content is being generated at an exponential rate. Image quality assessment (IQA) models, especially blind IQA (BIQA) models, have thus gained prime importance. However, in the literature, the enormous space of all possible natural images is usually represented by a handful of small-scale annotated IQA datasets, which are used to train and test BIQA models. This has imposed serious limitations on the development of robust and accurate deep neural networks (DNN) based BIQA methods as such models typically require an enormous amount of training data. It is difficult, if not impossible, to create large-scale human-rated IQA databases, composed of millions of images, due to the constraints of subjective testing. While considerable efforts have been made to enhance the performance of DNN based BIQA methods by focusing on the modeling aspect, efforts to address the scarcity of labeled IQA data remain surprisingly missing. To address this data challenge, we first construct a very large-scale dataset named the Waterloo Exploration-II database, which in its current state contains 3,570 pristine reference images and around 3.45 million singly and multiply distorted images created from them. Next, we develop a novel mechanism that synthetically assigns highly accurate perceptual quality labels to the distorted images, thereby allowing for the development of DNN based IQA models. To validate the utility of our very large-scale database and the synthetic quality annotation process, we construct a DNN based BIQA model called EONSS and train it on the Waterloo Exploration-II database. We extensively test EONSS on nine subject-rated IQA databases without any retraining or fine-tuning, and compare it with state-of-the-art BIQA methods. Our tests, that include a variety of evaluation criteria, reveal that EONSS, even with its simple network architecture, is able to outperform the very best of methods in the BIQA field, including DNN based BIQA models, and is much faster than other BIQA methods. This demonstrates the effectiveness of our approach to address the data challenge in the area of IQA by creating a new very large-scale IQA database with synthetically annotated quality labels for DNN based IQA model training.

3.1 Introduction

IQA can be classified into *subjective* and *objective* quality assessment (QA), and objective QA algorithms can be further categorized into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) or Blind IQA (BIQA) methods, as elaborated in Chapter 2. In the last two decades, significant progress has been made in the development of FR IQA algorithms. Many state-of-the-art FR methods (such as but not limited to [13–16]) are training-free and their predictions correlate well with human perception of quality while evaluating images afflicted with common distortion types. This is evident when they are tested on a wide variety of subject-rated datasets as was comprehensively shown in Chapter 2. Although the performance of FR methods has matured quite well, their practical application remains limited because in real-world media delivery systems, access to pristine reference images is either extremely rare or altogether nonexistent especially at the end-user level. In such practical scenarios, NR IQA or BIQA is the only feasible option. While a lot of work has also been done on the development of BIQA methods, significant room for improvement exists to further enhance their performance as was shown in Chapter 2 and in an earlier study [185]. This is understandable as BIQA is a much more difficult task owing to lack of access to the reference image.

While a few recent BIQA methods are either training-free [32, 145] or require training only to learn a universal model of pristine images [3, 144], a large number of BIQA methods try to alleviate the constraints posed by lack of access to the reference image by employing machine learning algorithms where training is done against human-annotated distorted content. Most training-based BIQA methods extract features from the distorted image and use standard regression methods such as SVR [149,153] in conjunction with subject-rated data to learn a quality model. These features are either domain knowledge based handcrafted features [139,140,142,147,176,177,179,182,186] or learned features [141,143]. We showed in Chapter 2 that BIQA models that employ learned features (such as [36,141,143]) offer better general-purpose performance compared to those that use handcrafted features. Since data-driven end-to-end optimized deep neural networks (DNNs) combine the tasks of learning goal-oriented features and regression, they have great potential to outperform the traditional two-stage approach where feature extraction and regression are optimized independently, but not jointly. A prerequisite requirement for using such deep networks is to have an adequately large set of training data. This is because such networks have hundreds of thousands, if not millions of parameters, and an insufficient amount of training data leads to overfitting, thereby degrading the generalizability of the trained model to unseen data. While a number of DNN based BIQA models have been proposed recently [187], they all encounter a significant hurdle: *the lack of large-scale perceptually annotated training data* [184,187]. Since obtaining large-scale subject-rated data (hundreds of thousands to millions of quality annotated images) is difficult, if not impossible, contemporary DNN based BIQA models focus on data augmentation techniques to enhance the size of the small-scale annotated IQA data that is available [184,187]. However, even with data augmentation, the size of the training data remains limited, and such augmentation techniques lead to their own issues. When tested on unseen data, the performance of DNN based BIQA models remains inadequate (as we showed in Chapter 2 for [146,148]).

In this chapter, we focus on the fundamental problem plaguing the development of high performance DNN based BIQA models, that is, the lack of large-scale training data. Through the development of a very large-scale synthetically annotated IQA dataset, we show that the performance of a DNN model with a simple architecture, when tested on wide-ranging subject-rated unseen data, can be elevated so much so that it not only outperforms recent DNN based BIQA models but also the very state-of-the-art in BIQA. The rest of the chapter is organized as follows. The data challenge in the area of IQA is discussed in Section 3.2 along with the contributions of this chapter. In Section 3.3 we discuss the construction of a very large-scale IQA dataset. In Section 3.4 a novel technique to synthetically annotate this dataset with perceptual quality ratings is presented and its performance is evaluated. Section 3.5 discusses the construction of a simple DNN based BIQA model that trains on the newly developed very large-scale IQA dataset along with extensive performance evaluation of this model on subject-rated IQA datasets in a variety of scenarios to demonstrate the effectiveness of our approach to elevate model performance by addressing the data challenge. The practical applications of the work done in this chapter are discussed in Section 3.6 while Section 3.7 concludes it.

3.2 The Data Challenge

3.2.1 DNNs in Visual Recognition: A Case Study

The application of DNNs has led to tremendous progress in the area of visual recognition, thus it is important to ascertain the reasons for this success. Like machine learning based BIQA, methods in visual recognition are composed of two important components, the model and the data used to train the model. While a lot of work has been done on the modeling component of the visual recognition task, a little more than ten years ago researchers started to focus on the data component of this task [18, 188]. While the Tiny Image dataset [188] has around 80 million loosely labeled low-resolution images, the ImageNet database [18, 189] is composed of more than 14 million more precisely labeled higher resolution images and has led to many breakthroughs in visual recognition. The images in ImageNet populate close to 22,000 [189] synonym sets (synsets) of the WordNet [190, 191] hierarchy with an average of 650 images per synset. For the image classification task, ImageNet first obtains images by crawling the Internet through synset specific search queries to several image search engines. Next, it engages human subjects, through the online crowdsourcing platform Amazon Mechanical Turk [80], to verify that images have been associated with the correct labels regardless of any distortions that may be present. Although it is not an easy task to ask human subjects to verify the labels associated with millions of images, it is still manageable because: 1) Subjects are not being asked to label an image from scratch, instead they are required to verify if an image contains an object associated with the given label. This simplifies the task. 2) The verification requires binary answers (Yes/No). 3) Each image can be treated as an independent entity. 4) Votes from only a few subjects are sufficient to verify the label of each image. Higher levels of the hierarchy are usually easier to verify and require votes from just a few subjects (much less than five), while deeper levels of the hierarchy may require votes from more subjects (five to ten) [18]. 5) Viewing conditions and devices do not impact the authenticity of label verification by subjects, and hence crowdsourcing can be conveniently used.

The annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [17] ran from 2010 to 2017 and was composed of subsets of ImageNet images for the purposes of algorithm training, validation, and testing. ILSVRC tasks included image classification, object localization, and object detection. With the availability of a large-scale dataset such as ImageNet along with computationally powerful GPUs coming of age, the stage was set for the development of effective DNN based visual recognition models. While the first two years of ILSVRC did not see DNN based entries, a significant turning point was observed in ILSVRC 2012, when a deep convolutional neural network (CNN) based model [192], with 60 million parameters, comprehensively won both the classification and localization challenges in terms of the top-5 error rate [17]. The margin with which [192] outperformed other models in the 2012 challenge had such an impact that submissions to the ILSVRCs in the subsequent years were predominantly deep CNN based. Since 2012, deep CNN based models have won the various ILSVRCs in terms of top-5 error rate for the image classification and object localization tasks, and in terms of mean average precision for the object detection task [17]. Thus, the development of high performance, generalizable and robust deep CNN based visual recognition models such as [156, 192–196] has become possible due to the ImageNet database [18] and the ILSVRC [17].

Finally, it is pertinent to mention that some other datasets in the area of visual recognition, such as the PASCAL VOC datasets [197], Caltech 101 dataset [198], and Caltech 256 dataset [199], that have between 9,000 to 31,000 images in 20 to 256 object classes, are considered small-scale and training DNN models from scratch on these datasets is considered infeasible due to overfitting concerns [156].

3.2.2 DNNs in BIQA: The Data Challenge

Small Scale of IQA Datasets

Compared to annotating datasets for visual recognition tasks such as ImageNet [18], obtaining subjective ratings of *image quality* is an altogether different and much more complex scenario because: 1) Subjects need to provide their *opinion* of an image's quality, which is a rather abstract concept and requires substantial critical thinking on the subject's part. 2) The quality scale is not binary, instead it either has a number of discrete levels (five or more) or is continuous. 3) A subject's opinion of quality needs to be *calibrated* before the experiment, so that they have a rough idea about the range of quality to expect. While subjects are asked to treat each image independently during the experiment, they still need to provide ratings relative to the quality range introduced to them. 4) To ensure reliability, it is recommended that at least 15 subjects rate each image in the subjective experiment [70]. 5) It is suggested that a test session should last no longer than 30-minutes to avoid fatigue effects and that participating subjects be screened for visual acuity and color vision [70]. 6) Viewing conditions play a crucial role in the appearance of visual content, and hence on its quality. Therefore, viewing conditions such as display luminance, background luminance, room illumination, observation angle, viewing distance, play an important role in subjective tests and need to be set according to established norms [70].

Considering the above-mentioned constraints, it is much more difficult, if not impossible, to carry out subjective testing for IQA datasets composed of millions of images, even with crowdsourcing. To-date IQA datasets consist of hundreds or a few thousands of distorted images. A summary of contemporary subject-rated IQA databases of 2D natural images is given in Table 3.1. IQA datasets are classified either as *simulated* or *authentic* distortion databases, depending upon whether distortions were simulated on a set of pristine reference images or if they were captured directly in the real-world environment, respectively. Simulated distortions datasets can further be classified into either *singly* or *multiply* distorted databases, where each distorted image is afflicted by a single distortion in the former case or by multiple simultaneous distortions in the latter. Among simulated distortion datasets, the multiply distorted ones are more accurate representations of practical content since visual content almost always undergoes multiple distortions in the real-world.

Table 3.1: Summary of contemporary subject-rated IQA databases.

Database Category	Database	Published Year	No. of Reference Images	No. of Distorted Images	No. of Distortion Types	No. of Distortion Levels	Images per Distortion Type	Subjective Data Type
Simulated Distortions (Singly Distorted)	A57 [75]	2007	3	54	6	3	9	DMOS
	CIDIQ [5]	2014	23	690	6	5	115	MOS
	CSIQ [26]	2010	30	866	6	4 to 5	116 to 150	DMOS
	IVC [30]	2005	10	185	4	5	20 to 50	MOS
	KADID-10K [22]	2019	81	10,125	25	5	405	DMOS
	LIVE R2 [24]	2006	29	779	5	Up to 5	145 to 175	DMOS
	MICT-Toyama [29]	2008	14	168	2	6	84	MOS
	PDSP-HDDS [20]	2018	250	12,000	10	4 to 5	1,000 to 1,250	MOS
	TID2008 [25]	2008	25	1,700	17	4	100	MOS
	TID2013 [19]	2013	25	3,000	24	5	125	MOS
	VCLFER [54]	2012	23	552	4	6	138	MOS
Simulated Distortions (Multiply Distorted)	LIVE MD [31]	2012	15	405	5	3	45 to 135	DMOS
	MDID2013 [32]	2014	12	324	1	3	324	DMOS
	MDID [33]	2017	20	1,600	1 to 4	4	N/A	MOS
	MDIVL [34]	2017	10	750	2	4 to 10	350 to 400	MOS
Authentic Distortions	BID [77]	2011	N/A	585	5	N/A	57 to 204	MOS
	CID2013 [78]	2015	N/A	480	12 to 14	N/A	N/A	MOS
	KonIQ-10K [81]	2018	N/A	10,073	N/A	N/A	N/A	MOS
	LIVE Challenge [79]	2016	N/A	1162	N/A	N/A	N/A	MOS

From Table 3.1 it can be seen that the largest singly distorted simulated distortion dataset, the recently released PDSP-HDDS [20], has only 12,000 distorted images, while the largest multiply distorted dataset, MDID [33], has only 1,600 distorted images. Among authentic distortion databases, the recently released KonIQ-10K [81], has only 10,073 distorted images. As is evident from recent surveys of DNN based BIQA models [184, 187], datasets such as LIVE R2 [24], LIVE MD [31], LIVE Challenge [79], CSIQ [26], and TID2013 [19] are used to train such models. The largest dataset among them is the singly distorted database TID2013 [19] which has only 3,000 distorted images. From Table 3.1, it can be observed that each dataset has only a limited number of images per distortion type. The number of images per distortion type per level of distortion is even smaller. For example, for singly distorted datasets, this is usually equal to the number of pristine images in the dataset. It can also be observed from Table 3.1 that simulated distortion datasets have a very limited amount of pristine reference content. The PDSP-HDDS [20] has 250 reference images, however it is itself an outlier as all other datasets have less than 100 reference images (usually 10 to 30). Since these pristine reference images are *supposed* to be representatives of the enormous space of all possible natural images, contemporary IQA datasets do rather inadequately in terms of overall content variation. These shortcomings of subject-rated IQA datasets create enormous hurdles in the development of generalizable and robust DNN based BIQA methods. Recently two large-scale singly distorted datasets have been constructed. These include the Waterloo Exploration-I dataset [21], which has 4,744 reference and 94,880 distorted images, and the KADIS-700k dataset [22], which has 140,000 reference and 700,000 distorted images. However, they cannot be used to train DNN based BIQA models as their distorted content has not been annotated with perceptual quality labels. Thus, the issue of large-scale *annotated* IQA data is an open problem which needs to be resolved.

Current Strategies to Deal with Lack of Data

Contemporary DNN based BIQA models employ a number of data augmentation techniques to deal with the issue of small-scale training data.

A widely used technique adopted by DNN based BIQA methods to increase the size of the training set is to extract multiple fixed size small patches from each labeled image [38, 39, 148, 155, 158, 160–163, 165, 167, 168, 200]. While a popular patch size is 32 × 32 [38, 39, 148, 155, 158, 160–163, 200], larger sized patches have also been employed [165, 167, 168]. Since local patch-level quality labels are not available in IQA databases, the global image-level quality score is usually applied to each patch extracted from an image. Due to the influence of distortions and the visual attention property of the HVS, some regions of an image might seem perceptually more relevant to a human subject while assigning global quality scores [116, 117]. An image might be assigned a low quality score, yet patches extracted from it might receive high quality scores when viewed independently. Thus, the assignment of the global quality score to local patches extracted from an image leads to a significant label noise problem. The method in [38, 39] tries to address this problem by splitting model training into two steps. In the first step, an FR method (FSIM [14]) was used to assign a quality score to each local patch, and a CNN was pre-trained using this patch-level data. In the second step, the model was fine-tuned on subject-rated datasets. Similar to [39], the method in [168] also follows a two-step training process, however, instead of using FR methods to label local patches, it uses the exponent difference function to generate objective error maps for these patches, which are then used as intermediate training targets. The model is fine-tuned on subject-rated datasets. The method in [148, 155] tries to alleviate the label noise problem by weighted-average patch quality aggregation. It estimates the quality of 32 × 32 image patches and also determines the relative weight of each patch to account for its contribution in the global quality of the parent image. Patch weight estimation is carried out by adding a parallel branch to the patch quality regression layers, and the whole network is optimized in an end-to-end manner. In [162], an image is segmented and the Prewitt operator is used to generate the gradient map through which patch weights are determined. The quality of each image patch is predicted by a CNN and the global image quality is computed through a weighted average of patch qualities. Even with the adoption of the patch-based data augmentation technique, the volume of training data remains limited given the small-scale of IQA datasets. While a recent method [200] tries to further increase the training data size by using various combinations of distorted and their corresponding reference image patches, to generate patch pairs which are used for CNN training, the overall amount of training data still remains limited. The very small amount of reference content and the label noise issue further impacts the utility of this data augmentation technique.

Some methods [39, 146, 166, 168] increase the size of the training data by horizontally flipping the images or image patches, and use the quality label assigned to the parent image. Other kinds of geometric transformations cannot be applied in the area of IQA as they can significantly impact the perceptual quality of an image [184]. In addition to horizontal flipping, the method in [146] creates additional training samples by changing the saturation and contrast of images as long as these changes do not impact their perceptual quality. Given the small-scale of IQA datasets, this data augmentation technique also leads to a limited expansion of training data and suffers from the limited nature of reference content.

Since very large-scale annotated databases are available in the area of visual recognition, some BIQA methods utilize DNNs that have been pre-trained for the visual recognition task. In [165], the Caffe network [195] that has been pre-trained on the ImageNet [18] and Places [201] visual recognition databases is used in two ways: 1) As a feature extractor, where SVR and IQA databases are used to map these features to perceived quality scores;

2) As an initialization, where the network is fine-tuned with respect to IQA databases. In [164], the VGG network [156], that has been pre-trained on the ImageNet [18] database for the object recognition task, is used where feature vectors are extracted from different layers of the network and form a multi-level representation of the image. Next, IQA databases are used along with SVR to learn a mapping from each feature vector to a quality score. A global quality score for an image is then computed as the average of quality scores predicted by different network layers. Instead of predicting a single quality score, the method in [167] predicts the quality distribution of a given image using CNNs. The output of the CNN model is in terms of probabilistic quality representation (PQR) vectors that are then mapped to scalar quality scores using SVR. Of the three CNNs used in [167], two are deep CNNs (AlexNet and ResNet50) that have been pre-trained for the image classification task on the ImageNet database [18]. These deep CNNs are then fine-tuned by using subject-rated IQA databases. Although features extracted from a DNN that has been trained for a particular visual recognition task, such as image classification, are known to be effective generic features for other visual recognition tasks [202,203], their use in an altogether different area, such as IQA, is open to doubt [187].

Some DNN based BIQA methods adopt a multi-task strategy to deal with the lack of quality-annotated training data. The work in [160] is a pioneering effort in this direction, where image quality and distortion type are simultaneously estimated. It is demonstrated in [160] that such a multi-task approach allows for a reduction in the model's learnable parameters without loss in model performance. The method in [146] uses the multi-task approach of distortion identification and quality prediction, however in a causal manner. It splits these two tasks between two sub-networks such that their early layers are shared. Sub-network 1, which identifies distortion type through a probability vector is fed into sub-network 2, which predicts image quality. Since a large amount of labeled data can be generated for the distortion identification task without the need for human annotations, 840 pristine images are degraded at five distortion levels for different distortion types in [146] to generate a large amount of training data, which is used to pre-train sub-network 1 along with the shared layers of the overall network. The entire network is subsequently joint optimized using subject-rated data. The method in [204] utilizes two deep CNNs to separately deal with the scenarios of synthetically (simulated distortions) and authentically distorted images. For synthetic distortions, the distortion type and level information is used for pre-training, where the training set includes 852,891 distorted images that have been obtained by using 9 distortion types to degrade 21,869 pristine images at various distortion levels. For authentic distortions, the CNN VGG-16 [156], that has been pre-trained on the ImageNet database [18] for the image classification task, is used. The feature sets from the two CNNs are transformed into one representation set through bilinear pooling. The entire network is fine-tuned on subject-rated IQA datasets. In these methods, especially [146,204], the use of multi-task learning, where distortion identification is carried out in addition to quality score prediction, has the benefit of enabling the training process to be split into pre-training and joint optimization stages. Labels for distortion identification do not require human annotation, and thus a large amount of data can be generated for the pre-training step. However, such an approach does not take into account the impact of content variations. Distortions of the same type and magnitude can lead to drastically different perceived quality results for two different contents. This is a fundamental limitation of such multi-task learning based approaches for data augmentation.

Since deep models require a large amount of training data due to the high-dimensional nature of images as model inputs, some techniques (such as [159]) use low-dimensional representations of images, by using NSS [123] features extracted from the images, as inputs to the model. While this reduces the training size requirements of training data, such models are unable to realize the full potential of DNNs since end-to-end learning is lacking.

While a lot of efforts have been made to construct DNN based BIQA methods that focus on the modeling part of the problem and try to alleviate the lack of training data by using data augmentation and multi-task learning techniques (as described above), efforts to address the fundamental problem of lack of large-scale quality-annotated IQA databases remain surprisingly missing. In this chapter we focus on addressing this fundamental problem plaguing the development of robust and generalizable DNN based BIQA models. Specifically, we make the following three novel contributions: 1) We construct the largest IQA dataset to-date, called the Waterloo Exploration-II database, which has 3,570 pristine and more than 3.45 million distorted images (including both singly and multiply distorted content). 2) Since annotating so many images through subjective testing is not possible, we devise a novel synthetic quality benchmark generation mechanism that annotates the images with perceptually oriented quality ratings. Our tests on a wide range of available subject-rated IQA datasets show that this mechanism leads to quality annotations that are highly correlated with human perception of quality, and thus they can be used as alternatives to human quality ratings. 3) To show the advantage of the large-scale synthetically annotated Waterloo Exploration-II database, and thus the strength of our approach to resolve the data challenge in IQA, we develop a DNN based BIQA model called EONSS and train it using the Waterloo Exploration-II dataset. Although EONSS has a simple architecture, we show that when tested across a wide range of available subject-rated IQA datasets, it not only comprehensively outperforms other DNN based BIQA models with more complex architectures that use data augmentation, but it also outperforms the very state-of-the-art in BIQA, thereby highlighting the significance of our approach to overcome the data challenge encountered when constructing DNN based BIQA models.

3.3 Waterloo Exploration-II Database Construction

The Waterloo Exploration-II database is a simulated distortions dataset which starts with a set of pristine reference images and simulates both singly and multiply distorted images at various distortion levels.

3.3.1 Reference Content

The *pristine* or *reference* content in an IQA database is representative of the enormous space of all possible natural images. As is evident from Table 3.1, contemporary IQA databases have only a small number of reference images, which can be regarded as a very sparse and inadequate representation of this enormous space. To ensure wide coverage of image content, we include 3,570 reference images in the Waterloo Exploration-II database, which we take from the Waterloo Exploration-I database [21]. To ensure images that are representatives of what humans see in their daily lives, the creators of the Waterloo Exploration-I database [21] use 196 keywords to search the Internet for images which broadly belong to seven categories (human, animal, plant, landscape, cityscape, still-life, and transportation) and obtain an initial set of 200,000 images. Next, they manually view each image and remove those that have any visible distortions, leading to a filtered set of 7,000 images. Finally, they carry out another round of filtering where they view each remaining image by zooming in multiple times and remove any images with visible compression distortions leaving 4,744 high quality natural images that become the reference image set of the Waterloo Exploration-I database [21]. For the Waterloo Exploration-II database, we start with the 7,000 images that were obtained after the first round of filtering while obtaining the Waterloo Exploration-I database reference image set. Using a similar manual procedure of viewing each of these images multiple times and zooming in, we carry out another round of filtering and select 3,570 pristine quality images as our pristine reference image set.

While the usual practice is to describe the variety of reference content by using subjective terms, a few quantitative descriptors have also been used to describe such content, such as image spatial information (SI) which is indicative of edge energy in an image, and colorfulness (CF) which represents the variety and intensity of colors in an image [88]. The 2D SI versus CF space has been used to represent and compare the diversity of source content in different IQA databases [88]. Three different SI measures were compared in [89] and it was found that a mean based SI measure (SI_{mean}) has the highest correlation with compression based image complexity measures. We use SI_{mean} [89] and a computationally efficient CF measure [90] to plot the reference image content of the Waterloo Exploration-II database in the SI versus CF space, as shown in Fig. 3.1, where the blue outer boundary marks the convex hull and the area inside is marked yellow. Fig. 3.1 suggests a comprehensively improved content representation in the Waterloo Exploration-II database in terms of both diversity and density in comparison with nine well-known IQA databases, whose SI versus CF plots are shown in Fig. 2.1 of Chapter 2.

3.3.2 Distorted Content

An ideal simulated distortions IQA dataset should be diverse in terms of distortion types and levels. The goal is to simulate varying degrees of distortions so that the perceptual quality scale is uniformly sampled, which ensures that objective IQA methods are tested

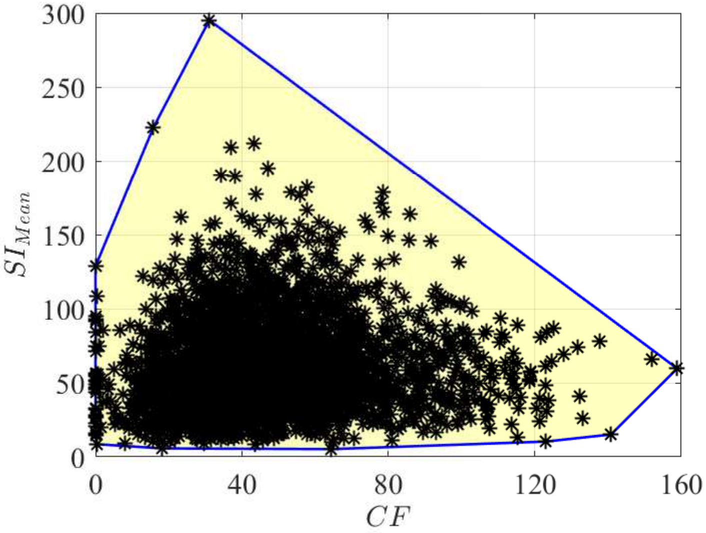


Figure 3.1: Spatial Information (SI_{Mean}) versus Colorfulness (CF) plot of the reference images of the Waterloo Exploration-II database. The blue lines represent the convex hull.

(and trained) across the quality spectrum. Distortions also need to be realistic, and thus multiply distorted content takes precedence over singly distorted images. Existing simulated distortion datasets, as summarized in Table 3.1, have the following shortcomings apart from their small-scale nature: 1) Most of them are singly distorted. 2) Usually 4 to 6 distortion levels per distortion type are used, which does not allow for a dense sampling of the perceptual quality scale. 3) Existing multiply distorted datasets usually have 3 to 4 distortion levels per distortion type per stage, leading to a sparse multiply distorted image set which is inadequate for learning how two or more different (or same) distortions interact with each other (this point will be further elaborated in Chapter 4). 4) Distorted content in most IQA datasets is not uniformly distributed across the quality spectrum as demonstrated in Chapter 2, which is because fixed distortion parameters are used to generate each distortion type. While this is a convenient approach, it does not adapt to the impact of content variations on the perceptual appearance of distortions. Since it is known that many objective IQA methods find it more difficult to evaluate better quality images compared to lower quality ones [19], effective representation of the entire quality scale, especially the higher quality region is necessary. To address the above-mentioned shortcomings of existing IQA datasets, we generate the distorted content of the Waterloo Exploration-II database in the following manner.

Content Adaptive Distortion Thresholds

To ensure uniform coverage of the entire quality spectrum, we use content adaptive distortion parameters instead of fixed ones. For its first version, we choose the following four base distortions for the Waterloo Exploration-II database: 1) Gaussian white noise, 2) Gaussian blur, 3) JPEG compression, and 4) JPEG2000 compression. We use one of the most advanced FR quality-of-experience (QoE) measures called SSIMplus [61], to identify distortion parameters that correspond to a particular level of distortion for each reference image. SSIMplus predicts the quality of images on a scale of 0-100 where 100 corresponds to the best while 0 corresponds to the worst quality. A significant advantage of using SSIMplus is that its quality scale was calibrated to be linear with respect to perceptual quality, which means that the loss of quality associated with x SSIMplus points has the same perceptual significance regardless of the starting point on the quality scale, allowing for a division of the quality scale into uniformly spaced intervals. To densely sample the quality spectrum, we choose to have 17 distortion levels for the four base distortion types. These distortion levels, their target SSIMplus scores, and quality categories are depicted in Table 3.2, where it can be seen that we do not go below the SSIMplus score of 20 as the resulting images are severely distorted and do not make a useful contribution to the dataset. For each reference image of the Waterloo Exploration-II database, we use different distortion parameters to create 15,000, 10,000, 101, and 20,000 distorted images for the base distortions of Gaussian white noise, Gaussian blur, JPEG compression, and JPEG2000 compression, respectively. Finally, distortion parameters for each base distortion that lead to SSIMplus scores closest to the target scores of the 17 distortion levels (see Table 3.2) are selected for subsequent database generation. Thus, each of the 3,570 reference images in the Waterloo Exploration-II database has its own set of distortion parameters for each of the four base distortion types.

Table 3.2: Distortion Levels and Target SSIMplus Scores.

Distortion Level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	–	–	–	–
Target SSIM+	100	95	90	85	80	75	70	65	60	55	50	45	40	35	30	25	20	15	10	5	0
Quality Category	Excellent				Good				Fair				Poor				Bad				

Table 3.3: Composition of the Waterloo Exploration-II database.

Reference Images (Pristine Quality)	Stage-1 Distorted Images (Singly Distorted)		Stage-2 Distorted Images (Multiply Distorted)	
Number of Images	Distortion	Number of Images	Distortion Combination	Number of Images
3,570	Blur	39,270	Blur-JPEG	667,590
			Blur-Noise	667,590
	JPEG	39,270	JPEG-JPEG	667,590
	Noise	39,270	Noise-JPEG	667,590
			Noise-JP2K	667,590
	Total	117,810	Total	3,337,950
	Overall 3,455,760 Distorted Images			

Dense Singly and Multiply Distorted Content

To better mimic real-world distortions, we construct the Waterloo Exploration-II database in two stages to include both singly and multiply distorted content, with emphasis on the latter. Table 3.3 outlines the composition of the database.

Stage-1 contains singly distorted images belonging to three distortion types:

1. Gaussian white noise
2. Gaussian blur
3. JPEG compression

Images for each of the three single distortion types mentioned above are obtained by distorting the reference images using their respective content adaptive distortion parameters belonging to Levels 1 to 11, as depicted in Table 3.2. Thus, the 11 Stage-1 distortion levels correspond to the SSIMplus [61] quality range of 50 to 100, which is representative of *fair* to *excellent* perceptual quality. We restrict ourselves to the top half of the perceptual quality spectrum as distorted images in the earlier part of the media distribution pipelines are expected to be in this quality range. This leads to 39,270 singly distorted images for each single distortion category for a total of 117,810 singly distorted images. We have restricted ourselves to just three distortion types for Stage-1 because even they lead to a final dataset consisting of more than 3.45 million distorted images (see Table 3.3), thereby creating significant storage requirements and enhanced time for training machine learning based models. Thus, we have selected distortion types that are most commonly found in IQA literature. Since earlier training-based BIQA models are trained and tested on images of such distortion types, choosing them also provides a fair ground for comparison. However, the inclusion of just three distortion types at Stage-1 remains a limitation of the current work and a more diverse set of distortion types should be included in the future. For example, while thermal noise is approximated as additive white Gaussian noise (AWGN), the noise distribution of real-world camera sensors is better represented by the Poisson distribution [205, 206]. Thus, future releases of the dataset should include a more diverse set of Stage-1 distortion types representing more noise types, blur types, contrast distortions, color distortions, transmission errors, and so on.

Stage-2 contains multiply distorted images belonging to five distortion combinations. Since images taken in the real-world are quite often afflicted with noise (due to limitations of the camera sensor and lighting conditions) and/or blur (due to movement of the photographer/target or limitations of the camera sensor) and are then almost always stored with some form of compression, we choose compression as the second distortion stage in four of the five cases. The distortion combinations are given below along with justifications for selecting them:

1. Gaussian blur followed by JPEG compression (Blur-JPEG) to mimic storing a blurry image through JPEG compression.

2. Gaussian blur followed by Gaussian white noise (Blur-Noise) to mimic different image capture scenarios. For example, capturing a photograph when the camera is moving and lighting conditions are inadequate.

3. JPEG compression followed by JPEG compression (JPEG-JPEG) to mimic multiple levels of compression. For example, a picture taken by a cell phone camera is usually stored after JPEG compression and may undergo another round of compression if it is uploaded to a social media platform.

4. Gaussian white noise followed by JPEG compression (Noise-JPEG) to mimic storing a noisy image through JPEG compression.

5. Gaussian white noise followed by JPEG2000 compression (Noise-JP2K) to mimic storing a noisy image through JPEG2000 compression.

Stage-2 multiply distorted images are obtained by starting from the respective Stage-1 singly distorted images and distorting them by using the content adaptive distortion parameters of the parent reference image belonging to Levels 1 to 17, as depicted in Table 3.2, where it can be seen that this covers SSIMplus quality range of 20 to 100, which is representative of *bad* to *excellent* perceptual quality. Thus, the first distortion in multiply distorted images belongs to the fair to excellent quality range and the subsequent distortion belongs to the entire meaningful quality spectrum (bad to excellent). Each of the five multiple distortion combinations has 667,590 images for a total of 3,337,950 multiply distorted images. Overall, the Waterloo Exploration-II database has 3,455,760 singly and multiply distorted images, which we annotate with synthetic perceptual quality ratings (explained in Section 3.4), making it by far the largest annotated dataset in IQA. As noted earlier for singly distorted Stage-1 images, having multiply distorted images belonging to only five distortion combinations can be considered as a limitation of the current work, and a more diverse set of multiply distorted images should be considered in the future. This can be done by having more combinations in a two-stage distortion pipeline, or by afflicting images with more than two distortions to mimic practical scenarios in an even better manner. For example, an image with blur and noise stored after compression.

Distorted Content Analysis

To observe how well the Waterloo Exploration-II database covers the perceptual quality spectrum, we plot the synthetic quality benchmark (SQB) histogram of the dataset in

Fig. 3.2. The generation of these synthetic quality labels will be explained in Section 3.4. The SQB has a quality range of 0 to 100, where 100 is representative of the best while 0 represents the worst quality. It can be seen from Fig. 3.2 that the Waterloo Exploration-II database has more than at least 10,000 annotated images for each integer quality value above 10, thereby ensuring adequate representation of each quality value. It can also be seen that the quality range of 50 to 100 has the most images, which ensures that the higher quality range, which is difficult to assess for objective IQA methods [19], is adequately represented.

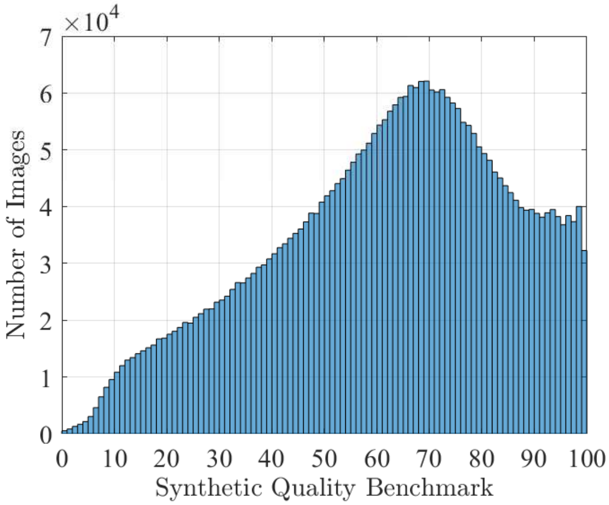


Figure 3.2: SQB histogram of the Waterloo Exploration-II database.

To see how well the Waterloo Exploration-II database covers the quality spectrum in comparison with other well-known IQA datasets, we compute the SQB values of all the distorted images in the Waterloo Exploration-II database and nine well-known IQA datasets, and provide the corresponding boxplots in Fig. 3.3, where the range of distortions in different databases can be directly compared. By observing these boxplots, it becomes clear that while most contemporary IQA datasets tend to favor either the higher or lower end of the quality spectrum, the Waterloo Exploration-II database offers a better spread and more balanced coverage, with the highest concentration at the practically most common mid-to-high quality range.

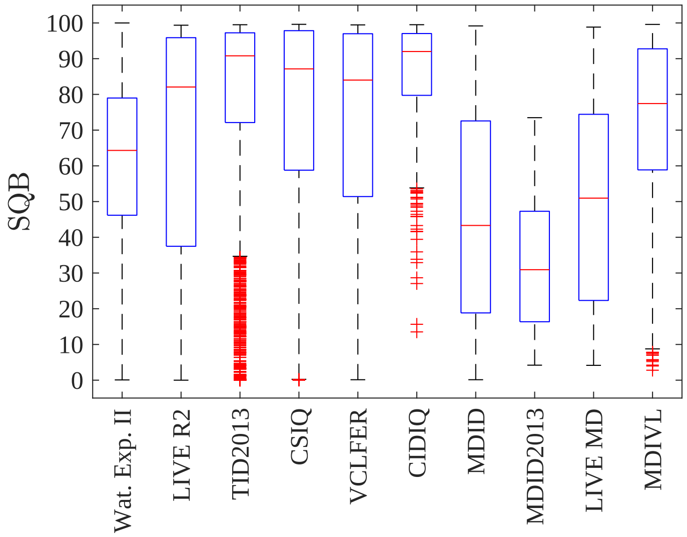


Figure 3.3: SQB box plot of the Waterloo Exploration-II (Wat. Exp. II) database in comparison with nine well-known IQA datasets. The top and bottom edges of the blue boxes represent the 75th and 25th percentiles, respectively. The red line represents the median (50th percentile). The top and bottom black lines represent the extreme data points while the outliers are represented by red + symbols.

3.4 Synthetic Quality Benchmark

3.4.1 Background and Extensive Review

As discussed in 3.2.2, it is not possible to annotate large-scale IQA datasets through human observers, and thus the assignment of perceptual quality annotations through alternative means is necessary. Given that the area of FR IQA has matured quite well (as shown in Chapter 2), one possible alternative is to replace subjective ratings with scores from reliable FR methods. In fact, a number of works in IQA literature have already taken this route. During its training phase, the BIQA method QAC [35] uses the FR method FSIM [14] to annotate image patches with quality scores based on which subsequent grouping is done. The BLISS framework [41] proposes a way to convert opinion-aware BIQA methods into opinion-unaware ones. It first fuses five FR methods (FSIM [14], FSIMc [14], GMSD [99], IWSSIM [13], and VIF [113]) to generate synthetic scores for a dataset composed of 100 reference and 3,200 distorted images. It then uses these synthetic scores to retrain a BIQA method CORNIA [141] which was previously trained through subjective ratings. The BIQA method dipIQ [36] uses three FR methods (GMSD [99], MSSSIM [4], and VIF [113]) to generate 80-million quality-discriminable image pairs (DIPs) from a dataset that has 840 reference and 16,800 distorted images, which are then used to learn a blind quality model. A recent BIQA method called Multiply and Singly distorted Image Quality Estimator (MUSIQUE) [37] uses the FR method VIF [113] in its training stage to find a relationship between estimated distortion parameters and VIF quality scores. The BIQA method in [38, 39] uses FR methods (FSIMc [14], GMSD [99], SSIM [111], and VSI [15]) to derive local scores of 32 × 32 patches and then pre-trains a CNN using these patches with corresponding FR scores. The model is then fine-tuned on a subject-rated dataset. In [40], the FR method MSSSIM [4] is used to annotate four large-scale databases of singly and multiply distorted images, the largest of which is composed of around 2 million images.

While the above-mentioned works demonstrate that FR scores may be used in place of subjective ratings, their choices of FR methods are rather ad hoc and deeper justification and analysis are lacking. The following questions arise when using FR scores for annotating large-scale IQA datasets as alternatives to subjective ratings: 1) Which FR method or methods can be reliably used? 2) Can fused FR methods, which combine the results of multiple FR methods, offer any further advantages over individual ones? We comprehensively answered these questions in Chapter 2, where we carried out the largest performance evaluation study to-date in IQA literature, as a prerequisite requirement of this chapter. In Chapter 2, we compared the performance of 43 FR and seven fused FR methods (22 versions) on nine subject-rated IQA databases (five singly and four multiply distorted) to ensure the diversity of test data. Our results indicated the following: 1) Among individual FR methods, the structural similarity based methods IWSSIM [13], FSIMc [14], VSI [15], and DSS [16], outperform others. 2) However, the performance of even the best individual FR methods varies, at times widely, across different IQA datasets, a point which has earlier been noted in [73]. This puts into question the robustness of individual FR methods, especially when using them as alternatives to human annotations. 3) Among FR fusion methods, learning based fusion methods such as MMF [130] and CNNM [128], and empirical fusion methods such as HFSIMc [129], CM3 [127], and CM4 [127], are outperformed by the best individual FR methods and thus do not offer any advantages. 4) However, the FR fusion method which we called RRF [23] based Adjusted Scores (RAS) in Chapter 2, is found to outperform not only the other fusion based methods, but more importantly, the best individual FR methods. In the literature of IQA, RAS was originally proposed as part of the BLISS framework in [41] and uses a rank aggregation based fusion strategy [23], but no deeper analysis, reasoning or empirical justification was provided. 5) The performance of RAS is found to be more stable across different IQA datasets relative to individual FR methods. Thus, the training-free rank aggregation based fusion strategy [23] is a strong candidate for synthetically annotating large-scale IQA datasets.

3.4.2 Synthetic Quality Benchmark Generation

At the core of the rank aggregation based fusion strategy is the training-free Reciprocal Rank Fusion (RRF) algorithm [23], which was first developed to combine document rankings from multiple information retrieval systems in an unsupervised manner. For a given set of test images and their associated quality scores as assigned by different FR IQA methods, a consensus ranking can be obtained in terms of RRF as follows [23]:

$$RRF_{score}(I_i) = \sum_{j=1}^{J} \frac{1}{k + r_j(i)} \tag{3.1}$$

where J is the number of FR methods being fused, $r_j(i)$ is the rank given by the j-th FR method to the image I_i, $RRF_{score}(I_i)$ is the RRF score of image I_i, and k is a stabilization constant. RRF was first used in IQA as part of the BLISS framework [41], which replaces human opinion scores with synthetic quality scores that act as ground truth data to train BIQA methods. The BLISS framework [41] produces synthetic quality scores in two steps for a given set of images: 1) Generation of a consensus ranking score through RRF [23], and 2) Since the ranking score is a measure of quality relative to other images and cannot be considered an independent quality measure, the scores of a base FR method are adjusted based on the consensus ranking, which then act as synthetic quality scores. The latter step is required because the BLISS framework operates in the absence of subject-rated datasets. The choice of FR methods to combine in [41] is ad hoc. To test the rank aggregation based fusion of FR methods more thoroughly, we performed an exhaustive search by testing 737,280 different combinations of 2 to 15 FR methods, and finalized 13 versions of RAS in Chapter 2. Among them, RAS6 was found to be the top performer and will be used as a basis for the generation of our synthetic data annotation, as explained next.

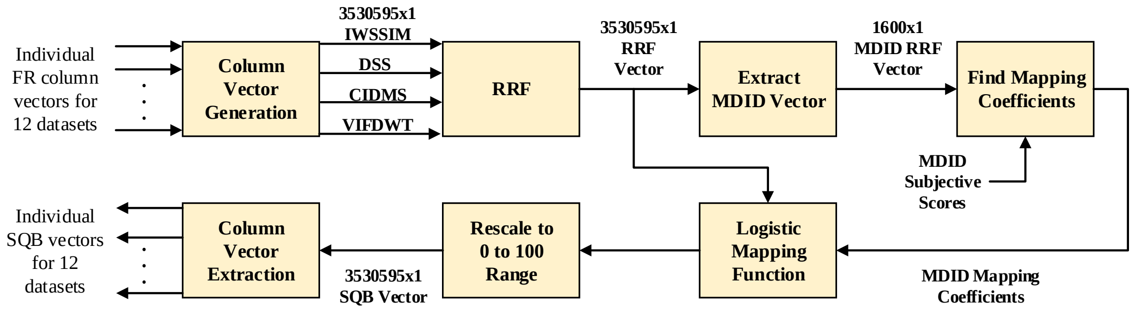


Figure 3.4: Synthetic Quality Benchmark (SQB) generation procedure.

To generate the Synthetic Quality Benchmark (SQB) for the very large-scale Waterloo Exploration-II database, we use RRF [23] to fuse the same four FR methods as in RAS6 (see

Table 3.4: Individual database and concatenated column vector sizes for SQB generation.

S.	Database	Subject	Database Column	Concatenated Column
No.		Rated	Vector Size	Vector Size
1	DR IQA V1	No	32912×1	3530595x1
2	DR IQA V2	No	32912×1	
3	Waterloo Exp.-II	No	3455760×1	
4	LIVE R2 [24]	Yes	779×1	
5	TID2013 [19]	Yes	3000×1	
6	CSIQ [26]	Yes	866×1	
7	VCLFER [54]	Yes	552×1	
8	CIDIQ [5]	Yes	690×1	
9	MDID [33]	Yes	1600×1	
10	MDID2013 [32]	Yes	324×1	
11	LIVE MD [31]	Yes	450×1	
12	MDIVL [34]	Yes	750×1	

Chapter 2), that is, IWSSIM [13], DSS [16], CID_MS [95], and VIF_DWT [93]. However, unlike RAS (BLISS framework [41]), we do not adjust the score of a base FR method to generate the synthetic quality scores. Instead, we use the novel framework shown in Fig. 3.4 to generate the SQB. First, we acquire the scores of the above-mentioned four FR methods for 12 databases mentioned in Table 3.4. These include nine subject-rated datasets (serial number 4 to 12) which have been described earlier in Section 2.2 and in Tables 2.2 and 3.1, and three large datasets which include the Waterloo Exploration-II database, and two other datasets called DR IQA V1 and DR IQA V2, which have been developed in a manner similar to the Waterloo Exploration-II database. These latter two datasets will be used in Chapter 4 and will be discussed more in detail there. For each dataset, we obtain scores for each FR method in terms of database-wide column vectors, which are all then concatenated into one large column vector of size 3530595×1 for each FR method, as depicted in Table 3.4. Next, RRF is used to fuse the four large column vectors, through Equation 3.1, resulting in an RRF vector which contains the consensus ranking. Since the RRF process involves sorting the constituents of individual vectors being fused, this results in punctuating the scores of the three large *unannotated* databases with scores of the nine databases that do have human annotations of quality. This is done without subjects rating the images of the three large databases, and meanwhile allows us to evaluate the performance of SQB compared to actual human annotations in Section 3.4.3. The RRF vector is normalized as follows:

$$RRF_{normalized} = \frac{RRF_{score}}{max(RRF_{score})} \tag{3.2}$$

As discussed earlier, the outcome of the RRF step leads to a quality rating for an image relative to other images. To be considered independently, the ratings from the consensus RRF ranking can be mapped to a subjective quality scale by using a subset of RRF scores for which subjective quality scores are available. Since the MDID database [33] has uniform representation from different parts of the quality spectrum (as discussed in Chapter 2), we choose it to learn this mapping through a five-parameter modified logistic function [24]:

$$S(R) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\{\beta_2 (R - \beta_3)\}}} \right] + \beta_4 R + \beta_5 \tag{3.3}$$

where R denotes the RRF scores of the MDID database that have been extracted from the overall normalized RRF vector, S denotes the predicted MDID subjective quality scores, and β_1, β_2, β_3, β_4, and β_5 are mapping coefficients that are found numerically to maximize the correlation between MDID subjective quality scores and its RRF scores. Since the MDID RRF scores punctuate the overall RRF vector in a regular manner, the MDID mapping coefficients are then used to map the entire RRF vector to the MDID subjective quality scale (0 to 8) by again using Equation 3.3. These quality scores, denoted by Q, are then rescaled to the 0 to 100 range as:

$$SQB = 100 \times \frac{Q - min(Q)}{max(Q - min(Q))} \tag{3.4}$$

Equation 3.4 results in the concatenated Synthetic Quality Benchmark (SQB) vector for all databases involved, and ensures that the rescaling process does not disturb the distribution of the quality scores. Finally, the SQB vectors for individual databases are extracted from the overall SQB vector.

It is mentioned in [23] that the constant k in Equation 3.1 counters the impact of high rankings by outliers and its value was set at 60 through a pilot investigation. This value of the constant k was also used in the BLISS framework [41]. While the value of k may not be critical when the number of data points is small, we believe that it takes on a more significant role when the number of objects to be ranked is large. In our case, with around 3.53 million images, a small value of $k = 60$ leads to weights assigned to rank 1 and to rank 3,530,595 that differ by several orders of magnitude. Thus, some ranks are more favored than others and a level playing field is absent. We believe that in order to rank a large number of objects, the value of the constant k should be proportionately higher. To test this hypothesis, we carry out an empirical study where the value of k was progressively increased in terms of order of magnitude, and the overall SQB vector was recomputed each time. The SQB scores for each of the nine subject-rated databases mentioned in Table 3.4 were extracted for each value of k. For each database we compute the SRCC of the SQB

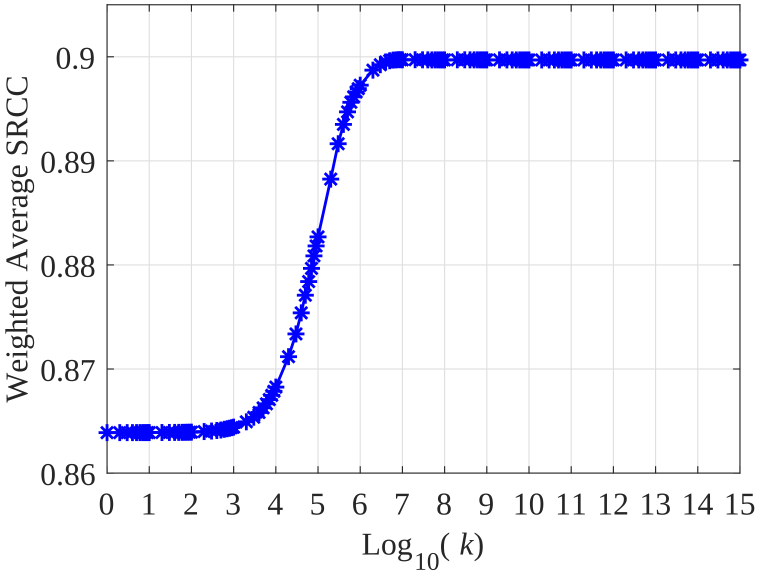


Figure 3.5: Weighted average SRCC of SQB with respect to subjective scores of the nine subject-rated databases (see Table 3.4) for different values of the RRF constant k.

with respect to the respective subjective scores. The weighted average SRCC values for the nine subject-rated databases for various values of k are depicted in Fig. 3.5, where it can be seen that our hypothesis is indeed correct. Given that we have around 3.53 million images for which RRF is being computed, the weighted average SRCC starts increasing as k goes beyond 10^4 and keeps on increasing until k attains a value of around 10^7 beyond which it remains constant. We believe that further increase of the value of k does not lead to further SRCC gain as the weights assigned to all the ranks remain within the same order of magnitude. Through our empirical investigation, we have found that for 3.53 million images, the weighted average SRCC does not increase beyond $k = 8 \times 10^6$, and hence this value of k has been used in this work.

3.4.3 SQB Performance

Databases and Methods used for Comparison

To comprehensively test the perceptual quality prediction performance of SQB, we test it on the nine subject-rated IQA databases that were made part of the SQB computation process as discussed in the previous section. Five of these databases include singly distorted content and include LIVE R2 [24], TID2013 [19], CSIQ [26], VCLFER [54], and CIDIQ [5], while four contain multiply distorted content and include MDID [33], MDID2013 [32], LIVE MD [31], and MDIVL [34]. It should be noted that the CIDIQ database [5] contains subject-ratings at two viewing distances, that of 50 cm and 100 cm, for all its images. We shall refer to results for these two sets of subjective data as CIDIQ50 and CIDIQ100, respectively. The main features of these databases are given in Table 3.1, while more details can be found in Chapter 2 or in their respective papers.

For a thorough comparison, we tested the performance of other state-of-the-art methods, including two fused and 14 individual FR methods, on the above-mentioned datasets. The fused FR methods include the rank aggregation based fusion method RAS6 which was the overall top performer in the performance evaluation survey that we performed in Chapter 2, and uses the approach described in [41]. We also include the learning based fusion method MMF [130] in our comparison. The 14 individual FR methods are top performers in the performance evaluation study of Chapter 2 and belong to three state-of-the-art FR design philosophies. Among them, eight methods belong to the *structural similarity* based design philosophy and include CID_MS [95], DSS [16], ESSIM [98], FSIMc [14], GMSD [99], IWSSIM [13], MCSD [102], and VSI [15], four are *natural scene statistics* (NSS) based and include QASD [107], SFF [109], VIF [113], and VIF_DWT [93], and two belong to the *mixed strategy* based design philosophy and include DVICOM [96] and MAD [26]. We also include the *error based* method PSNR for legacy purposes.

Evaluation Criteria

We use five evaluation criteria to evaluate the performance of methods under test. For assessing *prediction accuracy*, we use the Pearson Linear Correlation Coefficient (PLCC)

[72]. The scores generated by objective IQA methods are usually not linear with respect to subjective ratings. Thus, a nonlinear mapping step is required before the computation of PLCC. To do this, we adopt the five-parameter modified logistic function used in [24] and given in Equation 3.3. PLCC is then computed at the database-level between the subjective scores and the objective scores after passing through the nonlinear mapping step. We assess *prediction monotonicity* by using the Spearman Rank-order Correlation Coefficient (SRCC) [72]. SRCC is a non-parametric rank-order based correlation measure. It does not require the preceding nonlinear mapping step. The absolute value of both PLCC and SRCC lies in the 0 to 1 range. A better objective IQA method should have higher PLCC and SRCC values with respect to subjective scores, where a value of 1 would indicate perfect perceptual performance. Since we are using nine different IQA databases for performance evaluation (ten if the two viewing distance of CIDIQ database are considered separately) and PLCC/SRCC values are at the individual database-level, trying to make conclusions about the overall performance becomes cumbersome and a measure of aggregate performance is required. We provide this measure by calculating the weighted average (WA) PLCC and WA SRCC values for each IQA method across all databases (as in [13] and in Chapter 2). The total number of distorted images in a database defines the weight assigned to it in the weighted average computation. The CIDIQ database [5] is considered twice in this calculation due to its evaluations being at two viewing distances. Finally, we perform statistical significance testing (hypothesis testing) to draw statistically sound and generalizable inferences about the performance of an IQA method compared to another. We carried out these tests on the prediction residuals of different methods for each database. These residuals were obtained by first mapping the IQA method outcomes to subjective scores by using the nonlinear mapping approach described above for PLCC calculation, and then subtracting these predictions from the actual subjective scores. We use the one-sided (left-tailed) two-sample F-test [175] to statistically compare the performance of two IQA methods with each other at the 5% significance level (95% confidence) for each of the IQA databases. By carrying out this test twice, with the order of the methods reversed, we were able to determine if the method performance was statistically indistinguishable or if one method performed better than another. Since these tests assume the Gaussianity of residuals, we used a simple kurtosis

based check for Gaussianity (as in [24]), where Gaussianity is assumed if the kurtosis of the residuals is between 2 and 4. The databases and evaluation criteria used in this chapter are the same as in Chapter 2, thus results in this chapter can be directly compared with other methods discussed there. For a more detailed description of these evaluation criteria, refer to Section 2.4.1 of Chapter 2.

SQB Performance Evaluation

Since the very-large scale Waterloo Exploration-II database does not have subjective ratings, it is not possible to evaluate the performance of its SQB annotation scores directly. The reason why we concatenated the objective score vectors, belonging to the nine subject-rated IQA datasets, of the FR methods being fused in SQB computation, with those of the Waterloo Exploration-II, DR IQA V1, and DR IQA V2 databases, is that it allowed us to punctuate data without subject ratings with data that does have these ratings. Thus, in the overall SQB vector, the SQB scores for the nine subject-rated datasets act as regularly distributed samples. Since these samples also have subjective scores available, this allows us to comprehensively test the performance of SQB.

The perceptual quality prediction performance, of all methods under test for the nine IQA datasets, in terms of PLCC is given in Table 3.5. As mentioned earlier, the CIDIQ database [5] is considered as two datasets since it has subjective ratings at two viewing distances. For each IQA database, all of its constituent distortions were included for testing. The weighted average PLCC is provided in the rightmost column of Table 3.5 and is used to sort the methods in the descending order. Thus, the best performing methods are towards the top of the table. The names of the fused FR methods are mentioned in bold, to distinguish them from the individual FR methods. Similarly, Table 3.6 provides the perceptual quality prediction performance of all methods under test in terms of SRCC. Again, all distortions in each dataset were considered. The weighted average SRCC is provided in the rightmost column of Table 3.6 and methods have been sorted in the descending order with respect to these values. The results of statistical significance testing of SQB relative to the 17 other methods are shown in Table 3.7, where a “1”, “–”, or “0” means that the perceptual quality prediction performance of SQB is better, indistinguishable, or

Table 3.5: Test results of SQB, 2 fused FR, 14 state-of-the-art FR methods, and PSNR, on nine subject-rated IQA databases in terms of PLCC. All distortions in each dataset were considered. The Weighted Average PLCC (WA PLCC) is provided in the rightmost column and methods are sorted in descending order with respect to WA PLCC. Fused FR methods are highlighted in bold.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MD IVL	WA PLCC
SQB (Proposed)	0.9612	0.8917	0.9596	0.9408	0.8745	0.8742	0.9293	0.8152	0.9144	0.9126	0.9100
RAS6 [41]	0.9682	0.8488	0.9640	0.9408	0.8832	0.8585	0.9294	0.8181	0.9150	0.9202	0.8979
IWSSIM [13]	0.9522	0.8319	0.9144	0.9191	0.8476	0.8698	0.8983	0.8513	0.9109	0.9056	0.8787
FSIMc [14]	0.9613	0.8769	0.9191	0.9329	0.7583	0.8410	0.8998	0.6429	0.8965	0.9039	0.8786
DSS [16]	0.9618	0.8530	0.9612	0.9259	0.7715	0.8267	0.8733	0.8168	0.9023	0.8973	0.8757
VSI [15]	0.9482	0.9000	0.9279	0.9320	0.7226	0.8240	0.8703	0.5732	0.8789	0.8749	0.8714
MCSD [102]	0.9675	0.8648	0.9560	0.9217	0.7532	0.7727	0.8637	0.8281	0.8847	0.8787	0.8705
GMSD [99]	0.9603	0.8590	0.9541	0.9176	0.7387	0.7585	0.8776	0.8336	0.8808	0.8685	0.8672
ESSIM [98]	0.9566	0.8645	0.9224	0.9094	0.7953	0.8256	0.8451	0.6648	0.8861	0.9081	0.8664
SFF [109]	0.9632	0.8706	0.9643	0.7761	0.7834	0.7721	0.8590	0.7952	0.8893	0.8904	0.8658
QASD [107]	0.9574	0.8897	0.9481	0.9253	0.7257	0.8116	0.8063	0.6698	0.8966	0.8827	0.8638
DVICOM [96]	0.9734	0.8194	0.9191	0.9144	0.8035	0.8018	0.8919	0.8161	0.8873	0.8773	0.8632
MMF [130]	0.8561	0.9504	0.9262	0.8624	0.7326	0.7572	0.8185	0.6788	0.8523	0.8075	0.8600
CID_MS [95]	0.9159	0.8362	0.8732	0.9375	0.8364	0.8171	0.8414	0.6155	0.8917	0.8961	0.8510
MAD [26]	0.9675	0.8267	0.9502	0.9053	0.7809	0.8541	0.7552	0.7471	0.8944	0.8985	0.8464
VIF [113]	0.9604	0.7720	0.9278	0.8938	0.7267	0.6415	0.9367	0.8376	0.9030	0.8736	0.8388
VIF_DWT [93]	0.9657	0.7657	0.9123	0.8969	0.7259	0.5845	0.9031	0.7264	0.8839	0.8653	0.8211
PSNR	0.8723	0.7017	0.8000	0.8321	0.6302	0.6808	0.6164	0.5647	0.7398	0.6806	0.7065

worse, respectively, than that of the method in the row for a given database (with 95% confidence). We preceded the statistical significance testing with a kurtosis based check for Gaussianity of prediction residuals of all methods under test on all datasets (described earlier in this section) and found that the assumption of Gaussianity holds in around 79% cases.

It can be clearly seen from Tables 3.5 and 3.6 that SQB is the top performer in terms of both WA PLCC and WA SRCC. From Table 3.7, it can be observed that for the 170 method-database combinations, SQB performs statistically better than the best of all other methods in around 74% cases, while its performance is statistically indistinguishable or inferior than other methods in around 19% and 6% cases, respectively. This is no small

Table 3.6: Test results of SQB, 2 fused FR, 14 state-of-the-art FR methods, and PSNR, on nine subject-rated IQA databases in terms of SRCC. All distortions in each dataset were considered. The Weighted Average SRCC (WA SRCC) is provided in the rightmost column and methods are sorted in descending order with respect to WA SRCC. Fused FR methods are highlighted in bold.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MD IVL	WA SRCC
SQB (Proposed)	0.9665	0.8749	0.9542	0.9421	0.8760	0.8651	0.9252	0.8045	0.8857	0.8845	0.8997
RAS6 [41]	0.9680	0.7930	0.9603	0.9405	0.8840	0.8532	0.9250	0.8214	0.8867	0.8954	0.8761
VSI [15]	0.9524	0.8965	0.9422	0.9317	0.7213	0.8106	0.8569	0.5700	0.8414	0.8269	0.8631
FSIMc [14]	0.9645	0.8510	0.9309	0.9323	0.7608	0.8285	0.8904	0.5806	0.8666	0.8613	0.8628
IWSSIM [13]	0.9567	0.7779	0.9212	0.9163	0.8484	0.8564	0.8911	0.8551	0.8836	0.8588	0.8559
SFF [109]	0.9649	0.8513	0.9627	0.7738	0.7834	0.7689	0.8396	0.8005	0.8700	0.8535	0.8527
DSS [16]	0.9616	0.7921	0.9555	0.9272	0.7755	0.8246	0.8658	0.8078	0.8714	0.8759	0.8520
QASD [107]	0.9629	0.8674	0.9530	0.9231	0.7307	0.8079	0.7778	0.6687	0.8766	0.8315	0.8482
MMF [130]	0.8741	0.9409	0.9043	0.8594	0.7241	0.7379	0.8084	0.6799	0.8085	0.7703	0.8479
MCSD [102]	0.9668	0.8089	0.9592	0.9224	0.7562	0.7808	0.8451	0.8269	0.8517	0.8370	0.8464
CID_MS [95]	0.9103	0.8314	0.8789	0.9366	0.8350	0.8062	0.8330	0.6168	0.8608	0.8778	0.8445
GMSD [99]	0.9603	0.8044	0.9570	0.9177	0.7427	0.7675	0.8613	0.8283	0.8448	0.8210	0.8433
ESSIM [98]	0.9597	0.8035	0.9325	0.9075	0.7968	0.8253	0.8250	0.6966	0.8517	0.8682	0.8418
DVICOM [96]	0.9750	0.7598	0.9181	0.9155	0.8034	0.7903	0.8840	0.8168	0.8672	0.8374	0.8387
MAD [26]	0.9669	0.7807	0.9466	0.9061	0.7815	0.8391	0.7249	0.7507	0.8646	0.8643	0.8220
VIF [113]	0.9636	0.6769	0.9194	0.8866	0.7203	0.6257	0.9306	0.8444	0.8823	0.8381	0.8024
VIF_DWT [93]	0.9681	0.6439	0.9020	0.8930	0.7224	0.5826	0.8943	0.7553	0.8479	0.8243	0.7768
PSNR	0.8756	0.6394	0.8057	0.8246	0.6254	0.6701	0.5784	0.5604	0.6771	0.6136	0.6720

achievement given that all other methods included in the comparison, apart from PSNR, are considered state-of-the-art in FR and fused FR IQA. While RAS6 was the top performer in the comprehensive performance evaluation study in Chapter 2, it did not perform as well on the TID2013 database [19], as can be seen from its PLCC and SRCC values in Tables 3.5 and 3.6, respectively. With 3,000 distorted images and as many as 24 different distortion types, TID2013 can be considered as one of the largest and most diverse subject-rated IQA databases, making it quite challenging. It is clear from Tables 3.5 and 3.6 that SQB performs quite well on the TID2013 database, when compared to other methods. From Table 3.7, it can be seen that on the TID2013 database SQB is outperformed only by the fused FR method MMF [130] and the FR method VSI [15]. MMF is a learning-based fusion

Table 3.7: Statistical significance testing of SQB through the F-Test with respect to fused and individual FR methods on different IQA databases. A "1" means that SQB performance is statistically better than the method in the row, a "0" means that it is statistically worse, while a "–" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). Fused FR methods are highlighted in bold.

FR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
CID_MS [95]	1	1	1	–	1	1	1	1	1	1
DSS [16]	–	1	–	1	1	1	1	–	–	1
DVICOM [96]	0	1	1	1	1	1	1	–	1	1
ESSIM [98]	–	1	1	1	1	1	1	1	1	–
FSIMc [14]	–	1	1	–	1	1	1	1	1	–
GMSD [99]	–	1	1	1	1	1	1	–	1	1
IWSSIM [13]	1	1	1	1	1	–	1	0	–	–
MAD [26]	0	1	1	1	1	1	1	1	1	1
MCSD [102]	0	1	–	1	1	1	1	–	1	1
MMF [130]	1	0	1	1	1	1	1	1	1	1
PSNR	1	1	1	1	1	1	1	1	1	1
QASD [107]	–	–	1	1	1	1	1	1	1	1
RAS6 [41]	0	1	0	–	–	–	–	–	–	–
SFF [109]	–	1	0	1	1	1	1	–	1	1
VIF [113]	–	1	1	1	1	1	0	–	–	1
VIF_DWT [93]	0	1	1	1	1	1	1	1	1	1
VSI [15]	1	0	1	–	1	1	1	1	1	1

method and we trained it on the TID2013 database. Thus, it is unfair to compare other methods with MMF on TID2013. While VSI outperforms SQB on TID2013, SQB performs better than VSI on almost all other datasets. We believe that the performance gain of SQB on the TID2013 database, especially when compared to RAS6, is explained by the way we have selected the constant k in the RRF [23] computation (Equation 3.1), as explained in Section 3.4.2. From Table 3.7, we can see that SQB is outperformed by RAS6 on the LIVE R2 [24] and CSIQ [26] databases. Tables 3.5 and 3.6 show that RAS6 outperforms SQB only slightly in terms of PLCC and SRCC, respectively, on these datasets. Given that the TID2013 dataset is much more diverse, in terms of distortions, when compared to LIVE R2 and CSIQ, we believe that this performance compromise is justified. From Table 3.7, it can be seen that some individual FR methods statistically outperform SQB on at most

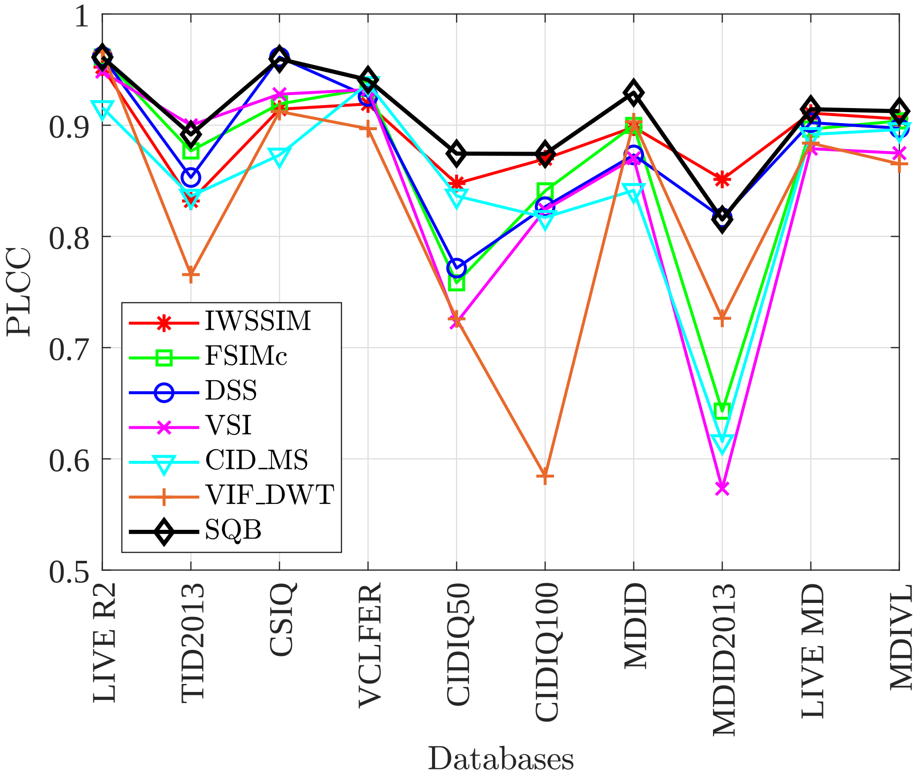


Figure 3.6: PLCC of SQB and selected FR methods for different IQA databases.

a single dataset, but are outperformed by SQB on almost all other datasets. For example, while IWSSIM [13] outperforms SQB on MDID2013 database [32], it is outperformed by SQB on six other datasets (Table 3.7), sometimes quite significantly, such as on TID2013 database [19] (Tables 3.5 and 3.6). In fact, all state-of-the-art individual FR methods perform inconsistently across different IQA datasets, where they perform well on some datasets but not on others. This behavior can be seen in Fig. 3.6, where the PLCC of SQB and some state-of-the-art FR methods is plotted for different datasets. These FR methods include IWSSIM [13], FSIMc [14], DSS [16], and VSI [15], which were found to be the top performers in our comprehensive study of Chapter 2, out of a total of 43 FR methods. We have also included the FR methods CID_MS [95] and VIF_DWT [93] in Fig. 3.6 as they, together with IWSSIM and DSS, are fused together in SQB. From Fig. 3.6, it can be seen that the six individual FR methods encounter wide swings in performance across different datasets which differ from each other in terms of their constituent distortions and content. Thus, the performance of these FR methods cannot be regarded as stable, which goes against their use as alternatives to human annotations for large-scale datasets. However, Fig. 3.6 also shows that the performance variations are much less pronounced for SQB across all datasets. Hence, the performance of SQB can be regarded as stable regardless of the distortions that afflict the images that it evaluates, thereby making it a much more suitable candidate to replace human annotations for labeling large-scale IQA datasets. We believe that SQB displays this superior performance relative to individual state-of-the-art FR methods because: 1) It uses a rank aggregation based fusion approach, RRF [23], that is unsupervised and training-free, which makes it robust to unseen data, and 2) The deficiencies of some FR methods being fused through RRF, for particular distortions, are supplemented by the strengths of other FR methods for those distortions, and thus the fused combination achieves stable performance for all distortions that usually afflict visual content. It is pertinent to mention here that the four FR methods being fused in SQB have not been randomly selected, but through an exhaustive search that included evaluating 737,280 FR fusion combinations in Chapter 2.

3.5 EONSS - A DNN Based BIQA Model

To test the validity of our hypothesis that a large-scale annotated training database will enhance the performance of DNN based BIQA models and to validate our SQB approach of synthetically annotating such a database, as described in Section 3.4, we build a DNN based BIQA model called End-to-end Optimized deep neural Network using Synthetic Scores (EONSS). It should be noted that EONSS aims to provide a transparent common testing platform on which the impact of training data on the performance of DNNs can be assessed, and thus special or sophisticated designs of DNN architectures are avoided that may complicate the interplay between training data and network architecture on their contributions to the overall performance. While we train EONSS on the Waterloo Exploration-II database, we test it on nine subject-rated IQA datasets. It should be noted that there is no overlap between the training and testing data. This allows us to not only rigorously test EONSS on unseen human-rated data but also to make comparisons with other BIQA methods on such data. This also enables us to show the strength of our synthetic quality annotation process.

3.5.1 Network Architecture and Implementation Details

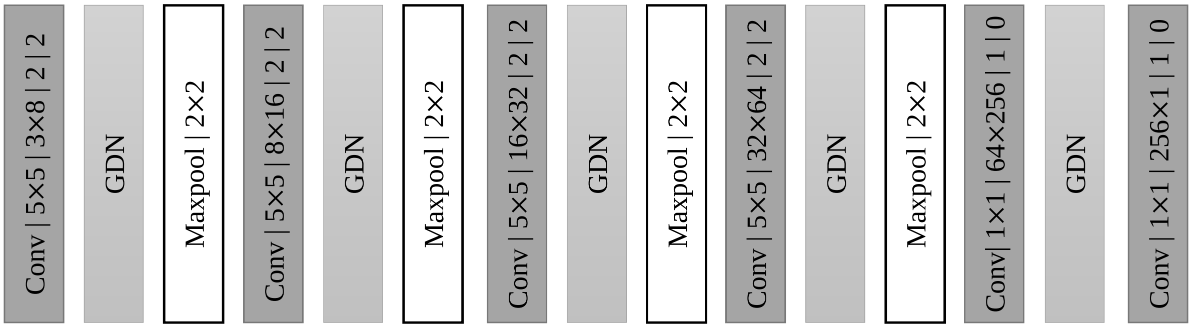

Conv | 5×5 | 3×8 | 2 | 2
GDN
Maxpool | 2×2
Conv | 5×5 | 8×16 | 2 | 2
GDN
Maxpool | 2×2
Conv | 5×5 | 16×32 | 2 | 2
GDN
Maxpool | 2×2
Conv | 5×5 | 32×64 | 2 | 2
GDN
Maxpool | 2×2
Conv| 1×1 | 64×256 | 1 | 0
GDN
Conv | 1×1 | 256×1 | 1 | 0

Figure 3.7: The architecture of the EONSS network for the BIQA task. We adopt the style and convention of [6] and denote the parameterization of the convolutional layer as: *Conv* | *kernel height* × *kernel width* | *input channel* × *output channel* | *stride* | *padding*.

The architecture of the EONSS network is illustrated in Fig. 3.7. The network takes a 235 × 235 RGB color image patch as input and predicts its quality in terms of a scalar value. With a few exceptions, most DNN based BIQA methods that have been proposed so far, use a smaller input patch size of 32 × 32, as discussed in Section 3.2.2. On the other hand, patches of larger size, such as 235 × 235, contain visually more meaningful content than smaller patches and can better represent the parent image, thereby reducing the label noise problem. Since earlier DNN based BIQA methods train on small-scale subject-rated datasets, they use a smaller input patch size, as a larger patch size would dramatically reduce the overall number of patches available for training. However, our model does not suffer from this issue since the Waterloo Exploration-II database has a sufficiently large number of training images. As can be seen from Fig. 3.7, the EONSS network consists of six stages of processing. The first four stages each contain a convolutional, a generalized divisive normalization (GDN) [122], and a max-pooling layer. The purpose of these four stages is to map the $235 \times 235 \times 3$ raw pixels from the image space to a lower-dimensional feature space where the impact of distortions on image quality can be more easily quantified in a perceptually aware manner. In the first four stages, the network reduces the spatial dimension through the use of convolution with stride 2×2 and employs 2×2 max-pooling after each GDN layer to select neurons with the highest local response. The last two stages, which consist of two fully connected layers and a GDN transform layer in between, map the extracted features to a single quality score. Before being sent to the last two fully connected layers, the spatial size of the features is reduced to 1×1, so that the number of weights in the fully connected layers is considerably reduced. Instead of using ReLU [157], we use GDN [122] as the activation function after the convolution layers in the first five stages of the network to add non-linearity to the model. While ReLU [157] is widely used as the activation function in CNNs, it suffers from strong higher-order dependencies, thus requiring a much larger network to achieve good performance for a given task [6]. We utilize a bio-inspired normalization transform, GDN [122], as the activation function because it helps decorrelate the high-dimensional features by using a joint nonlinear gain control mechanism, thereby enabling a much smaller network to achieve competitive performance. The GDN transform has been previously used effectively in image compression [6] and has also been used in a DNN based BIQA model [146]. We define the loss function as the negative of PLCC. The advantages of choosing PLCC over MAE or MSE are [207]: 1) The range of predictions is no longer restricted to the range of targets, since it is known that the range of subjective quality scores could be set arbitrarily and does not have any physical meaning. 2) It automatically normalizes the loss to the range [-1, 1], which gives more stability and flexibility to the training process. 3) PLCC is differentiable and is a frequently used evaluation criteria in the area of perceptual IQA. To empirically verify that our choice of PLCC as the loss function is valid, we also trained EONSS with MSE as the loss function. By using the nine subject-rated datasets (all distortions), mentioned in Section 3.4.3, and the WA SRCC evaluation criteria, also mentioned in Section 3.4.3, we found that the WA SRCC of EONSS relative to subjective data is 0.6183 and 0.6509 when using MSE and PLCC as loss functions, respectively. This clearly demonstrates the superiority of using PLCC as the loss function for EONSS.

To train the EONSS model, we randomly split the Waterloo Exploration-II database into training, validation and testing sets that consist of 60%, 20% and 20% of the dataset, respectively. Since the network accepts images of size $235 \times 235 \times 3$, for the sake of speeding up the training phase, we randomly sample one 235×235 patch from each training image if its dimensions are larger. This does not prevent us from creating a sufficiently large pool of training data given the very large-scale nature of the Waterloo Exploration-II database. This also allows us to obtain a batch of image patches that have greater diversity, thereby helping to prevent model overfitting. Since the 235×235 patch size can cover a relatively large area of the original image, thereby containing perceptually meaningful content, we assign the SQB quality score of the original image as the image quality label of the sampled patch. During the validation and testing phases, we consider the entire image instead of just one patch from it. Thus, for images with larger dimensions, we sample 235×235 patches from the original image with a stride of 128×128 in an overlapping manner, and consider the average of the predicted quality scores of all patches as the predicted quality of the original image. This ensures a more rigorous validation process and also that all parts of an image are considered while testing. We initialize the weights of the convolution layers by following the approach in [208] and use Adam [209] for optimization. The training batch size is chosen to be 50 and the image patches in each batch are randomly sampled from the training set only. We start with a learning rate of 0.001 which is decreased by a factor of 10 after every two epochs. Other parameters of Adam [209] are set as default. The model performance, in terms of PLCC and SRCC, is tested on the validation set after each epoch and we stop training after 10 epochs when the performance on the validation set reaches a plateau. Finally, the model after 10 epochs of training, is applied to the testing set.

When compared to many recent DNN based BIQA models, discussed in Section 3.2.2, it can be seen that EONSS uses a relatively simple network architecture. We have favored simplicity because our focus is not on the design of DNN architectures, but instead on the impact of training data on this task. While it will become apparent in Section 3.5.2 that EONSS outperforms the very state-of-the-art in BIQA, our primary goal of constructing it, is to validate our approach of using synthetically annotated very-large scale IQA datasets for DNN model training.

3.5.2 EONSS Performance Evaluation

To analyze the performance of EONSS and other BIQA methods, we use the same set of test datasets as mentioned in Section 3.4.3, which includes five singly and four multiply distorted subject-rated IQA databases. We also use the same evaluation criteria as described in Section 3.4.3. However, we compute the evaluation metrics for two categories of data: 1) The *all distortions category* includes all distorted images within each test dataset, that is, all distortion types are considered while computing PLCC, SRCC, and performing statistical significance testing. 2) The *subset distortions category* includes a subset of distortion types in each dataset for which evaluation metrics are calculated. For singly distorted IQA datasets (LIVE R2 [24], TID2013 [19], CSIQ [26], VCLFER [54], and CIDIQ [5]), images belonging to the following four common distortion types form the subset: 1) Noise, 2) Gaussian Blur, 3) JPEG compression, and 4) JPEG2000 compression. Although Poisson noise is used in the CIDIQ database [5] and additive white Gaussian noise is used in the other four singly distorted datasets, we do not make a distinction between the two for the purpose of subset performance evaluation. For multiply distorted datasets, subsets of distorted images are formed by separately considering individual distortion combinations (if possible). Thus, we separately consider the Blur-JPEG and Blur-Noise combinations in the LIVE MD database [31], and the Blur-JPEG and Noise-JPEG combinations in the MDIVL database [34]. Since images in the MDID [33] and MDID2013 [32] databases cannot be split into subsets, the entire datasets are considered for the subset case as well.

The motivation for conducting performance evaluation on the *subset distortions category* of IQA datasets, especially for singly distorted datasets, stems from the fact that most training-based opinion-aware BIQA methods are trained for the above-mentioned common distortion types that are present in almost all singly distorted datasets. Therefore, these subsets of distortions provide a fair ground for comparison. However, the ultimate goal of NR or *blind* IQA methods is to be robust to *unseen* data, thus, the *all distortions category* of IQA datasets, allows for more rigorous testing of BIQA methods. Any gap in performance for these two categories of test data would highlight directions for future research. We do not retrain BIQA methods on individual datasets but use the original versions, that is EONSS trained on the Waterloo Exploration-II database and author-trained versions of other BIQA methods, again to ensure rigorous testing.

Performance Comparison with State-of-the-Art BIQA Methods

In addition to evaluating the performance of EONSS, we also tested the performance of 14 other state-of-the-art BIQA methods on the test data so that we can situate EONSS relative to the best in the field. Among them, eight methods belong to the *opinion-aware* (OA) BIQA category and include BIQI [139], BRISQUE [140], CORNIA [141], GWHGLBP [142], HOSA [143], MEON [146], NRSL [147], and WaDIQaM-NR [148], while six methods belong to the *opinion-unaware* (OU) category and include dipIQ [36], ILNIQE [144], LPSI [145], NIQE [3], QAC [35], and SISBLIM [32]. It should be noted that among these methods, MEON [146] and WaDIQaM-NR [148] are DNN based BIQA methods. While a number of other deep learning based BIQA methods have recently been proposed, as discussed in Section 3.2.2, we have tested the performance of MEON [146] and WaDIQaM-NR [148] as their author-trained models are publicly available. As an additional comparison point, in subsequent analysis we also include results for FR methods IWSSIM [13] and PSNR in order to compare the performance of EONSS and other BIQA methods with a state-of-the-art (IWSSIM) and legacy (PSNR) FR method.

For all datasets, the test results for the *all distortions category* are given in Tables 3.8 and 3.9 in terms of PLCC and SRCC, respectively. The test results for the *subset distortions category* are given in Tables 3.10 and 3.11 in terms of PLCC and SRCC, respectively. In each of these tables, the weighted average (WA) PLCC/SRCC values are provided in the rightmost column and the methods have been sorted in descending order with respect to these values. The results of statistical significance testing of EONSS relative to the two FR and 14 BIQA methods for both the *all distortions* and *subset distortions* categories are provided in Table 3.12, where a “1”, “–”, or “0” means that the perceptual quality prediction performance of EONSS is better, indistinguishable, or worse, respectively, than that of the method in the row for a given database (with 95% confidence). Each entry in the table may be composed of more than one symbol, each of which represents the outcome of the test for either the *all distortions* and *subset distortions* categories, as explained in the table caption. We preceded the statistical significance testing with a kurtosis based check

Table 3.8: PLCC of EONSS in comparison with 2 FR and 14 NR methods on nine subject-rated IQA databases. All distortions in each test dataset were considered. The Weighted Average PLCC (WA PLCC) is provided in the rightmost column and methods are sorted in descending order with respect to it. FR methods are highlighted in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MD IVL	WA PLCC
IWSSIM [13]	0.9522	0.8319	0.9144	0.9191	0.8476	0.8698	0.8983	0.8513	0.9109	0.9056	0.8787
PSNR	0.8723	0.7017	0.8000	0.8321	0.6302	0.6808	0.6164	0.5647	0.7398	0.6806	0.7065
EONSS	0.9244	0.5442	0.7660	0.9120	0.5798	0.4821	0.8374	0.3020	0.8437	0.8744	0.6933
CORNIA [141]	0.9665	0.5729	0.7593	0.8366	0.4496	0.3530	0.8074	0.6935	0.8679	0.8277	0.6878
ILNIQE [144]	0.9022	0.5883	0.8538	0.7289	0.3124	0.3390	0.7245	0.5146	0.8923	0.6303	0.6452
HOSA [143]	0.9991	0.5521	0.7560	0.8496	0.4969	0.3761	0.6590	0.2513	0.6768	0.7167	0.6328
dipIQ [36]	0.9348	0.4774	0.7720	0.8942	0.5223	0.3889	0.6789	0.4376	0.7669	0.7627	0.6284
NRSL [147]	0.9815	0.5338	0.7456	0.8905	0.4672	0.4034	0.6566	0.3088	0.5183	0.6794	0.6182
SISBLIM [32]	0.8077	0.4805	0.7378	0.7574	0.4909	0.4671	0.6321	0.8135	0.8948	0.5723	0.6077
GWHGLBP [142]	0.8079	0.4982	0.7104	0.6427	0.3653	0.2978	0.7108	0.7443	0.9655	0.5966	0.5991
BIQI [139]	0.9224	0.4678	0.6916	0.6106	0.3596	0.2661	0.6763	0.3369	0.7389	0.6215	0.5648
NIQE [3]	0.9052	0.4001	0.7188	0.8040	0.3703	0.2708	0.6728	0.5634	0.8387	0.5688	0.5646
MEON [146]	0.9389	0.4919	0.7865	0.9221	0.4774	0.3854	0.5250	0.2430	0.2684	0.5722	0.5630
WaDIQaM-NR [148]	0.9341	0.5712	0.6882	0.7862	0.4133	0.3481	0.4631	0.1371	0.2685	0.5214	0.5457
BRISQUE [140]	0.9671	0.4747	0.7006	0.8208	0.4155	0.3257	0.4450	0.1403	0.6045	0.6517	0.5429
QAC [35]	0.8625	0.4371	0.7067	0.7615	0.3573	0.2856	0.6043	0.4240	0.4145	0.5713	0.5338
LPSI [145]	0.8280	0.4892	0.7216	0.6020	0.4037	0.3981	0.4335	0.1765	0.5464	0.5715	0.5204

for Gaussianity of prediction residuals of all methods under test on all datasets (described in Section 3.4.3) and found that the assumption of Gaussianity holds in around 89% cases in the *all distortions category* and in around 83% cases in the *subset distortions category*, thereby allowing us to use the F-test.

From the above-mentioned tables, the following observations can be made: 1) Tables 3.8 and 3.9 reveal that EONSS outperforms all other state-of-the-art BIQA methods in the *all distortions category*, both in terms of WA PLCC and WA SRCC. 2) Similarly, in the *subset distortions category*, which can be considered a more *fair* ground for comparison as stated earlier, Tables 3.10 and 3.11 show that EONSS considerably outperforms all other BIQA methods both in terms of WA PLCC and WA SRCC. 3) Table 3.12 shows that for the 160 method-database combinations of the *all distortions category*, EONSS performs statistically better than other methods in around 62% cases, while its perfor-

Table 3.9: SRCC of EONSS in comparison with 2 FR and 14 NR methods on nine subject-rated IQA databases. All distortions in each test dataset were considered. The Weighted Average SRCC (WA SRCC) is provided in the rightmost column and methods are sorted in descending order with respect to it. FR methods are highlighted in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MD IVL	WA SRCC
IWSSIM [13]	0.9567	0.7779	0.9212	0.9163	0.8484	0.8564	0.8911	0.8551	0.8836	0.8588	0.8559
PSNR	0.8756	0.6394	0.8057	0.8246	0.6254	0.6701	0.5784	0.5604	0.6771	0.6136	0.6720
EONSS	0.9267	0.5045	0.6774	0.9063	0.4991	0.3448	0.8297	0.2874	0.7260	0.8833	0.6509
CORNIA [141]	0.9681	0.4288	0.6534	0.8354	0.3727	0.2071	0.7918	0.7055	0.8340	0.8336	0.6147
ILNIQE [144]	0.8975	0.4939	0.8144	0.7391	0.2997	0.3127	0.6900	0.5148	0.8778	0.6238	0.6031
HOSA [143]	0.9990	0.4705	0.5925	0.8574	0.4494	0.3248	0.6412	0.2993	0.6393	0.7399	0.5851
dipIQ [36]	0.9378	0.4377	0.5266	0.8957	0.4135	0.2100	0.6612	0.4153	0.6678	0.7131	0.5620
NRSL [147]	0.9796	0.4277	0.6750	0.8930	0.4249	0.2894	0.6458	0.4088	0.4145	0.6047	0.5589
SISBLIM [32]	0.7741	0.3177	0.6603	0.7622	0.4435	0.4098	0.6554	0.8089	0.8770	0.5375	0.5408
GWHGLBP [142]	0.7410	0.3844	0.5773	0.6243	0.3337	0.2412	0.7032	0.7555	0.9698	0.5841	0.5377
NIQE [3]	0.9073	0.3132	0.6271	0.8126	0.3458	0.2212	0.6523	0.5451	0.7738	0.5713	0.5181
BIQI [139]	0.9198	0.3935	0.6186	0.6170	0.3433	0.2353	0.6276	0.0077	0.5556	0.5711	0.5007
MEON [146]	0.9409	0.3750	0.7248	0.9215	0.4101	0.2497	0.4861	0.2980	0.1917	0.5466	0.4969
BRISQUE [140]	0.9654	0.3672	0.5563	0.8130	0.3640	0.2496	0.4035	0.2209	0.5018	0.6647	0.4792
WaDIQaM-NR [148]	0.9417	0.4393	0.6388	0.7524	0.3588	0.2235	0.4040	0.1316	0.2379	0.5614	0.4782
QAC [35]	0.8683	0.3722	0.4900	0.7686	0.3196	0.1944	0.3239	0.2272	0.3579	0.5524	0.4292
LPSI [145]	0.8181	0.3949	0.5303	0.5865	0.2060	0.1411	0.0306	0.0168	0.2717	0.5736	0.3558

mance is statistically indistinguishable or inferior than other methods in around 19% and 19% cases, respectively. Similarly, for the 192 method-database combinations of the *subset distortions category*, EONSS performs statistically better than other methods in around 67% cases, while its performance is statistically indistinguishable or inferior than other methods in around 13% and 20% cases, respectively. This again demonstrates the superiority of EONSS when compared to the very state-of-the-art in the BIQA field. 4) While considering Tables 3.8, 3.9, 3.10, 3.11, and 3.12, it should be noted that the OA BIQA methods BIQI [139], BRISQUE [140], NRSL [147], CORNIA [141], HOSA [143], WaDIQaM-NR [148], and MEON [146] are trained on the LIVE R2 database [24], and GWHGLBP [142] is trained on the LIVE MD database [31]. Thus, comparing these OA BIQA methods with other approaches on these respective databases is unreliable and unfair to those other methods. Disregarding the results of these methods on the said datasets

Table 3.10: PLCC of EONSS in comparison with 2 FR and 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each test dataset were considered. The Weighted Average PLCC (WA PLCC) is provided in the rightmost column and methods are sorted in descending order with respect to it. FR methods are highlighted in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG	WA PLCC
IWSSIM [13]	0.9556	0.9407	0.9655	0.9191	0.8745	0.8536	0.8983	0.8513	0.9164	0.9117	0.9269	0.9101	0.9116
EONSS	0.9462	0.8751	0.9291	0.9120	0.7973	0.8082	0.8374	0.3020	0.8622	0.8337	0.9232	0.8918	0.8430
CORNIA [141]	0.9715	0.8868	0.9257	0.8366	0.5898	0.5480	0.8074	0.6935	0.8774	0.8723	0.9419	0.7900	0.8145
dipIQ [36]	0.9559	0.8879	0.9481	0.8942	0.7475	0.6706	0.6789	0.4376	0.8235	0.7895	0.8311	0.7882	0.7839
HOSA [143]	0.9992	0.8901	0.9384	0.8496	0.6774	0.6597	0.6590	0.2513	0.8968	0.6728	0.9005	0.7022	0.7600
ILNIQE [144]	0.9164	0.8576	0.9070	0.7289	0.3860	0.4598	0.7245	0.5146	0.9048	0.8968	0.8293	0.5759	0.7263
PSNR	0.8699	0.8912	0.9079	0.8321	0.6532	0.5560	0.6164	0.5647	0.7409	0.7751	0.7143	0.6645	0.7241
NRSL [147]	0.9887	0.9153	0.9133	0.8905	0.6236	0.6145	0.6566	0.3088	0.3516	0.6263	0.6418	0.7334	0.7239
SISBLIM [32]	0.8220	0.7896	0.7967	0.7574	0.5899	0.6844	0.6321	0.8135	0.9030	0.8913	0.8056	0.4871	0.7194
NIQE [3]	0.9162	0.8091	0.8767	0.8040	0.4994	0.4712	0.6728	0.5634	0.9099	0.8481	0.7996	0.4507	0.7135
GWHGLBP [142]	0.8088	0.7675	0.8052	0.6427	0.5196	0.5347	0.7108	0.7443	0.9677	0.9684	0.7745	0.4943	0.7113
BIQI [139]	0.9534	0.7772	0.8224	0.6106	0.4957	0.5164	0.6763	0.3369	0.7743	0.7404	0.7398	0.6035	0.6827
MEON [146]	0.9907	0.9053	0.9423	0.9221	0.6620	0.6510	0.5250	0.2430	0.2675	0.4927	0.3875	0.7405	0.6763
QAC [35]	0.8777	0.8051	0.8736	0.7615	0.4512	0.5068	0.6043	0.4240	0.5378	0.6722	0.6765	0.6090	0.6637
BRISQUE [140]	0.9760	0.8659	0.9239	0.8208	0.5257	0.5421	0.4450	0.1403	0.8663	0.4594	0.8249	0.6511	0.6564
WaDIQaM-NR [148]	0.9302	0.8994	0.8860	0.7862	0.5137	0.5530	0.4631	0.1371	0.6842	0.3921	0.6415	0.5231	0.6251
LPSI [145]	0.8440	0.8114	0.8657	0.6020	0.5508	0.6289	0.4335	0.1765	0.8820	0.1182	0.7959	0.5075	0.5991

further increases the demonstrated superiority of EONSS. It is also pertinent to mention that the nine subject-rated IQA datasets have only been used to test EONSS, without any retraining or fine-tuning. 5) Even though EONSS has been trained on the Waterloo Exploration-II dataset, which predominantly consists of multiply distorted images, it performs well on even the singly distorted test datasets. This is explained by the wide density of distorted images in the Waterloo Exploration-II dataset, which includes 117,810 singly distorted images and a large number of multiply distorted images that have a small amount of stage-1 distortion, thereby allowing the DNN model to learn effectively for the single distortion scenario. 6) It can be clearly seen that EONSS comprehensively outperforms the two other DNN based models, MEON [146] and WaDIQaM-NR [148], statistically and in terms of WA PLCC, WA SRCC, for both the *all* and *subset distortions categories*. Since, MEON [146] and WaDIQaM-NR [148], are trained on a small-scale singly distorted dataset (LIVE R2 [24]), they do not perform well on multiply distorted datasets, which is not the

Table 3.11: SRCC of EONSS in comparison with 2 FR and 14 NR methods on nine subject-rated IQA databases. A subset of distortions in each test dataset were considered. The Weighted Average SRCC (WA SRCC) is provided in the rightmost column and methods are sorted in descending order with respect to it. FR methods are highlighted in bold.

NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG	WA SRCC
IWSSIM [13]	0.9616	0.9262	0.9603	0.9163	0.8755	0.8374	0.8911	0.8551	0.8700	0.8933	0.8778	0.8713	0.9002
EONSS	0.9499	0.8446	0.8969	0.9063	0.7885	0.7553	0.8297	0.2874	0.7348	0.7331	0.8754	0.9085	0.8205
CORNIA [141]	0.9732	0.8727	0.8987	0.8354	0.5740	0.5053	0.7918	0.7055	0.8278	0.8523	0.9254	0.8027	0.8007
dipIQ [36]	0.9574	0.8720	0.9290	0.8957	0.7460	0.6433	0.6612	0.4153	0.6979	0.7391	0.6512	0.7730	0.7562
HOSA [143]	0.9991	0.8681	0.9111	0.8574	0.6677	0.6236	0.6412	0.2993	0.8437	0.5357	0.8789	0.7150	0.7438
ILNIQE [144]	0.9153	0.8417	0.8802	0.7391	0.3669	0.4248	0.6900	0.5148	0.8915	0.8821	0.7915	0.5797	0.7078
PSNR	0.8731	0.9073	0.9218	0.8246	0.6553	0.5763	0.5784	0.5604	0.6621	0.7088	0.6572	0.5841	0.7048
SISBLIM [32]	0.7835	0.7703	0.8059	0.7622	0.5565	0.6314	0.6554	0.8089	0.8746	0.8782	0.7584	0.3320	0.7008
NRSL [147]	0.9880	0.8965	0.8874	0.8930	0.5732	0.5564	0.6458	0.4088	0.2634	0.5991	0.4684	0.7125	0.6996
NIQE [3]	0.9168	0.7972	0.8710	0.8126	0.4703	0.4180	0.6523	0.5451	0.8713	0.7938	0.7625	0.4510	0.6954
GWHGLBP [142]	0.7447	0.6538	0.6728	0.6243	0.4768	0.4454	0.7032	0.7555	0.9640	0.9751	0.7584	0.4502	0.6672
MEON [146]	0.9906	0.9012	0.9300	0.9215	0.6421	0.5830	0.4861	0.2980	0.0476	0.3257	0.3255	0.7397	0.6441
BIQI [139]	0.9528	0.7763	0.7972	0.6170	0.4976	0.4849	0.6276	0.0077	0.6542	0.4902	0.6591	0.5302	0.6272
BRISQUE [140]	0.9757	0.8401	0.8992	0.8130	0.4727	0.4771	0.4035	0.2209	0.7923	0.2991	0.7385	0.6612	0.6239
WaDIQaM-NR [148]	0.9399	0.8646	0.8636	0.7524	0.4777	0.4691	0.4040	0.1316	0.5012	0.2502	0.6121	0.4830	0.5786
QAC [35]	0.8857	0.8055	0.8415	0.7686	0.4450	0.4566	0.3239	0.2272	0.3959	0.4707	0.5537	0.5282	0.5529
LPSI [145]	0.8333	0.7046	0.7711	0.5865	0.3382	0.3949	0.0306	0.0168	0.8387	0.0012	0.7348	0.4692	0.4254

case with EONSS. By using MEON [146] as an example, we show in the next sub-section, that the performance of pre-existing DNN based BIQA models can indeed be elevated by retraining on the Waterloo Exploration-II database. 7) While the performance of EONSS is a considerable distance away from the state-of-the-art FR method IWSSIM [13] in the *all distortions category* (Tables 3.8 and 3.9), its performance is relatively closer to IWSSIM in the *subset distortions category* (Tables 3.10 and 3.11). Since the Waterloo Exploration-II database does not have the wide-ranging distortions of the *all distortions category*, this shows that it is possible for a DNN based BIQA method to approach FR performance for distortion types for which sufficient annotated training data is available. This is no small achievement for a BIQA method, given that it has no access to the reference image.

We evaluated the computational complexity of all IQA methods under test in terms of their execution time to determine the quality of a 1024 × 1024 test color image on a desktop computer with a 3.5 GHz Intel Core i7-7800X processor, 16 GB of RAM, NVIDIA

Table 3.12: Statistical significance testing of EONSS through the F-Test with respect to 2 FR and 14 NR methods on different IQA databases, for *All* and subset (*SS*) distortions. The order of symbols within each entry is as follows: LIVE R2 (All, SS), TID2013 (All, SS), CSIQ (All, SS), VCLFER (All), CIDIQ50 (All, SS), CIDIQ100 (All, SS), MDID (All), MDID2013 (All), LIVE MD (All, Blur-JPEG, Blur-Noise), MDIVL (All, Blur-JPEG, Noise-JPEG). A "1" means that EONSS performance is statistically better than the method in the row, a "0" means that it is statistically worse, while a "–" means that it is statistically indistinguishable. Testing was done at the 5% significance level (95% confidence). Methods are listed in alphabetical order and FR methods are in bold.

FR/NR Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL
BIQI [139]	–0	11	11	1	11	11	1	–	111	111
BRISQUE [140]	00	1–	1–	1	11	11	1	–	1–1	111
CORNIA [141]	00	– –	– –	1	11	11	1	0	0–0	101
dipIQ [36]	00	1–	–0	1	–1	–1	1	–	11–	111
GWHGLBP [142]	11	11	11	1	11	11	1	0	000	111
HOSA [143]	00	– –	–0	1	11	–1	1	–	101	111
ILNIQE [144]	11	0–	01	1	11	11	1	0	000	111
IWSSIM [13]	00	00	00	–	00	00	0	0	000	0–0
LPSI [145]	11	11	11	1	11	–1	1	–	1–1	111
MEON [146]	00	10	–0	–	11	–1	1	1	111	111
NIQE [3]	11	11	11	1	11	11	1	0	–0–	111
NRSL [147]	00	–0	–1	1	11	–1	1	–	111	111
PSNR	11	0–	01	1	–1	01	1	0	111	111
QAC [35]	11	11	11	1	11	11	1	–	111	111
SISBLIM [32]	11	11	–1	1	11	–1	1	0	000	111
WaDIQaM-NR [148]	01	–0	11	1	11	11	1	–	111	111

GeForce GTX 1050Ti GPU, and Ubuntu 18.04 operating system. The execution times of all methods are given in Table 3.13, where methods have been sorted in ascending order with respect to execution time. Since the FR PSNR is the fastest method, we also provide the execution time relative to PSNR for ease in comparison. The time for DNN based methods, EONSS, MEON [146], and WaDIQaM-NR [148], was evaluated both on the GPU and CPU, while that of all other methods was evaluated on the CPU only. It should be noted that the execution time of some other well-known BIQA methods including

Table 3.13: Execution Time of FR and NR methods on a test image. Methods are sorted in ascending order with respect to the execution time. FR methods are highlighted in bold.

FR/NR Method	Processing Unit	Execution Time (Seconds)	Execution Time Relative to PSNR
PSNR	CPU	0.0013	1.00
LPSI [145]	CPU	0.0397	30.54
EONSS	GPU	0.0604	46.46
EONSS	CPU	0.0817	62.85
MEON [146]	CPU	0.0819	63.00
MEON [146]	GPU	0.0876	67.38
HOSA [143]	CPU	0.1309	100.69
QAC [35]	CPU	0.1357	104.38
NRSL[1] [147]	CPU	0.1421	109.31
GWHGLBP[1] [142]	CPU	0.1469	113.00
WaDIQaM-NR [148]	GPU	0.1549	119.15
BRISQUE [140]	CPU	0.1823	140.23
NIQE [3]	CPU	0.2941	226.23
BIQI [139]	CPU	0.4634	356.46
IWSSIM [13]	CPU	0.6067	466.69
dipIQ [36]	CPU	1.6592	1276.31
CORNIA [141]	CPU	2.0304	1561.85
SISBLIM [32]	CPU	2.2005	1692.69
ILNIQE [144]	CPU	2.5227	1940.54
WaDIQaM-NR [148]	CPU	6.2818	4832.15

[1] Feature extraction time only.

BLIINDS2 [176], DIIVINE [177], FRIQUEE [179], MS-LQAF [181], NFERM [182], and TCLT [183], is even more than that of ILNIQE [144], making them infeasible for large-scale or real-time use, which is why we have not included them in our analysis. It can be seen from Table 3.13 that the execution time of EONSS is approximately 20 to 30 times faster than competitive BIQA methods, such as CORNIA [141], dipIQ [36], ILNIQE [144], and SISBLIM [32]. Thus when Tables 3.8, 3.9, 3.10, 3.11, and 3.12 are considered in conjunction with Table 3.13, it becomes clear that EONSS not only outperforms the very best methods in the BIQA field in terms of perceptual quality prediction accuracy on unseen test data, but that it is also the fastest among them by a wide margin, making it an excellent choice for practical applications.

Table 3.14: PLCC and SRCC values for EONSS, EON_L, MEONSS, and MEON when tested on nine subject-rated IQA databases. All distortions in each test dataset were considered. The Weighted Average PLCC/SRCC are provided in the rightmost column and methods are sorted in descending order with respect to them.

Evaluation Criteria	Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL	Weighted Average
PLCC	EONSS	0.9244	0.5442	0.7660	0.9120	0.5798	0.4821	0.8374	0.3020	0.8437	0.8744	0.6933
	MEONSS	0.8975	0.4270	0.7359	0.9079	0.4459	0.3218	0.6916	0.2855	0.7885	0.9012	0.6059
	MEON	0.9389	0.4919	0.7865	0.9221	0.4774	0.3854	0.5250	0.2430	0.2684	0.5722	0.5630
	EON_L	0.8586	0.3772	0.6214	0.7557	0.2758	0.1942	0.6293	0.1865	0.4963	0.4238	0.4833
SRCC	EONSS	0.9267	0.5045	0.6774	0.9063	0.4991	0.3448	0.8297	0.2874	0.7260	0.8833	0.6509
	MEONSS	0.9060	0.3796	0.6547	0.9087	0.3768	0.1748	0.6985	0.2709	0.6211	0.8918	0.5615
	MEON	0.9409	0.3750	0.7248	0.9215	0.4101	0.2497	0.4861	0.2980	0.1917	0.5466	0.4969
	EON_L	0.8636	0.2464	0.5284	0.7456	0.2245	0.1395	0.5453	0.1343	0.3670	0.3643	0.4006

Waterloo Exploration-II versus a Contemporary IQA Dataset: Impact on DNN performance

The superior performance of EONSS, as demonstrated in the previous sub-section, can be directly attributed to the large-scale Waterloo Exploration-II database. We demonstrate this point more explicitly in this section by comparing the following four models: 1) EONSS that has been trained on the Waterloo Exploration-II database, 2) We retrain the DNN architecture employed by EONSS (as described in Section 3.5.1) on the small-scale subject-rated LIVE R2 database [24] and call this model EON_L. 3) As another comparison point, we consider MEON [146] trained on LIVE R2 database [24]. 4) We retrain the MEON DNN on the Waterloo Exploration-II database and call it MEONSS. Tables 3.14 and 3.15 show the results for the *all* and *subset distortions categories*, respectively, both in terms of WA PLCC and WA SRCC, where we have sorted methods in the descending order with respect to their WA PLCC/SRCC values.

It is clear from Tables 3.14 and 3.15 that EONSS massively outperforms EON_L in terms of WA PLCC and WA SRCC, both in the *all* and *subset distortions categories*. The only difference between EONSS and EON_L is the training data used (both use exactly the same DNN). Thus, the enormous superiority of EONSS when compared to EON_L can only

Table 3.15: PLCC and SRCC values for EONSS, EON_L, MEONSS, and MEON when tested on nine subject-rated IQA databases. A subset of distortions in each test dataset were considered. The Weighted Average PLCC/SRCC are provided in the rightmost column and methods are sorted in descending order with respect to them.

Evaluation Criteria	Method	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG	Weighted Average
PLCC	EONSS	0.9462	0.8751	0.9291	0.9120	0.7973	0.8082	0.8374	0.3020	0.8622	0.8337	0.9232	0.8918	0.8430
	MEONSS	0.9213	0.8255	0.9125	0.9079	0.6307	0.5836	0.6916	0.2855	0.8019	0.7911	0.9104	0.9333	0.7668
	MEON	0.9907	0.9053	0.9423	0.9221	0.6620	0.6510	0.5250	0.2430	0.2675	0.4927	0.3875	0.7405	0.6763
	EON_L	0.8769	0.8006	0.7562	0.7557	0.3551	0.3854	0.6293	0.1865	0.5556	0.5342	0.5663	0.3602	0.6039
SRCC	EONSS	0.9499	0.8446	0.8969	0.9063	0.7885	0.7553	0.8297	0.2874	0.7348	0.7331	0.8754	0.9085	0.8205
	MEONSS	0.9280	0.8151	0.9095	0.9087	0.6199	0.5428	0.6985	0.2709	0.5756	0.6722	0.8437	0.9365	0.7479
	MEON	0.9906	0.9012	0.9300	0.9215	0.6421	0.5830	0.4861	0.2980	0.0476	0.3257	0.3255	0.7397	0.6441
	EON_L	0.8882	0.7822	0.7706	0.7456	0.3274	0.3495	0.5453	0.1343	0.4309	0.3074	0.3928	0.2620	0.5472

be attributed to the large-scale synthetically-annotated training data that it utilizes, that is, the Waterloo Exploration-II database, which allows the DNN to learn a robust quality model. From Tables 3.14 and 3.15 it is also clear that MEONSS outperforms MEON in terms of WA PLCC and WA SRCC, both in the *all* and *subset distortions categories*. Again, the only difference between MEONSS and MEON is the training data (both use exactly the same DNN). This again demonstrates the superiority of using the Waterloo Exploration-II database for BIQA model training.

From Tables 3.14 and 3.15 we can also make the following three observations: 1) It is evident that the margin with which MEONSS outperforms MEON is smaller than the one with which EONSS outperforms EON_L. 2) It is clear that although both EONSS and MEONSS are trained on the very large-scale Waterloo Exploration-II database, EONSS significantly outperforms MEONSS in terms of WA PLCC and WA SRCC, both in the *all* and *subset distortions categories*. 3) However, it can also be seen that although both EON_L and MEON are trained on the small-scale LIVE R2 database [24], MEON significantly outperforms EON_L in terms of WA PLCC and WA SRCC, both in the *all* and *subset distortions categories*. This is a significant finding as it shows that the choice of DNN network architecture for the BIQA task is strongly impacted by the amount of available quality annotated training data. As we have discussed before, MEON [146] takes a multi-task approach and utilizes two sub-networks, where sub-network 1 performs the task of distortion type identification for which a large amount of non-quality annotated training data is made available, and sub-network 2 performs quality prediction using the results from sub-network 1. On the other hand, EONSS takes a single task approach of quality prediction and hence its network is simpler compared to MEON. Our results show that the multi-task DNN model (MEON) performs better when only a small amount of quality annotated training data is available (MEON outperforms EON_L), while the single-task DNN model (EONSS) performs much better when a very large amount of quality annotated training data is present (EONSS outperforms MEONSS). This shows that even a simple single-task network architecture is able to learn an effective quality model in a truly end-to-end manner given the availability of a large amount of quality annotated training data, thereby establishing the strength of the large-scale Waterloo Exploration-II database.

While considering model performance on individual datasets in Tables 3.14 and 3.15, it can be seen that MEON performs better than EONSS and MEONSS on the singly distorted subject-rated databases LIVE R2 [24], TID2013 [19], CSIQ [26], and VCLFER [54], especially in the *subset distortions category*. Since MEON is trained on LIVE R2, it is unfair to compare other models with MEON on this dataset. It is pertinent to mention that the distortion type distributions, for the *subset distortions category*, of the TID2013, CSIQ, and VCLFER databases are similar to that of LIVE R2. Thus, MEON also performs well on these datasets. However, Waterloo Exploration-II, which is a predominantly multiply distorted dataset and is used to train EONSS and MEONSS, has very different distortion type distributions compared to these singly distorted subject-rated datasets. It should also be noted that the reference content of LIVE R2 and TID2013 databases has a partial overlap. Thus, it is not completely fair to compare the performance of EONSS and MEONSS with MEON on LIVE R2, TID2013, CSIQ, and VCLFER, as these datasets are biased in favor of MEON. However, even then the performance of EONSS and MEONSS is satisfactory and in most cases not that far behind that of MEON on these singly distorted subject-rated datasets. Further, EONSS performs better than MEON on the TID2013 database in the difficult *all distortions category*, while it significantly outperforms MEON on the singly distorted CIDIQ database [5] and the multiply distorted datasets MDID [33], LIVE MD [31], and MDIVL [34]. This shows that the alignment of the content and distortion type distributions becomes a crucial factor when training with small-scale datasets. On the other hand, using a very large-scale dataset for training mitigates the impact of such distribution misalignment between the training and testing data, thereby leading to more robust models.

Impact of Training Dataset Size on EONSS Performance

While it is difficult to determine how large the training dataset size should be to learn effective DNN based BIQA models, we try to answer this question empirically. Specifically, we consider four subsets of the Waterloo Exploration-II database which contain 1%, 5%, 10%, and 20% reference images of the original dataset along with their respective distorted versions. Next, we retrain EONSS on these dataset subsets and call the trained versions EONSS_1, EONSS_5, EONSS_10, and EONSS_20, respectively. While training, we further split each subset into training, validation, and testing sets which are composed of 60%, 20%, and 20% of subset images, respectively. Tables 3.16 and 3.17 show the results for the *all* and *subset distortions categories*, respectively, both in terms of WA PLCC and WA SRCC. We have repeated the results for EONSS in these tables, which utilizes 100% of the Waterloo Exploration-II database for its training, validation, and testing. It can be observed from these tables that the model performance increases dramatically from EONSS_1 to EONSS_5 for both the *all* and *subset distortions categories*. Substantial performance increase is further seen from EONSS_5 to EONSS_20 for the *subset distortions category*, which we believe is a more accurate category to consider for these experiments given that training and testing distortion types are more closely aligned. For both the *all* and *subset distortions categories* further performance gain can be seen from EONSS_20 to EONSS (that uses the entire dataset for training, validation, and testing), however it is not by a wide margin. While definitive conclusions are hard to make, it can be said that using more than 20% of the Waterloo Exploration-II database may be a bit redundant. It should be noted that 20% of the dataset still includes a substantial amount of annotated data (691,152 distorted images). However, for future dataset releases, this indicates that instead of having such a large amount of images per distortion type, images belonging to a more diverse set of distortion types should be considered.

Table 3.16: PLCC and SRCC values for various versions of EONSS trained on different subsets of the Waterloo Exploration-II database and tested on nine subject-rated IQA databases. All distortions in each test dataset were considered. The Weighted Average PLCC/SRCC are provided in the rightmost column and methods are sorted in descending order with respect to them.

Evaluation Criteria	EONSS Version	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD	MDIVL	Weighted Average
PLCC	EONSS	0.9244	0.5442	0.7660	0.9120	0.5798	0.4821	0.8374	0.3020	0.8437	0.8744	0.6933
	EONSS_20	0.9231	0.5637	0.7678	0.9050	0.5800	0.4629	0.8014	0.2346	0.8330	0.8749	0.6890
	EONSS_10	0.9146	0.5788	0.8112	0.8809	0.4723	0.4081	0.7701	0.5231	0.8089	0.8538	0.6856
	EONSS_5	0.8972	0.5807	0.7775	0.8839	0.4359	0.3693	0.7634	0.2867	0.7922	0.8699	0.6681
	EONSS_1	0.7637	0.5026	0.7438	0.6292	0.2608	0.2608	0.5662	0.0570	0.6038	0.6988	0.5334
SRCC	EONSS	0.9267	0.5045	0.6774	0.9063	0.4991	0.3448	0.8297	0.2874	0.7260	0.8833	0.6509
	EONSS_20	0.9214	0.5165	0.6651	0.9053	0.5097	0.3530	0.7919	0.2163	0.7278	0.8736	0.6451
	EONSS_10	0.9107	0.5232	0.7260	0.8879	0.4395	0.3314	0.7615	0.4050	0.6923	0.8495	0.6420
	EONSS_5	0.8895	0.5407	0.6794	0.8899	0.4094	0.3057	0.7544	0.2622	0.6907	0.8724	0.6335
	EONSS_1	0.7373	0.4377	0.6600	0.6432	0.2426	0.2164	0.5163	0.0268	0.5289	0.6899	0.4866

Table 3.17: PLCC and SRCC values for various versions of EONSS trained on different subsets of the Waterloo Exploration-II database and tested on nine subject-rated IQA databases. A subset of distortions in each test dataset were considered. The Weighted Average PLCC/SRCC are provided in the rightmost column and methods are sorted in descending order with respect to them.

Evaluation Criteria	EONSS Version	LIVE R2	TID2013	CSIQ	VCLFER	CIDIQ50	CIDIQ100	MDID	MDID2013	LIVE MD BJPG	LIVE MD BN	MDIVL BJPG	MDIVL NJPG	Weighted Average
PLCC	EONSS	0.9462	0.8751	0.9291	0.9120	0.7973	0.8082	0.8374	0.3020	0.8622	0.8337	0.9232	0.8918	0.8430
	EONSS_20	0.9403	0.8705	0.9244	0.9050	0.7899	0.7925	0.8014	0.2346	0.8753	0.7971	0.9116	0.8882	0.8250
	EONSS_10	0.9385	0.8687	0.9264	0.8809	0.6121	0.6421	0.7701	0.5231	0.8301	0.7963	0.8685	0.8719	0.8007
	EONSS_5	0.9117	0.8515	0.9146	0.8839	0.5641	0.5904	0.7634	0.2867	0.8297	0.7717	0.9100	0.8719	0.7762
	EONSS_1	0.8038	0.7663	0.8037	0.6292	0.2450	0.3184	0.5662	0.0570	0.6429	0.5981	0.7439	0.6957	0.5883
SRCC	EONSS	0.9499	0.8446	0.8969	0.9063	0.7885	0.7553	0.8297	0.2874	0.7348	0.7331	0.8754	0.9085	0.8205
	EONSS_20	0.9402	0.8411	0.8875	0.9053	0.7749	0.7331	0.7919	0.2163	0.7839	0.6893	0.8655	0.8907	0.8010
	EONSS_10	0.9385	0.8258	0.8871	0.8879	0.6112	0.6115	0.7615	0.4050	0.7020	0.6949	0.8257	0.8800	0.7737
	EONSS_5	0.9070	0.8049	0.8635	0.8899	0.5664	0.5569	0.7544	0.2622	0.7236	0.6657	0.8711	0.8803	0.7528
	EONSS_1	0.7803	0.6809	0.7306	0.6432	0.2677	0.2811	0.5163	0.0268	0.5545	0.5157	0.6809	0.7137	0.5498

3.6 Practical Application

The practical applications of the work done in this chapter are as follows:

- Since EONSS has outperformed the very state-of-the-art in BIQA, both in terms of perceptual quality prediction and speed, it can be used in real-world scenarios that require BIQA, as long as the expected distortion types have an overlap with what EONSS has been trained for, i.e., the distortions found in the Waterloo Exploration-II database.
- As mentioned in Section 3.5, EONSS has a relatively simple architecture since our focus was to investigate the impact of data on DNN performance in BIQA. Since we are publicly releasing the Waterloo Exploration-II database, more sophisticated DNN based IQA models (FR, RR, or NR) can be developed and trained on this dataset with the aim to perform better than the performance baseline established by EONSS.
- The alternative quality annotation mechanism developed in this chapter, SQB, can be used to annotate any number of new IQA datasets which can be even larger than the Waterloo Exploration-II database, thereby leading to even more diverse datasets for training machine learning based IQA models.

3.7 Summary

Although DNN based models have led to tremendous progress in the area of visual recognition, such breakthroughs have not been witnessed thus far in the area of DNN based BIQA models, mainly due to the lack of large-scale annotated training data in the IQA field. Researchers have tried to address this issue by relying on data augmentation and primarily focusing on the design of DNN architectures and training methods, but have achieved only limited success. Perhaps the biggest contribution of the current work is to show that the quality and quantity of the training data plays an even more important role in the success of DNN approaches. In this chapter we have developed the largest IQA dataset to-date, called the Waterloo Exploration-II database, which has 3,570 pristine reference and around 3.45 million, singly and multiply, distorted images. Since it is not possible to quality-annotate such a large number of images through subjective experiments, we have developed a novel alternative mechanism, based on reciprocal rank fusion, to synthetically assign quality labels to the images of this dataset. Extensive tests on subject-rated datasets, reveal that these synthetic quality benchmark labels are highly accurate in perceptual quality prediction and perform better than the very best of FR IQA methods. To demonstrate the validity of our approach, we have developed a new DNN based BIQA model called EONSS, which is trained on the Waterloo Exploration-II database and tested on nine subject-rated IQA datasets without any retraining or fine-tuning.

We have comprehensively demonstrated in Section 3.5.2 that EONSS not only outperforms other methods, that are regarded as the very state-of-the-art in BIQA, in terms of perceptual quality prediction performance, but is also the fastest among them by a wide margin. These characteristics make EONSS the very best in the field of blind image quality assessment as it exists today. As discussed in Section 3.5.1, EONSS has a relatively simple network architecture, when compared to other DNN based BIQA methods, such as MEON [146], WaDIQaM-NR [148], and other methods discussed in Section 3.2.2. Therefore, the success of EONSS can be attributed to the synthetically labeled Waterloo Exploration-II database, whose enormity and content-diversity has provided sufficient data to the DNN to learn a robust BIQA model in a truly end-to-end manner. The overall high performance of EONSS, especially on the *subset distortion category*, shows that it does not suffer from overfitting issues compared to other BIQA methods. Since we have trained EONSS on the synthetically annotated Waterloo Exploration-II database, and have used the nine subject-rated datasets only for testing, this also validates the effectiveness of our synthetic annotation approach for labeling very large-scale IQA datasets.

Chapter 4

Degraded Reference Image Quality Assessment

In practical media distribution systems, visual content usually undergoes multiple stages of quality degradations along the delivery chain between the source and destination. In addition to the final version, a number of earlier degraded versions of such content are available as it passes through the distribution system, however, the pristine original version is seldom available. The inaccessibility to the pristine quality version of visual content renders the full-reference (FR) and reduced-reference (RR) image quality assessment (IQA) methods infeasible for practical application. While no-reference (NR) or BIQA methods are readily applicable at the final destination, these methods have not yet reached a robust level of performance. While the availability of additional degraded versions of visual content may be beneficial to the task of IQA of the final distorted images, none of the major IQA paradigms (FR, RR, NR) have the capability to utilize this additional information. Thus, practically applicable IQA models are still lacking. In this chapter, we analyze the performance of contemporary FR and NR methods in evaluating the quality of multiply distorted content. Next, we make one of the first attempts to comprehensively study the behavior of five different multiple distortion combinations in a two-stage distortion pipeline. We use the insights thus gained to introduce a major new paradigm which we call degraded-reference (DR) IQA and develop first-of-their-kind IQA models that estimate the quality of

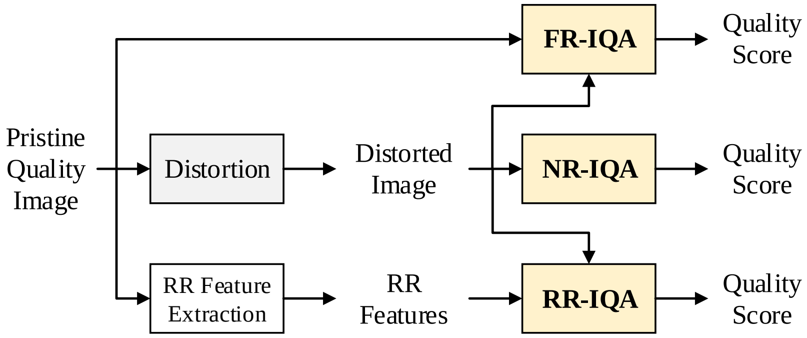


Figure 4.1: General framework of FR, RR and NR IQA.

the final distorted images by incorporating information from earlier degraded references. To aid in our development of such models, we also develop two new DR IQA databases that are used for model parameter estimation, and for model training and validation. These datasets have more than 30,000, mostly multiply distorted, images each, and we annotate them with the synthetic quality benchmark (SQB) developed in the previous chapter. Extensive performance evaluation of the DR IQA models reveals that they perform significantly better than contemporary FR and NR methods when applied in a multiple distortions environment.

4.1 Introduction

Objective IQA methods aim to predict the quality of images perceived by human eyes. As defined in earlier chapters, depending upon the accessibility to the pristine reference content, they are traditionally classified into *full-reference* (FR), *reduced-reference* (RR) and *no-reference* (NR) or BIQA methods [11,12], as illustrated in Figure 4.1. These three different categories of objective IQA methods essentially constitute three major paradigms in which contemporary image and video quality assessment research is ongoing. However, each of them has certain limitations:

- *Full-Reference* quality assessment algorithms need access to a *pristine reference* image in order to compute the quality score of distorted content. This has two drawbacks:

    1. In practice, perfect-quality *pristine reference* images may not exist because all digital images captured from the real world are affected by sensor noise. Even if a digital camera sensor of exceptionally high quality is used, a number of other factors such as exposure conditions, stability of camera platform, etc., may not be perfect.
    2. Even if we regard a high quality image with an acceptable amount of distortion as a “pristine” reference, access to such content may be limited in practical image distribution systems. Such images typically have very large data rate, which restricts their transmission over various networks, effectively limiting access to them.

- Although *reduced-reference* quality assessment algorithms only require some features from the reference image, access to the *pristine reference* image is still needed in order to extract those reference features. Moreover, additional cost such as an error-free ancillary channel needs to be paid to transmit the RR features. This also restricts the use of reduced-reference quality assessment algorithms.

- *No-reference* quality assessment algorithms do not suffer from the limitations of full-reference or reduced-reference algorithms since they do not need access to pristine reference content. However, their performance does not match that of the full-reference algorithms (as shown in Chapter 2). In addition, no-reference methods cannot check the fidelity against the original signals. As a result, a high-score of no-reference models cannot ensure the authenticity of the image data.

In the literature, the development of FR, RR, and NR IQA algorithms usually follows the general framework depicted in Fig. 4.1, that is, they are usually tested and at times trained on image databases of different distortion types, but typically, each distorted image has undergone a single stage of distortion. This is in clear contrast to real-world visual content distribution scenarios, as illustrated in Figure 4.2, where visual content may have

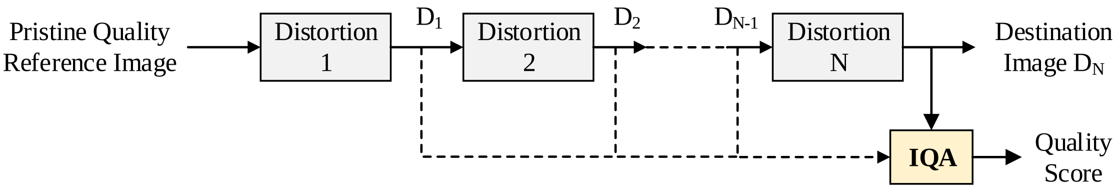


Figure 4.2: The framework of practical media distribution systems.

undergone multiple stages of distortions before reaching the target consumer devices, casting major challenges for the single distortion IQA framework of Fig. 4.1. Some examples are given as follows:

- Most consumer cameras and camcorders, including mobile phone cameras, store captured content using lossy compression standards such as JPEG and H.264/MPEG-4. When these images and videos are uploaded to a social networking website or a video-sharing website, they usually undergo another round of compression. For example:
  - YouTube recommends that 1080p videos having a standard frame rate (24, 25 and 30 frames per second) should have a bit rate of 8 Mbps if they are to be uploaded to YouTube [210]. In practice, this is often not satisfied. Moving forward, YouTube decodes and then transcodes such videos into a set of derivative video streams of different bandwidths and resolutions for onward delivery to viewers. This essentially means multiple levels of compression.
  - A similar example applies to images as well. It is known that Facebook compresses images if the file size is above a threshold [211]. Thus, a JPEG compressed image that is uploaded to Facebook may undergo further compression.
- A content producer or provider may send compressed content to a video distributor, who may subsequently compress the content again before transmitting to end users.
- An image or video maybe contaminated by noise or blur during acquisition because of different factors such as the limitation of the digital camera sensor [206], the

lack of sufficient exposure conditions, inadequate lighting, motion of photographer or object being photographed, etc. The camera will store this content in compressed form which may be followed by further compression during its distribution. This essentially means noise contamination followed by compression or blur followed by compression.

- Compressed medical images provide another example of content afflicted by multiple distortion stages. It is known that magnetic resonance (MR) images are affected by noise that has a Rician probability density function (PDF) [212], low-dose computed tomography (CT) images are affected by noise that has a Gaussian PDF [213], and Ultrasound images are affected by speckle noise [214]. With the rapid increase in the resolution and volume of medical images and with the emergence of tele-medicine, it is now desirable to reduce the data rate of medical images by lossy image compression as long as it does not affect the diagnostic quality [215,216]. This leads to a distortion combination of noise followed by lossy compression.
- Compressed astronomical images provide yet another example of noise followed by lossy compression since astronomical images are contaminated by noise [217].

From the above discussion it can be seen that even if we start with a *pristine reference* image, it may be affected by multiple stages of distortions by the time it reaches the end user. The distortions at different stages may be similar or different giving rise to a number of distortion combinations. The requirement for IQA methods that deal with multiple simultaneous distortions is not new (for example, see [218]), however, designing such IQA methods is quite challenging since the interactions of different distortions need to be accounted for. Thus, IQA for images with multiple simultaneous distortions has been a major challenge that future research needs to address [219]. As discussed earlier, in practical media delivery systems, access to pristine reference images is either extremely rare or altogether nonexistent, especially at the end user level. This, coupled with the multiple distortion nature of such systems, makes the use of FR and RR IQA infeasible. While NR IQA methods can be used to determine the quality of the final distorted image, most NR methods are trained and tested on subject-rated databases that have images with a single stage of distortion (see Section 2.3.3). Although there have been recent advances in the design of NR IQA methods to handle multiply distorted images using some new databases, such progress remains limited in scope. SISBLIM [32] is a training-free metric designed for singly and multiply distorted images through the fusion of estimates of noise, blur, JPEG compression, and joint effects. BoWSF [220] selects features sensitive to different distortion types, which are encoded through a Bag-of-Words model and mapped to a quality score. LQAF [221] uses SVR to map features such as phase congruency, gradient magnitude, gray level gradient co-occurrence matrix and the contrast sensitivity function to quality scores. An enhanced and multi-scale version of LQAF, called MS-LQAF is proposed in [181]. The training-based GWHGLBP [142] uses the gradient-weighted histogram of the local binary pattern (LBP) generated on the gradient map of the distorted image to capture the effects of multiple distortions. Jet-LBP [180] uses color Gaussian jets to generate feature maps from a distorted image. The LBP is applied to these feature maps to ascertain the effect of multiple distortions, leading to a weighted histogram which is mapped to quality scores through SVR. MUSIQUE [37] handles multiply distorted images and operates by performing distortion identification followed by distortion parameter estimation and score generation. However, due to their fundamental design philosophy, a major limitation of NR IQA algorithms is that they are incapable of incorporating various versions of an image as it progresses through the media delivery chain in the quality assessment task, even if such additional information is available.

With regard to the framework for practical media distribution systems depicted in Fig. 4.2, the question is: *How should the available information about distorted images at mid-stages be best utilized to ascertain the quality of the final multiply distorted image in the absence of the pristine reference?* A pioneering work in this direction is the corrupted-reference (CR) IQA scheme laid out in the context of an image restoration problem [205, 222, 223]. The quality of the denoised image with respect to an absent pristine reference image is estimated by using a Gaussian or Poisson noise contaminated corrupted reference image. However, CR IQA does not apply when determining the quality of a general final distorted image, outside of the restoration context. The recently developed two-step quality assessment (2stepQA) scheme [1, 2] is directly relevant to the practical quality assessment framework of Fig. 4.2. It is developed for images that have been afflicted with two distortions, where the second distortion is compression. 2stepQA operates in the absence of pristine reference images but assumes that both stage-1 (distorted reference) and stage-2 (final compressed) images are available. It uses an FR method to determine the quality of the final compressed image with respect to the non-compressed yet distorted reference image. The quality of the distorted reference image is itself determined through the use of an NR method. The FR and NR quality scores are then combined as a weighted product. The publicly released version of 2stepQA is as follows:

$$Q_{2stepQA} = \text{MSSSIM} \cdot \left(1 - \frac{\text{NIQE}}{\alpha}\right) \tag{4.1}$$

where the FR score is obtained by using MSSSIM [4] (i.e., Q_{FR} = MSSSIM), and the NR score is obtained by using NIQE [3] and rescaled so that it is in the same range as MSSSIM (i.e., $Q_{NR} = 1 - \frac{\text{NIQE}}{\alpha}$, where $\alpha = 100$ is used). Apart from this publicly released version of 2stepQA, other combinations of NR (NIQE [3], BRISQUE [140], CORNIA [141], PQR [167]) and FR (PSNR, MSSSIM [4], FSIM [14], VSI [15]) methods are also evaluated in [2] on the LIVE Wild Compressed Picture Quality Database. While 2stepQA [1,2] is a pioneering work to access the quality of a multiply distorted compressed image given its earlier distorted reference, it does not take into account how different distortions behave in conjunction with each other and is a rather ad hoc combination of an NR and FR method.

In this chapter, we make one of the first attempts to develop IQA models that evaluate the quality of multiply distorted images by taking into account how different distortions interact with each other. We start by restricting ourselves to two stages of distortion. Thus, a *pristine reference* image will lead to a *degraded reference* image after passing through stage-1 distortion and a degraded reference image will lead to a *final distorted* image after passing through stage-2 distortion. We discussed earlier that access to *pristine reference* images is limited in practice. Therefore, in a practical two-stage distortion scenario, the task of image quality assessment may be defined as follows:

Definition 1 *Degraded-reference image quality assessment - Determining the quality of a final distorted image given access to a degraded reference image, but with no access to the pristine reference image.*

The above-mentioned definition leads to a *degraded reference image quality assessment* (DR IQA) framework, which we believe is the fourth major paradigm in IQA research, the other three being FR, RR, and NR IQA.

4.2 Baseline Performance Evaluation

Before moving on to the development of DR IQA models, we first evaluate the performance of some FR and NR methods, and the 2stepQA [1, 2] model on multiply distorted images in the next section. This will establish a baseline against which the performance of DR IQA models, developed later in this chapter, will be evaluated.

4.2.1 Databases, Methods and Criteria used for Comparison

In this sub-section, we define the IQA databases and evaluation criteria that will be used to not only evaluate the performance of baseline methods in this section, but will also be used to evaluate the performance of DR IQA methods built later in this chapter.

Databases

We will use the following four IQA databases for performance evaluation:

1. The **Waterloo Exploration-II (Waterloo Exp-II)** database that we constructed and described in detail in Chapter 3 (see Section 3.3). From Table 3.3 it can be seen that this dataset has 3,570 pristine, 39,270 singly distorted images each for Blur, JPEG compression, and Noise, and 667,590 multiply distorted images each for the distortion combinations of Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, and Noise-JPEG2000. The singly distorted images can essentially be regarded as *degraded references* while the multiply distorted images can be regarded as *final distorted* images in a 2-stage distortion process. As noted in Chapter 3, it is not possible to acquire annotations for such a large dataset from human subjects, and thus the

quality labels for this dataset, which we called synthetic quality benchmark (SQB), have been generated by fusing the results from four state-of-the-art FR methods. The SQB generation process has been described in detail in Section 3.4.2.

2. The recently released **LIVE Wild Compressed (LIVE WCmp)** database [1, 2, 224] is composed of 400 images. It starts with 80 authentically distorted images that it takes from the LIVE Wild Challenge database [79] (for the definition of authentically distorted images, see Section 2.2.3), which can be regarded as degraded references. Each of the 80 authentically distorted images are further compressed using JPEG compression at four fixed compression levels regardless of content, leading to a total of 320 final distorted images. Subjective testing was carried out by using the single stimulus methodology [70] and each subject participated in two 30-minute test sessions. After undergoing a training session, subjects rated the quality of test images by moving a slider on a continuous scale that had been marked with five adjectives: Bad, Poor, Fair, Good, and Excellent (from left to right). Numerical quality scores in the range of 1 to 100 were sampled from the location of the slider for each test image. Subjective scores were then computed for each test image in the form of MOS according to the procedures outlined in [70, 225]. It should be noted that this dataset does not have pristine reference images.

3. The **LIVE Multiply Distorted (LIVE MD)** database [31, 66] has been described earlier in Section 2.2.2 and in Table 2.2. Suffice it to say that this database consists of 15 pristine reference images, 45 singly distorted images each for Blur, JPEG Compression and Noise, and 135 multiply distorted images each for the distortion combinations of Blur-JPEG and Blur-Noise. For our analysis, we will consider the singly distorted Blur images as degraded references and the multiply distorted Blur-JPEG and Blur-Noise images as the final distorted images. LIVE MD provides subjective ratings for all its images in the form of DMOS.

4. The **Multiple Distorted IVL (MDIVL)** database [34, 68, 69] has been described earlier in Section 2.2.2 and in Table 2.2, where we note that it consists of 10 pristine reference and 750 multiply distorted images of which 350 belong to the Blur-JPEG combination while 400 belong to the Noise-JPEG combination. In both these combi-

nations, the second distortion is JPEG compression. Since the MDIVL database does not contain any singly distorted images, this limitation apparently makes it infeasible to include it in our analysis. However, we note that the 350 Blur-JPEG images are obtained by first distorting the pristine references at seven levels of Gaussian blur and then further distorting these singly distorted images at five levels of JPEG compression. Similarly, the 400 Noise-JPEG images are obtained by first distorting the pristine references at ten levels of Gaussian noise and then further distorting these singly distorted images at four levels of JPEG compression. In both the Blur-JPEG and Noise-JPEG combinations, the distortion level leading to the least compression utilizes the MATLAB JPEG compression quality factor of 100 at which compression artifacts are perceptually unapparent. Thus, we can regard 70 out of 350 Blur-JPEG images as singly distorted Blur images, and 100 out of 400 Noise-JPEG images as singly distorted Noise images, thereby providing us with degraded references and final distorted images. This enables the use of the MDIVL database in our analysis. MDIVL provides subjective ratings for all its images in the form of MOS.

We believe that the above-mentioned four datasets allow us to: 1) Test the performance of different models on a very large amount of data by utilizing the synthetically annotated Waterloo Exploration-II dataset, and 2) Test the performance of different models with respect to human-rated datasets (LIVE Wild Compressed, LIVE MD, MDIVL). These are also the only datasets that provide both singly distorted degraded references and their respective multiply distorted final distorted images. Although two other IQA datasets, the MDID database [33] and the MDID2013 database [32], contain multiply distorted images, they do not provide degraded references and hence cannot be used. We will also describe the construction of two other datasets later in this chapter, which are designed on the pattern of the very large-scale Waterloo Exploration-II dataset, albeit at a smaller scale. In terms of content, these datasets do not have any overlap with the four testing databases mentioned above, and will only be used for model development in the subsequent sections, not for testing.

Methods used for Comparison

Even though the FR methods IWSSIM [13], DSS [16], CID_MS [95], and VIF_DWT [93] were found to perform well in Section 2.4 (especially IWSSIM and DSS) we do not include them in our baseline performance analysis here as they are part of the SQB used to annotate the quality of images in the Waterloo Exploration-II database which is one of the testing datasets. Outside of these methods, we choose FSIMc [14] as our main FR method since it outperforms most other FR methods as can be seen in Section 2.4. We also analyze the performance of the FR method MSSSIM [4] as it is used as the FR component in the 2stepQA model [2]. Among NR methods, we evaluate the performance of CORNIA [141] and dipIQ [36] as they were found to be the top performers in Section 2.5. We also analyze the performance of the NR method NIQE [3] as it is used as the NR component in the 2stepQA model [2]. Finally, we evaluate the performance of the 2stepQA model [1, 2] as well.

Evaluation Criteria

We use the Pearson Linear Correlation Coefficient (PLCC) and the Spearman Rank-order Correlation Coefficient (SRCC) as measures of a model's prediction accuracy and prediction monotonicity, respectively [72]. Both PLCC and SRCC are computed between a model's predicted quality scores for dataset images and their MOS/DMOS/SQB values. PLCC is computed after a nonlinear mapping step whereas SRCC is computed directly. These evaluation criteria have been described in detail in Section 2.4.1. The computation of PLCC and SRCC is done at the level of the entire dataset and also for each distortion combination contained within a dataset. Thus, for the Waterloo Exp-II database, PLCC and SRCC are computed for the distortion combinations of Blur-JPEG (B-JPG), Blur-Noise (B-N), JPEG-JPEG (JPG-JPG), Noise-JPEG (N-JPG), Noise-JPEG2000 (N-JP2), and for the entire dataset (All Data). Additionally, where possible, to allow for better comparisons with [1, 2], we combine images belonging to distortion combinations where the final distortion is JPEG compression, i.e., Noise-JPEG, Blur-JPEG, and JPEG-JPEG and call this subset of multiply distorted images as Noise/Blur/JPEG-JPEG (NBJ-JPG).

4.2.2 Performance of FR Methods

FR IQA methods require the availability of a reference image in order to give a quality score to a distorted image. In a two-stage distortion scenario, where a pristine reference leads to a degraded reference which further leads to a final distorted image, and given that FR methods can only compare two images at a time, there are two possibilities when it comes to determining the quality of the final distorted image by using FR methods:

1. Use FR methods to determine the quality of the final distorted image with respect to the pristine reference. The quality estimate thus obtained can be regarded as the absolute quality of the final distorted image as comparison is being made with the pristine reference. However, as discussed earlier, practically pristine references do not exist or are unavailable and thus such absolute quality estimates of the final distorted image are not feasible.
2. Another possibility is to use FR methods to ascertain the relative quality of the final distorted image with respect to the degraded reference. Such a comparison is possible owing to the availability of degraded references, however, it does not lead to an absolute quality estimate since comparison with pristine references is missing.

The premise of DR IQA is that we do not have ready access to the pristine reference image and thus the absolute quality of the final distorted image cannot be determined directly. A straightforward method is to use the relative quality of the final distorted image with respect to the degraded reference as a predictor of its absolute quality. We regard this approach as the first baseline model and it is depicted in Fig. 4.3. To evaluate the performance of this baseline model, we use FR methods FSIMc [14] and MSSSIM [4] to compute the quality of the final distorted images in the Waterloo Exp-II, LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases with respect to their degraded references. These relative quality scores of the final distorted images are then compared with their respective SQB values (in case of the Waterloo Exp-II database) or with their respective MOS/DMOS values (in case of the LIVE MD, MDIVL, and LIVE WCmp databases) through the computation of PLCC and SRCC. The results are given in Table 4.1. However,

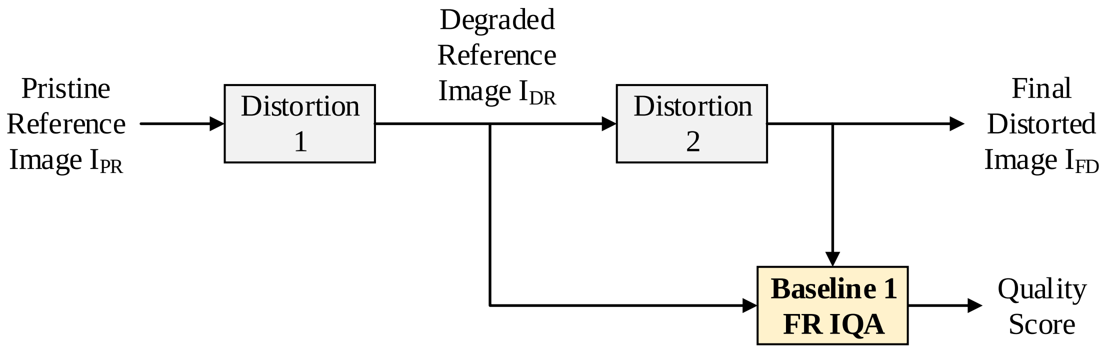

Pristine Reference Image I_{PR}
Distortion 1
Degraded Reference Image I_{DR}
Distortion 2
Final Distorted Image I_{FD}
Baseline 1 FR IQA
Quality Score

Figure 4.3: Baseline Model 1: Applying FR IQA methods between the degraded reference and final distorted images.

Table 4.1: Performance of FR methods when used to determine the quality of final distorted images with respect to degraded references and using it as the final quality measure.

Database	Correlation Metric	FR Method	Distortion Combination						
			B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	FSIMc	0.8436	0.8826	0.8276	0.8280	0.8223	0.7980	0.7926
		MSSSIM	0.8567	0.9473	0.8340	0.8809	0.8039	0.7498	0.7425
	SRCC	FSIMc	0.8442	0.8843	0.7544	0.8195	0.8154	0.7964	0.7816
		MSSSIM	0.8561	0.9481	0.7467	0.8803	0.7972	0.7579	0.7293
LIVE MD[b]	PLCC	FSIMc	0.2256	0.3882	–	–	–	–	0.3045
		MSSSIM	0.2366	0.4270	–	–	–	–	0.2254
	SRCC	FSIMc	0.1923	0.3336	–	–	–	–	0.2446
		MSSSIM	0.1370	0.3671	–	–	–	–	0.2076
MDIVL[b]	PLCC	FSIMc	0.5207	–	–	0.8111	–	–	0.6238
		MSSSIM	0.4984	–	–	0.8770	–	–	0.5985
	SRCC	FSIMc	0.4870	–	–	0.8243	–	–	0.6316
		MSSSIM	0.4619	–	–	0.8781	–	–	0.5698
LIVE WCmp[b,c]	PLCC	FSIMc	–	–	–	–	–	–	0.9030
		MSSSIM	–	–	–	–	–	–	0.8498
	SRCC	FSIMc	–	–	–	–	–	–	0.9024
		MSSSIM	–	–	–	–	–	–	0.8469

[a] PLCC and SRCC are computed with respect to SQB.

[b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its PLCC and SRCC values cannot be placed in a particular distortion combination.

Table 4.2: Performance of FR methods when used to determine the quality of final distorted images with respect to pristine references.

Database	Correlation Metric	FR Method	Distortion Combination						
			B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	FSIMc	0.9153	0.8990	0.9157	0.8932	0.9077	0.9110	0.9094
		MSSSIM	0.9363	0.9804	0.9470	0.8980	0.8989	0.9178	0.9043
	SRCC	FSIMc	0.9120	0.8956	0.9119	0.8878	0.9065	0.9076	0.9080
		MSSSIM	0.9366	0.9809	0.9488	0.8906	0.9100	0.9204	0.9126
LIVE MD[b]	PLCC	FSIMc	0.7563	0.7884	–	–	–	–	0.7690
		MSSSIM	0.7074	0.7738	–	–	–	–	0.6990
	SRCC	FSIMc	0.7066	0.7850	–	–	–	–	0.7517
		MSSSIM	0.6844	0.7614	–	–	–	–	0.6941
MDIVL[b]	PLCC	FSIMc	0.8909	–	–	0.9193	–	–	0.8874
		MSSSIM	0.8370	–	–	0.8996	–	–	0.8645
	SRCC	FSIMc	0.8402	–	–	0.8765	–	–	0.8354
		MSSSIM	0.7978	–	–	0.8175	–	–	0.8041

[a] PLCC and SRCC are computed with respect to SQB.

[b] PLCC and SRCC are computed with respect to MOS/DMOS.

these results cannot be considered independently as they are for the relative quality scenario whereas FR methods require the reference image to be of pristine quality. Thus, it is hard to pinpoint, solely by looking at Table 4.1, whether any loss of performance is due to the nature of the baseline model (Fig. 4.3) or due to any issues within the FR methods themselves. To alleviate this issue, we also use the above-mentioned FR methods to compute the absolute quality scores of the final distorted images with respect to their pristine references for each database and present the PLCC and SRCC results in Table 4.2 which is a true representation of the performance of these FR methods when used to determine the quality of multiply distorted images. Since the LIVE WCmp database does not have pristine references, absolute quality scores for its final distorted images cannot be computed and hence it is not present in Table 4.2. When Tables 4.1 and 4.2 are considered together, it can be observed that the first baseline model, based on the relative FR quality scores between the final distorted and degraded reference images, is not a good predictor of the absolute FR quality scores between the final distorted and pristine reference images. This is apparent for all individual distortion combinations, for the NBJ-JPG case (Waterloo Exp-II database), and for the entire dataset case, for the Waterloo Exp-II, LIVE MD, and MDIVL databases (the only exception being the SRCC value of MDIVL for the N-JPG

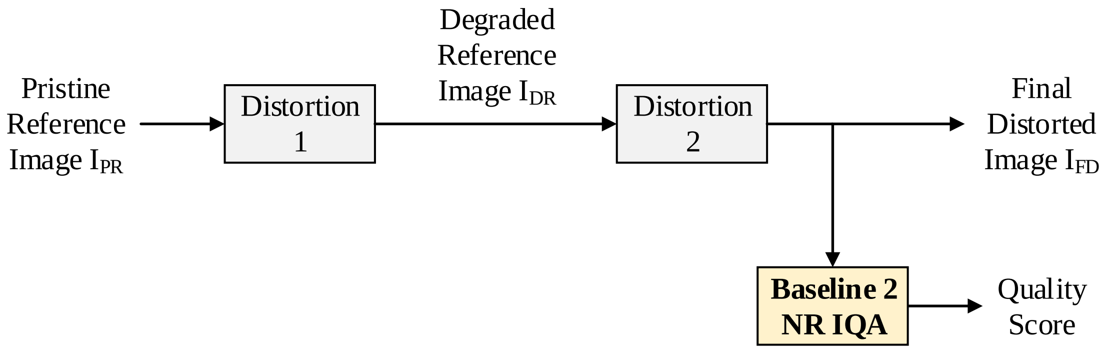

Pristine Reference Image I_{PR}
Distortion 1
Degraded Reference Image I_{DR}
Distortion 2
Final Distorted Image I_{FD}
Baseline 2 NR IQA
Quality Score

Figure 4.4: Baseline Model 2: Applying NR IQA methods directly to final distorted images.

case). The loss in performance is at times quite significant, for example, for the entire LIVE MD database and its two distortion combinations, for the entire MDIVL database and its B-JPG distortion combination, for the entire Waterloo Exp-II database and its B-JPG, JPG-JPG, N-JPG (only for FSIMc), N-JP2, and NBJ-JPG distortion combinations. Thus, it can be concluded that the first baseline model depicted in Fig. 4.3 is not a good predictor of the quality of multiply distorted images.

4.2.3 Performance of NR Methods

A simple counterargument to the entire premise of DR IQA is why not simply use NR IQA methods to predict the quality of the final distorted images directly. Many NR IQA algorithms exist currently and they do not need any reference image to calculate the quality score for a given distorted image. Thus, we consider the use of NR IQA methods as our second baseline model as depicted in Fig. 4.4. To evaluate the performance of this baseline model, we use NR methods CORNIA [141], dipIQ [36], and NIQE [3] to directly compute the quality of the final distorted images in the Waterloo Exp-II, LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases. Table 4.3 depicts the PLCC and SRCC values computed between these NR predicted quality scores and the SQB (Waterloo Exp-II database) or MOS/DMOS (LIVE MD, MDIVL, LIVE WCmp databases) values of the

Table 4.3: Performance of NR methods when used to determine the quality of final distorted images and using it as the final quality measure.

Database	Correlation Metric	NR Method	Distortion Combination						
			B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	CORNIA	0.8918	0.6205	0.7512	0.7832	0.6943	0.8172	0.7553
		dipIQ	0.8522	0.9414	0.8790	0.8462	0.8380	0.8422	0.8532
		NIQE	0.7741	0.8941	0.7084	0.6368	0.6913	0.7030	0.7137
	SRCC	CORNIA	0.8955	0.5940	0.7539	0.7826	0.7131	0.8101	0.7576
		dipIQ	0.8508	0.9443	0.8786	0.8403	0.8549	0.8437	0.8516
		NIQE	0.7670	0.8939	0.6978	0.6150	0.6975	0.6890	0.6891
LIVE MD[b]	PLCC	CORNIA	0.7141	0.8144	–	–	–	–	0.7360
		dipIQ	0.5238	0.6603	–	–	–	–	0.5531
		NIQE	0.7677	0.6670	–	–	–	–	0.5802
	SRCC	CORNIA	0.6897	0.7997	–	–	–	–	0.7278
		dipIQ	0.4823	0.5706	–	–	–	–	0.4548
		NIQE	0.7487	0.6359	–	–	–	–	0.5512
MDIVL[b]	PLCC	CORNIA	0.9331	–	–	0.7748	–	–	0.7963
		dipIQ	0.8298	–	–	0.8074	–	–	0.7514
		NIQE	0.7910	–	–	0.5357	–	–	0.5731
	SRCC	CORNIA	0.9202	–	–	0.8101	–	–	0.8157
		dipIQ	0.6561	–	–	0.8393	–	–	0.7423
		NIQE	0.7558	–	–	0.5791	–	–	0.5946
LIVE WCmp[b,c]	PLCC	CORNIA	–	–	–	–	–	–	0.8424
		dipIQ	–	–	–	–	–	–	0.7978
		NIQE	–	–	–	–	–	–	0.8314
	SRCC	CORNIA	–	–	–	–	–	–	0.8471
		dipIQ	–	–	–	–	–	–	0.7868
		NIQE	–	–	–	–	–	–	0.8327

[a] PLCC and SRCC are computed with respect to SQB.

[b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its PLCC and SRCC values cannot be placed in a particular distortion combination.

final distorted images. The following observations can be made: 1) The performance of all three NR methods being tested shows significant room for improvement when it comes to evaluating the quality of multiply distorted images. While there are exceptions, for example, CORNIA performs well for the B-JPG case of Waterloo Exp-II and MDIVL databases, NIQE performs well for the B-N case of the Waterloo Exp-II database, and both CORNIA and NIQE perform satisfactorily on the LIVE WCmp database, these NR methods perform unsatisfactorily for other distortion combinations and for the entire dataset case of the Waterloo Exp-II, LIVE MD and MDIVL databases. Such inconsistencies indicate that these methods do not offer reliable performance. 2) It can be seen from Table 4.3 that dipIQ performs rather well across all distortion combinations and for the entire data case of the Waterloo Exp-II database. However, it should be noted that the SQB quality labels of images in the Waterloo Exp-II database were generated through a rank-based method (RRF [23]) while dipIQ [36] also uses a rank-based algorithm (RankNet [170]). Thus, the Waterloo Exp-II database may not lead to a completely unbiased evaluation of dipIQ and other datasets should also be considered for its performance evaluation. 3) A comparison of Table 4.3 with Table 4.1 shows that the NR based baseline model performs better than the FR based baseline model on the entire LIVE MD database and both its distortion combinations, and also on the entire MDIVL database and its B-JPG distortion combination. However, the FR based baseline model does better than the NR based baseline model in case of the LIVE WCmp database and the N-JPG distortion combination of the MDIVL database. For the Waterloo Exp-II database, the FR based baseline model performs better on the JPG-JPG, N-JPG and N-JP2 distortion combinations and on the entire dataset, while the NR based baseline model performs better on the B-JPG (except NIQE), B-N (except CORNIA) and the NBJ-JPG (except NIQE) distortion combinations. 4) A comparison of Table 4.3 with Table 4.2 shows that the NR based baseline model cannot outperform the FR based absolute quality scores between the final distorted and pristine reference images on the Waterloo Exp-II database. On the LIVE MD database, CORNIA and NIQE offer performance comparable to the FR based absolute quality scores for the B-JPG distortion combination. For the B-N combination and the all data case of the LIVE MD database and for both distortion combinations and the all data case of the MDIVL database, the NR based baseline model cannot outperform the FR based absolute quality scores with the exception of CORNIA which either performs better or in a comparable manner. From the above discussion, it can be concluded that the NR based baseline model, depicted in Fig. 4.4, does not offer robust performance while evaluating the quality of multiply distorted images. This can be attributed to the difficult nature of the NR IQA design philosophy, where the quality of an image has to be determined without the help of any side information, which is why NR IQA is also referred to as *blind* IQA. The case of multiply distorted images further complicates the NR IQA task.

Table 4.4: Performance of the LIVE 2stepQA model [1, 2]. NIQE [3] is used to determine the quality of the degraded reference and MSSSIM [4] is used to determine the quality of the final distorted image with respect to the degraded reference.

Database	Correlation	Distortion Combination						
	Metric	B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	0.9340	0.9696	0.8951	0.8420	0.7213	0.7709	0.7140
	SRCC	0.9337	0.9708	0.8669	0.8342	0.7162	0.7797	0.7274
LIVE MD[b]	PLCC	0.7746	0.6730	–	–	–	–	0.6500
	SRCC	0.7530	0.5356	–	–	–	–	0.5318
MDIVL[b]	PLCC	0.8697	–	–	0.7964	–	–	0.8149
	SRCC	0.8539	–	–	0.7685	–	–	0.7713
LIVE WCmp[b,c]	PLCC	–	–	–	–	–	–	0.9229
	SRCC	–	–	–	–	–	–	0.9246

[a] PLCC and SRCC are computed with respect to SQB.

[b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its PLCC and SRCC values cannot be placed in a particular distortion combination.

4.2.4 Performance of the 2stepQA Model

Since the 2stepQA [1, 2] (discussed in Section 4.1) is the only model that utilizes the quality information about the degraded reference while determining the quality of a multiply distorted image, establishing its performance on our test datasets as a third baseline is essential so that the performance of DR IQA models developed later in this chapter can be compared with it. We also compare the performance of 2stepQA with that of the first and second baseline models discussed in Sections 4.2.2 and 4.2.3, respectively.

As discussed earlier, 2stepQA evaluates the NR score of a degraded reference image by using NIQE [3] and the relative FR score of the final distorted image with respect to the degraded reference by using MSSSIM [4], and then combines these two scores as given in Equation 4.1. We apply the 2stepQA approach to determine the quality of the final distorted images of the Waterloo Exp-II, LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases. Table 4.4 depicts the PLCC and SRCC values computed between these 2stepQA model predicted quality scores and the SQB (Waterloo Exp-II database) or MOS/DMOS (LIVE MD, MDIVL, LIVE WCmp databases) values of the final distorted images. The following observations can be made: 1) The 2stepQA model performs well on the B-JPG, B-

N, and JPG-JPG cases of the Waterloo Exp-II database, on the B-JPG case of the MDIVL database, and on the LIVE WCmp database. It also performs satisfactorily on the N-JPG combination of the Waterloo Exp-II database. However, it does not perform well on the N-JP2, NBJ-JPG, and the all data cases of the Waterloo Exp-II database. It also does not perform well on the entire LIVE MD database and its distortion combinations, and on the N-JPG and all data cases of the MDIVL database. 2) When Table 4.4 is compared with Table 4.1, it can be seen that the 2stepQA model is unable to perform better than the FR based baseline model on the Waterloo Exp-II database for the cases of N-JPG (MSSSIM performs better), N-JP2, NBJ-JPG (FSIMc performs better), and the all data case, and on the MDIVL database for the case of N-JPG. However, it performs better than the FR based baseline model in all other cases. 3) When Table 4.4 is compared with Table 4.3, it can be seen that while the 2stepQA model performs better than the NR based baseline model for quite a number of cases, it is itself outperformed by NR models on some cases, such as by CORNIA on the NBJ-JPG and all data cases of the Waterloo Exp-II database, by CORNIA on the B-N and all data cases of the LIVE MD database, by CORNIA on the B-JPG case and by dipIQ on the N-JPG case of the MDIVL database. It should be noted that we are not comparing the performance of dipIQ with 2stepQA on the Waterloo Exp-II database as this dataset may favor the former. 4) When Table 4.4 is compared with Table 4.2, it can be observed that 2stepQA is only able to perform better than the FR based absolute quality scores on the B-JPG cases of the LIVE MD and MDIVL databases. For all other cases, the 2stepQA model is outperformed by the FR based absolute quality scores, sometimes quite significantly. It can be concluded from the above analysis that while the 2stepQA model performs adequately for some distortion combinations (B-JPG, B-N, JPG-JPG), its performance remains lacking in other distortion combinations (N-JPG, N-JP2, NBJ-JPG, all data) and there is substantial room for improvement, which highlights the need for further research in the area of IQA of multiply distorted images.

4.3 Degraded Reference IQA: A New Paradigm

Research on the development of IQA models for multiply distorted images has thus far almost exclusively focused on the formation of NR models, as discussed in Section 4.1.

Given the framework of practical media distribution systems, as depicted in Fig. 4.2, earlier degraded versions of a final distorted image maybe available for the task of quality assessment of the final image. For example, at an encoder, both the input and output images are available for the task of quality assessment of the output image. However, due to their design philosophy, NR methods are unable to use this *additional* information about the final distorted image. Thus, the area of image quality assessment of a final distorted image, that takes into account its earlier degraded versions, is quite new. The only known work in this area is the development of the 2stepQA model [1,2], whose performance shows significant room for further development as evident from Section 4.2.4. Our aim in this chapter is to make a first attempt to comprehensively explore this new area.

Given the diverse nature of this topic and the lack of substantial research in it thus far, the first questions that arise are: *Where should we start*? *How many stages of distortions should be considered*? *What kind of distortions should be considered*? We answer these questions as follows:

- While practical images and videos may undergo many stages of distortions between the source and the end user, we begin by considering only two stages of distortions between the original source and the final destination. Since the interaction of even two simultaneous distortions has not been studied in depth so far, we believe this to be a logical starting point. Thus, an original source will generate a Pristine Reference (PR) image which will undergo stage-1 distortion and lead to a Degraded Reference (DR) image, which will itself undergo stage-2 distortion and lead to the Final Distorted (FD) image.

- As discussed in Section 2.2.3, IQA data can either be composed of *authentic* or *simulated* distortions. As the names imply, the former kind of distortions are captured in the real world whereas the latter are added to source content in a controlled manner, i.e., the distortion process is known. Intuitively, research on IQA models for multiply distorted content should focus on authentically distorted images as they are multiply distorted by their very nature. However, given the diverse nature of authentic distortions and the lack of understanding of how even well-known simulated distortions behave in conjunction with each other, we restrict ourselves to simulated distortion content. Specifically, we consider Gaussian noise, Gaussian blur, and JPEG compression to be stage-1 distortions, and Gaussian Noise, JPEG compression, and JPEG2000 compression to be stage-2 distortions. The multiply distorted images that we deal with include the following five distortion combinations: 1) Blur-JPEG (B-JPG), 2) Blur-Noise (B-N), 3) JPEG-JPEG (JPG-JPG), 4) Noise-JPEG (N-JPG), and 5) Noise-JPEG2000 (N-JP2). These distortion combinations have already been introduced in the Waterloo Exp-II database in Section 3.3, where justifications for the choice of distortion types for singly distorted images and distortion combinations for multiply distorted images have been provided in Section 3.3.2. While these multiple distortion combinations are able to represent various multiple distortion scenarios discussed in Section 4.1, we believe that an even more diverse set of distortion combinations should be considered in the future.

The task of degraded reference (DR) IQA is to determine the quality of the FD image given access to the DR image but without accessing the PR image. For now, to facilitate multiple distortions behavior analysis, let us assume that the PR image is also available. Thus, in a two-stage distortion pipeline, three kinds of images exist, i.e., PR, DR, and FD. Among the three major IQA frameworks (FR, RR, and NR), the performance of FR IQA is well established (see Chapter 2). Since FR methods can only evaluate the quality of a distorted image with respect to a reference image (i.e., only two images can be compared through FR IQA at a time), three kinds of FR comparisons are possible in a two-stage distortion pipeline. Two of these comparisons involve the pristine reference and thus, they will lead to *absolute* quality scores with respect to the PR image, as given in Equations 4.2 and 4.3:

$$\mathrm{AS_{DR}} = \mathrm{FR}(\mathrm{I_{PR}}, \mathrm{I_{DR}}), \tag{4.2}$$

$$\mathrm{AS_{FD}} = \mathrm{FR}(\mathrm{I_{PR}}, \mathrm{I_{FD}}), \tag{4.3}$$

where $\mathrm{I_{PR}}$ is the pristine reference image, $\mathrm{I_{DR}}$ is the degraded reference image, $\mathrm{I_{FD}}$ is the final distorted image, FR can be any state-of-the-art full-reference method, $\mathrm{AS_{DR}}$ is the absolute score of the DR image with respect to the PR image, and $\mathrm{AS_{FD}}$ is the absolute score of the FD image with respect to the PR image. The third possible FR comparison is between the DR and FD images. Since this comparison does not involve the pristine reference, the quality scores generated as a result cannot be regarded as absolute quality scores. Instead, they can only be regarded as *relative* scores. This comparison is given in Equation 4.4:

$$\mathrm{RS_{FD}} = \mathrm{FR}(\mathrm{I_{DR}}, \mathrm{I_{FD}}), \tag{4.4}$$

where $\mathrm{I_{PR}}$, $\mathrm{I_{DR}}$, $\mathrm{I_{FD}}$, and FR have already been defined above, and $\mathrm{RS_{FD}}$ is the relative score of the FD image with respect to the DR image.

Ideally, in a two-stage distortion pipeline, $\mathrm{AS_{FD}}$ is the score that would lead to the best estimate of the quality of the FD image when using FR IQA. However, in the absence of the PR image, only $\mathrm{RS_{FD}}$ can be obtained. In Section 4.2.2, we considered $\mathrm{RS_{FD}}$ as our first FR based baseline model and have already evaluated its performance (see Table 4.1). We have also evaluated the performance of $\mathrm{AS_{FD}}$ in Section 4.2.2 (see Table 4.2), where we observed that $\mathrm{RS_{FD}}$ is not a good predictor of $\mathrm{AS_{FD}}$. We will elaborate this in more detail in the next subsection.

4.3.1 Multiple Distortions Behavior Analysis

To observe how different distortions interact with each other and why $\mathrm{RS_{FD}}$ is not a good predictor of $\mathrm{AS_{FD}}$, let us first visually examine how two distortions belonging to fixed distortion levels interact with each other. We use the *Barbara* image as our pristine test image and distort it using various stage-1 distortions at distortion level-7 to create degraded references which are then further distorted using various stage-2 distortions at distortion level-11 to create the final distorted images. For a detailed discussion of what these distortion levels mean, refer to Section 3.3.2 and to Table 3.2. To highlight the impact of various distortions on each other and on image content, we utilize the *quality map* feature of the FR method SSIM [111]. SSIM uses a sliding window approach across the images that it compares and a local SSIM index is computed pixel by pixel. This results in an SSIM index map or *quality map*, where dark regions signify loss of quality while bright regions represent better quality. We also use the FR method FSIMc [14] to compute image-level AS_{DR}, AS_{FD}, and RS_{FD} scores. By using these tools, the following five examples are generated:

1. **Blur-JPEG** (Fig. 4.5): The PR *Barbara* image of Fig. 4.5 (a) is distorted at Gaussian blur level-7 to generate the DR image of Fig. 4.5 (b), which is then further distorted at JPEG compression level-11 to generate the FD image of Fig. 4.5 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.5 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.5 (e) and 4.5 (f), respectively.

2. **Blur-Noise** (Fig. 4.6): The PR *Barbara* image of Fig. 4.6 (a) is distorted at Gaussian blur level-7 to generate the DR image of Fig. 4.6 (b), which is then further distorted at Gaussian noise level-11 to generate the FD image of Fig. 4.6 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.6 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.6 (e) and 4.6 (f), respectively.

3. **JPEG-JPEG** (Fig. 4.7): The PR *Barbara* image of Fig. 4.7 (a) is distorted at JPEG compression level-7 to generate the DR image of Fig. 4.7 (b), which is then further distorted at JPEG compression level-11 to generate the FD image of Fig. 4.7 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.7 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.7 (e) and 4.7 (f), respectively.

4. **Noise-JPEG** (Fig. 4.8): The PR *Barbara* image of Fig. 4.8 (a) is distorted at Gaussian noise level-7 to generate the DR image of Fig. 4.8 (b), which is then further distorted at JPEG compression level-11 to generate the FD image of Fig. 4.8 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.8 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.8 (e) and 4.8 (f), respectively.

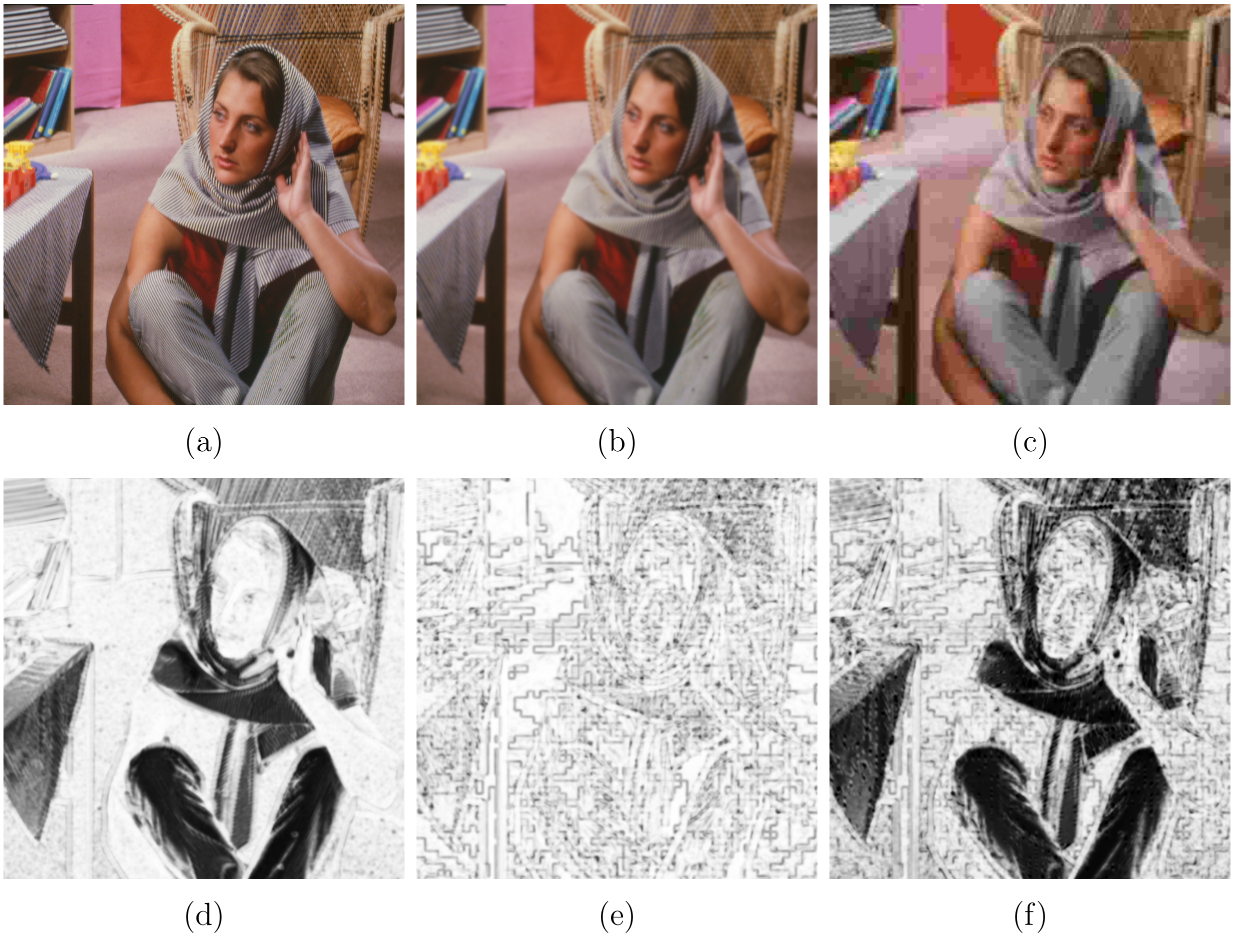


Figure 4.5: Example of the *Blur-JPEG* distortion combination. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by contaminating the image in (a) with Gaussian blur (level 7). (c) Final distorted *Barbara* image obtained by compressing the image in (b) by using JPEG compression (level 11). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).

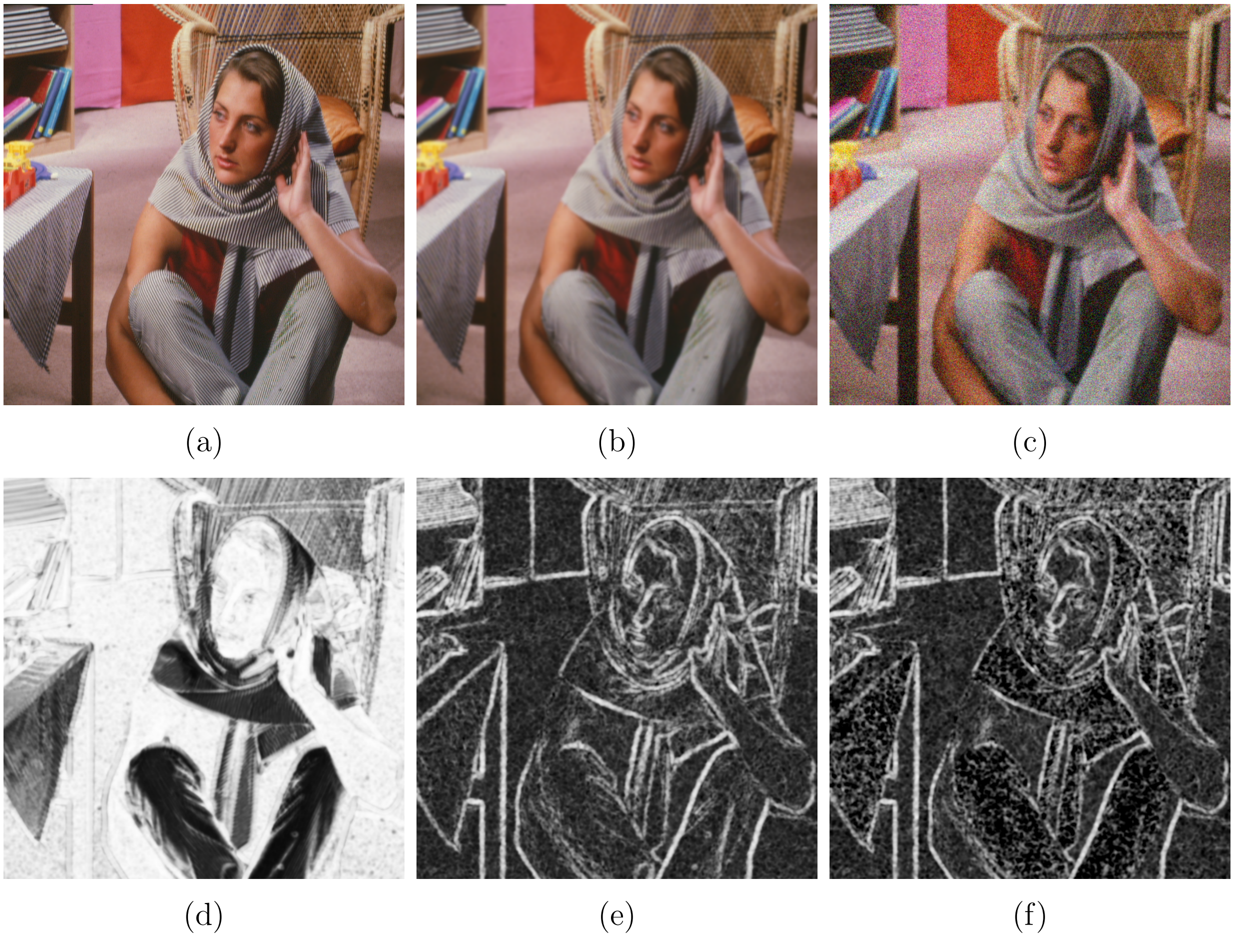


Figure 4.6: Example of the *Blur-Noise* distortion combination. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by contaminating the image in (a) with Gaussian blur (level 7). (c) Final distorted *Barbara* image obtained by contaminating the image in (b) with white Gaussian noise (level 11). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).

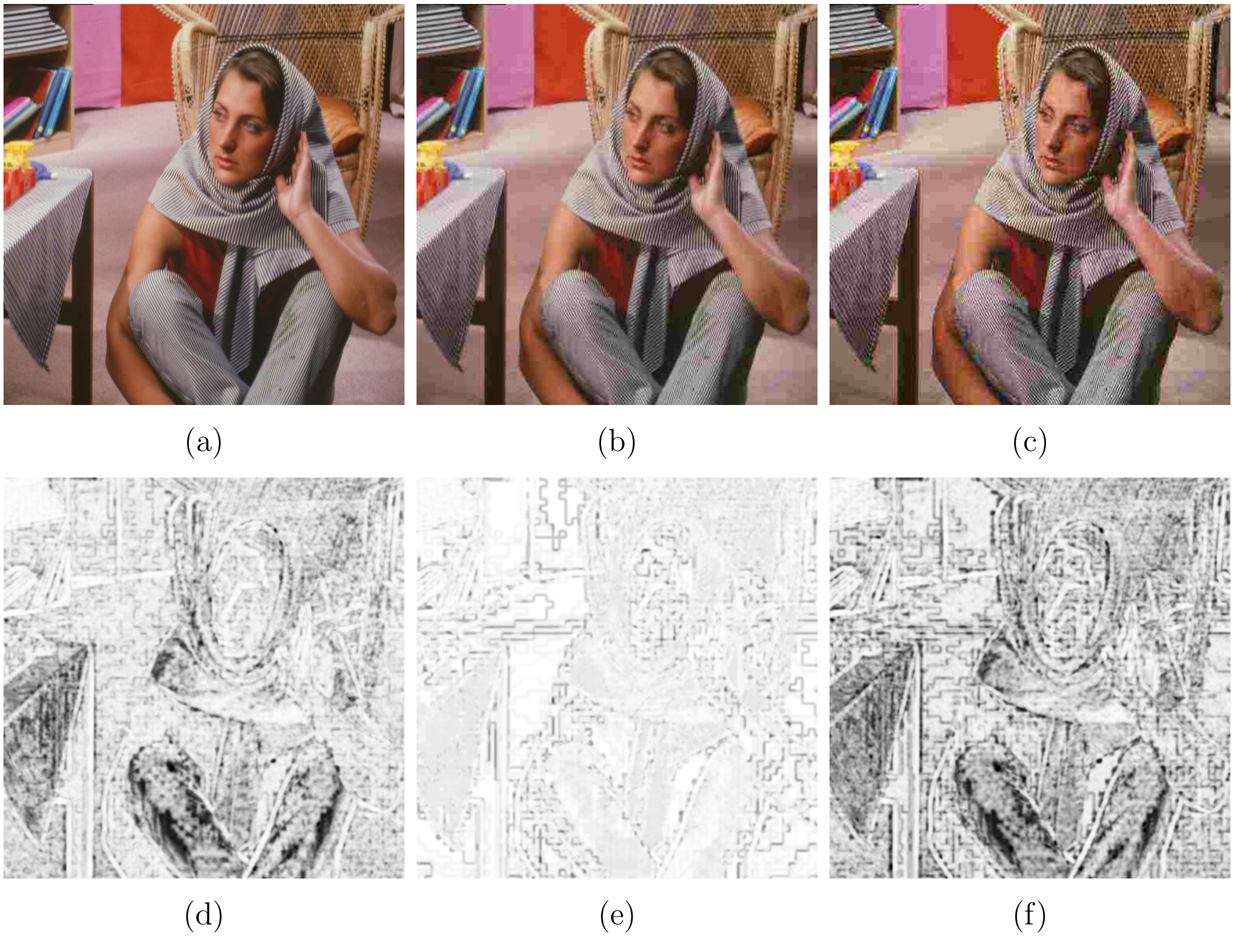
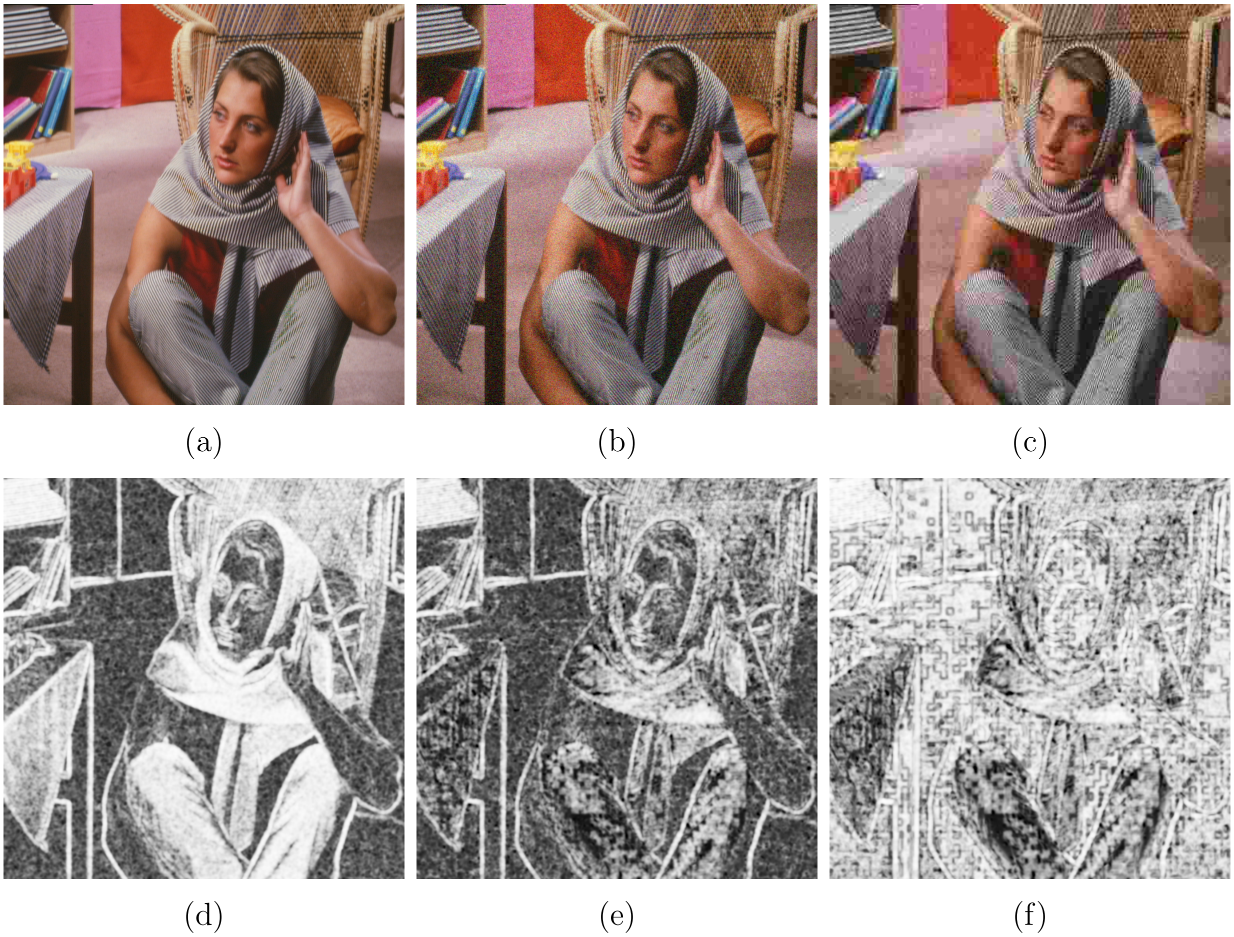


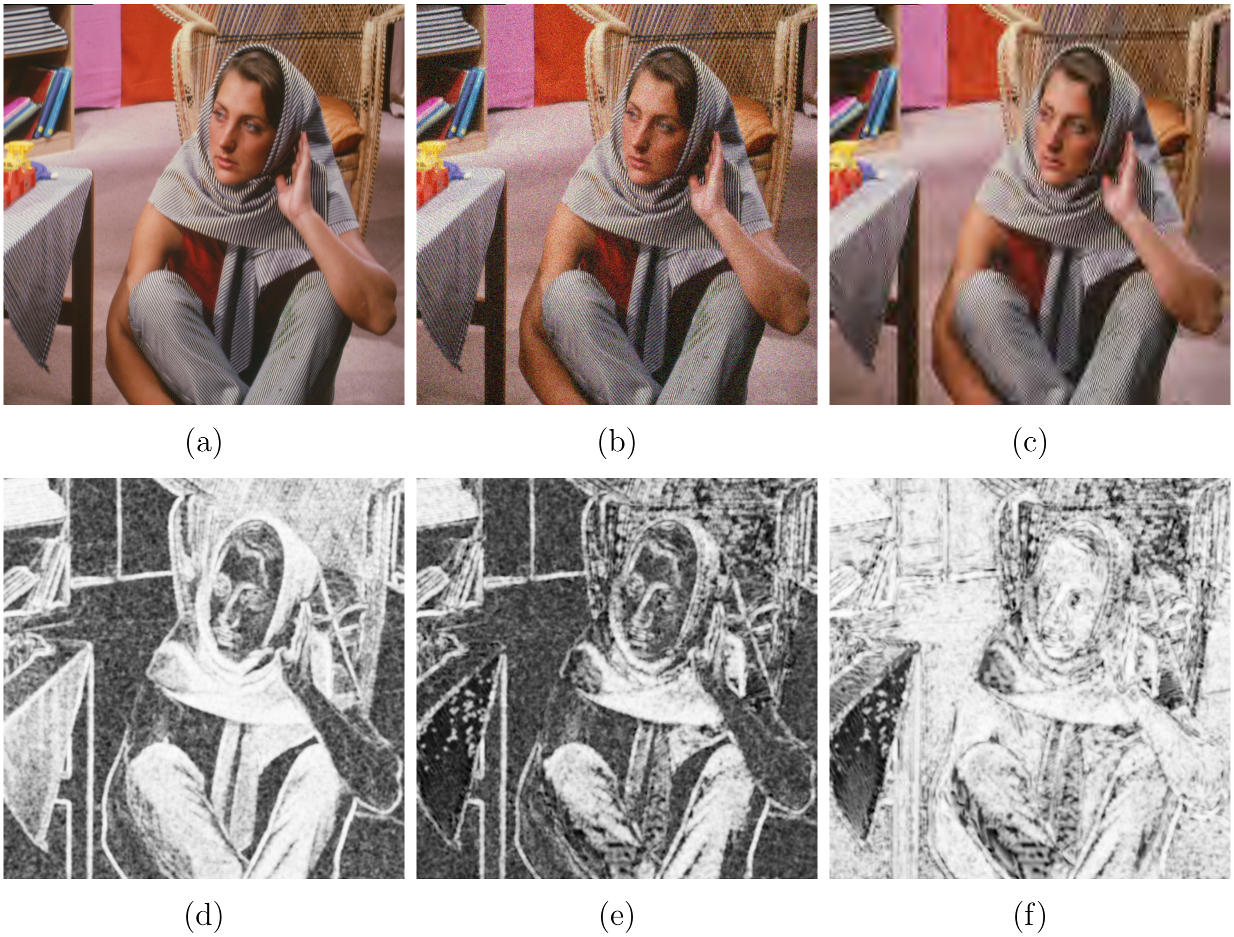
Figure 4.7: Example of the *JPEG-JPEG* distortion combination. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by compressing the image in (a) by using JPEG compression (level 7). (c) Final distorted *Barbara* image obtained by compressing the image in (b) by using JPEG compression (level 11). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).



(a) (b) (c)

(d) (e) (f)

Figure 4.8: Example of the *Noise-JPEG* distortion combination. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by contaminating the image in (a) with white Gaussian noise (level 7). (c) Final distorted *Barbara* image obtained by compressing the image in (b) by using JPEG compression (level 11). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).


(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.9: Example of the *Noise-JPEG2000* distortion combination. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by contaminating the image in (a) with white Gaussian noise (level 7). (c) Final distorted *Barbara* image obtained by compressing the image in (b) by using JPEG2000 compression (level 11). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).

Table 4.5: FSIMc $\mathrm{AS}_{\mathrm{DR}}$, $\mathrm{RS}_{\mathrm{FD}}$, and $\mathrm{AS}_{\mathrm{FD}}$ scores for the examples in Figures 4.5, 4.6, 4.7, 4.8, and 4.9.

Distortion	FSIMc $\mathrm{AS}_{\mathrm{DR}}$		FSIMc $\mathrm{RS}_{\mathrm{FD}}$		FSIMc $\mathrm{AS}_{\mathrm{FD}}$	
Combination	Between	Score	Between	Score	Between	Score
Blur-JPEG	Fig. 4.5 (a) & (b)	0.9495	Fig. 4.5 (b) & (c)	0.9253	Fig. 4.5 (a) & (c)	0.8815
Blur-Noise	Fig. 4.6 (a) & (b)	0.9495	Fig. 4.6 (b) & (c)	0.8646	Fig. 4.6 (a) & (c)	0.8570
JPEG-JPEG	Fig. 4.7 (a) & (b)	0.9525	Fig. 4.7 (b) & (c)	0.9538	Fig. 4.7 (a) & (c)	0.9066
Noise-JPEG	Fig. 4.8 (a) & (b)	0.9311	Fig. 4.8 (b) & (c)	0.8969	Fig. 4.8 (a) & (c)	0.9016
Noise-JPEG2000	Fig. 4.9 (a) & (b)	0.9311	Fig. 4.9 (b) & (c)	0.8907	Fig. 4.9 (a) & (c)	0.9148

5. **Noise-JPEG2000** (Fig. 4.9): The PR *Barbara* image of Fig. 4.9 (a) is distorted at Gaussian noise level-7 to generate the DR image of Fig. 4.9 (b), which is then further distorted at JPEG2000 compression level-11 to generate the FD image of Fig. 4.9 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.9 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.9 (e) and 4.9 (f), respectively.

The FSIMc [14] image-level $\mathrm{AS}_{\mathrm{DR}}$, $\mathrm{AS}_{\mathrm{FD}}$, and $\mathrm{RS}_{\mathrm{FD}}$ scores for the above-mentioned examples are given in Table 4.5. From Figures 4.5 to 4.9 and from Table 4.5, the following observations can be made: 1) For all five distortion combinations, the relative quality maps of the FD images with respect to their DR images are always different when compared to their respective absolute quality maps which are generated between the FD images and their PR images. This visually shows why $\mathrm{RS}_{\mathrm{FD}}$ is not a good predictor of $\mathrm{AS}_{\mathrm{FD}}$. Since FR methods can only compare two images at a time, they consider one of these images as having pristine quality and compute the quality of the other *distorted* image with respect to the *perfect* reference, regardless of whether the reference is itself of degraded quality. Thus, when used in a standalone manner, FR methods are only effective when the PR image is available, which is a major limitation of the FR IQA paradigm. 2) When the relative quality maps of the FD images with respect to their DR images for the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG shown in Figures 4.5 (e), 4.6 (e), and 4.7 (e), respectively, are compared with their respective absolute quality maps (i.e., quality maps of the FD images with respect to their PR images) shown in Figures 4.5 (f), 4.6 (f), and 4.7 (f), it can be seen that the relative quality maps are lighter compared to their respective absolute quality maps. This indicates that for these particular distortion combinations, RS_{FD} over-estimates AS_{FD}. This behavior can also be observed from Table 4.5 for the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG, where the RS_{FD} FSIMc scores are higher than their respective AS_{FD} scores (where a higher FSIMc score is indicative of better quality). 3) However, the opposite can be observed for the cases of Noise-JPEG and Noise-JPEG2000, where their relative quality maps shown in Figures 4.8 (e) and 4.9 (e), respectively, are darker when compared to their respective absolute quality maps shown in Figures 4.8 (f) and 4.9 (f). This indicates that for these particular distortion combinations RS_{FD} under-estimates AS_{FD}. This behavior can also be observed from Table 4.5 for the cases of Noise-JPEG and Noise-JPEG2000, where the RS_{FD} FSIMc scores are lower than the respective AS_{FD} scores (where a smaller FSIMc score is indicative of lower quality). 4) For the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG shown in Figures 4.5, 4.6, and 4.7, respectively, it seems that the absolute quality map of the FD image is a linear combination of the absolute quality map of the DR image and the relative quality map of the FD image, indicating that the individual distortions making up these distortion combinations impact the content rather independently. 5) However, this cannot be said for the cases of Noise-JPEG and Noise-JPEG2000 shown in Figures 4.8 and 4.9, respectively, indicating that the individual distortions making up these distortion combinations impact the content in a joint manner.

While the distorted images and quality maps of Figures 4.5 to 4.9 have given valuable insights into the behavior of multiple distortions, the observations made for these images are only illustrative and may not generalize since the above analysis has only been carried out for one pristine reference image and at a fixed stage-1 and stage-2 distortion level. To study the behavior of multiple distortions in a more general manner, multiple pristine contents need to be analyzed for a wide range of stage-1 and stage-2 distortion levels. To do this, we select four well-known pristine reference images from different IQA datasets that not only depict very different scenes, but also vary widely in terms of spatial information (SI) and colorfulness (CF). For the definition and meaning of SI and CF, refer to Section 2.2.4. The four pristine reference images are: 1) *Ocean* depicted in 4.10 (a) is an outdoor image showing a natural landscape and has low SI and CF. 2) *Buildings* depicted in 4.10 (b) is an outdoor image showing various buildings with a natural background and has high

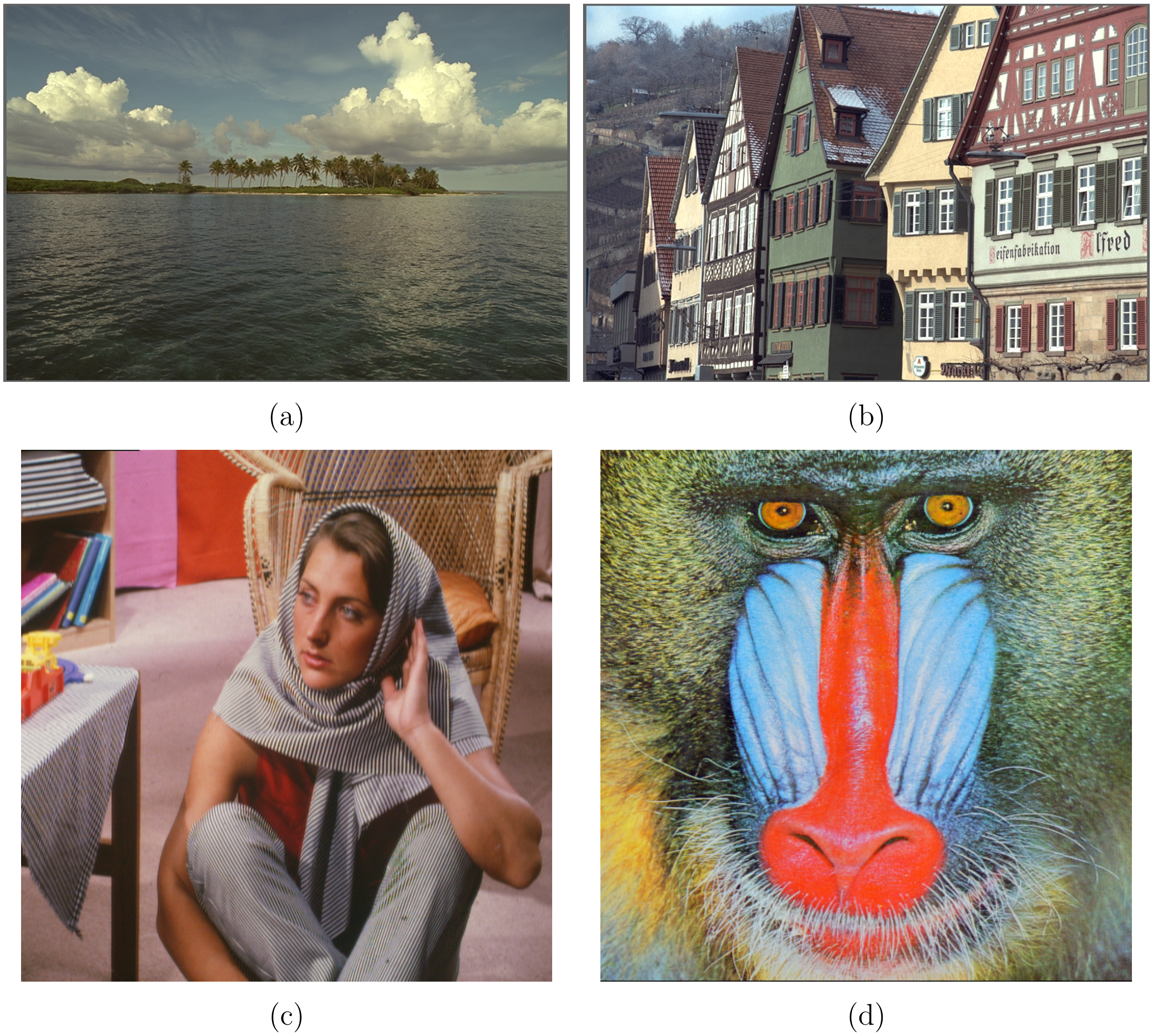
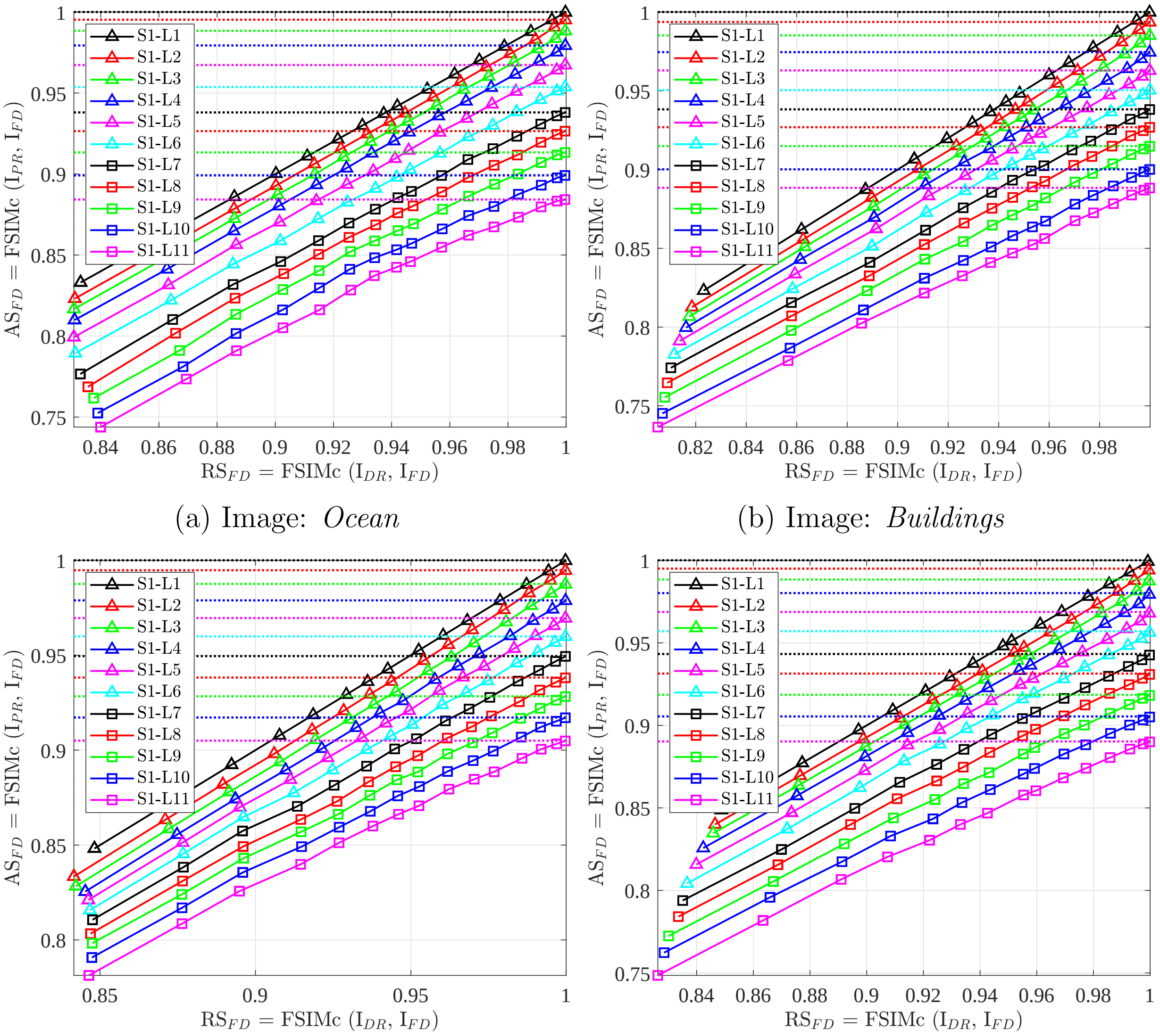


Figure 4.10: Pristine reference images being used for multiple distortions behavior analysis. (a) *Ocean* (Low spatial information (SI) and colorfulness (CF)). (b) *Buildings* (High SI and low CF). (c) *Barbara* (Mid-level SI and CF). (d) *Mandrill* (High SI and CF).

SI but low CF. 3) *Barbara* depicted in 4.10 (c) is an indoor image of a woman surrounded by furniture and has mid-level SI and CF. The earlier analysis carried out on the basis of Figures 4.5 to 4.9 and Table 4.5 also utilized this pristine reference image. 4) *Mandrill* depicted in 4.10 (d) is the close-up image of an animal's face and has high SI and CF. Utilizing images with a wide range of SI and CF would allow for observing the impact of different multiple distortions in a comprehensive manner by factoring in any masking effect that the content itself might have.

To create distorted images, we determine 17 content adaptive distortion thresholds corresponding to target quality levels depicted in Table 3.2 for the distortions of Gaussian noise, Gaussian blur, JPEG compression, and JPEG2000 compression, by following the same procedure as described earlier in Section 3.3.2. Among these 17 distortion levels, level-1 corresponds to minimum distortion whereas level-17 corresponds to highest distortion. Next, for the distortion types of Gaussian noise, Gaussian blur, and JPEG compression, we use their first 11 distortion levels to distort each of the four PR images into 11 stage-1 distorted or DR images that cover the top half of the quality spectrum. Multiply distorted images belonging to five distortion combinations are then created by distorting: 1) each blur DR image at 17 levels of JPEG compression to create 187 Blur-JPEG FD images for each PR, 2) each blur DR image at 17 levels of Gaussian noise to create 187 Blur-Noise FD images for each PR, 3) each JPEG compressed DR image at 17 levels of JPEG compression to create 187 JPEG-JPEG FD images for each PR, 4) each noise DR image at 17 levels of JPEG compression to create 187 Noise-JPEG FD images for each PR, and 5) each noise DR image at 17 levels of JPEG2000 compression to create 187 Noise-JPEG2000 FD images for each PR. Thus, for each PR image and for each distortion combination, there are 11 DR images and thus 11 $\mathrm{AS}_{\mathrm{DR}}$ quality scores. For each DR image, there are 17 FD images leading to a total of 187 FD images for each PR image in each distortion combination and thus there are 187 $\mathrm{RS}_{\mathrm{FD}}$ and $\mathrm{AS}_{\mathrm{FD}}$ scores (for each $\mathrm{AS}_{\mathrm{DR}}$ score, there are 17 $\mathrm{RS}_{\mathrm{FD}}$ and $\mathrm{AS}_{\mathrm{FD}}$ scores). In the analysis that follows, we use FSIMc [14] to compute the $\mathrm{AS}_{\mathrm{DR}}$, $\mathrm{RS}_{\mathrm{FD}}$, and $\mathrm{AS}_{\mathrm{FD}}$ scores.

To thoroughly analyze the behavior of different distortion combinations, we plot $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ scores for each of the four PR images and for each distortion combination. These plots for the Blur-JPEG distortion combination are shown in Fig. 4.11. For each PR image, there are 11 different curves, each of which corresponds to one of the 11 $\mathrm{AS}_{\mathrm{DR}}$ scores. The 17 different points on each curve represent the $\mathrm{RS}_{\mathrm{FD}}$ and $\mathrm{AS}_{\mathrm{FD}}$ scores corresponding to one particular $\mathrm{AS}_{\mathrm{DR}}$ score. The dotted lines in Fig. 4.11 represent the $\mathrm{AS}_{\mathrm{DR}}$ scores for the



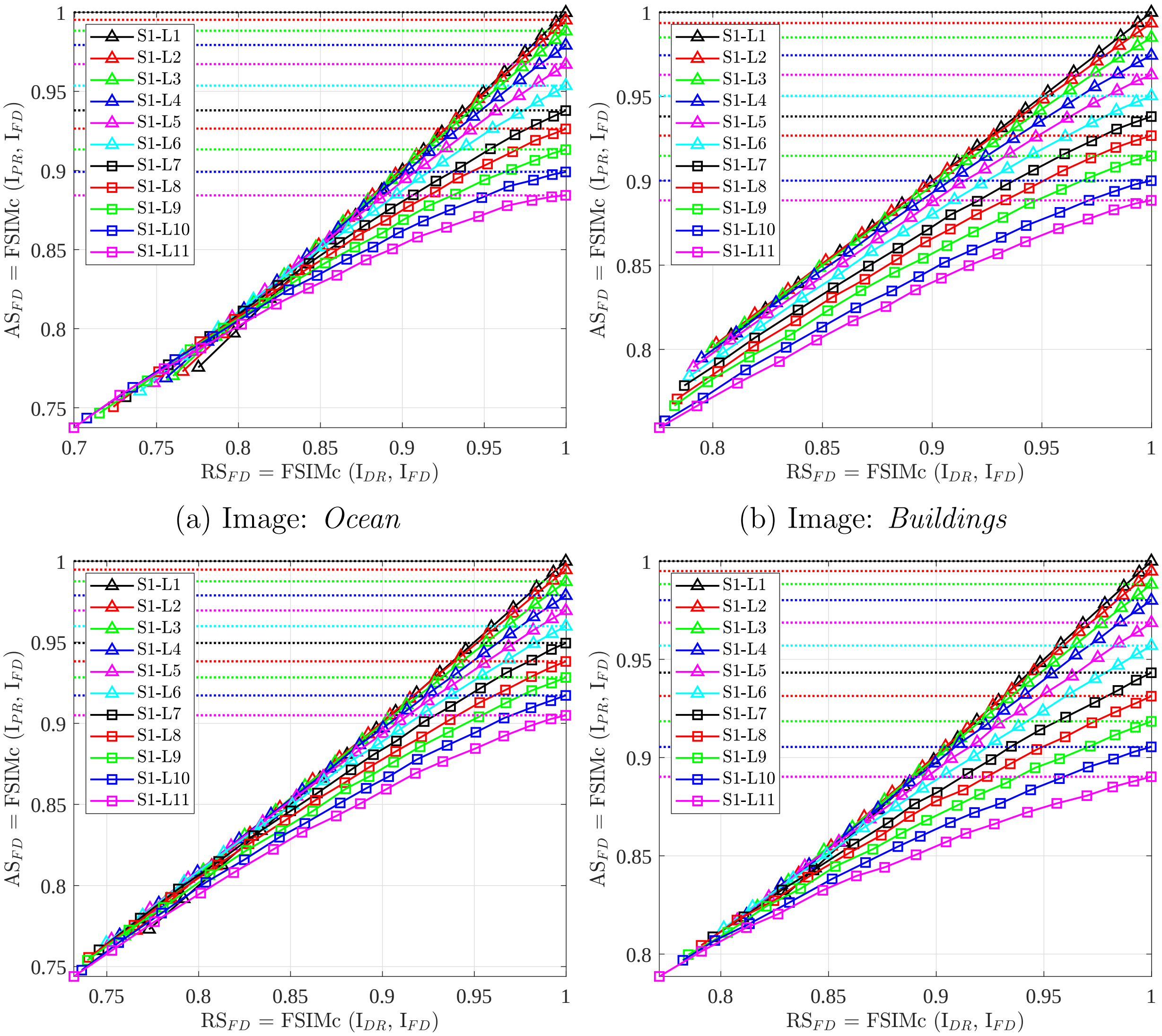
(a) Image: *Ocean*

(b) Image: *Buildings*

(c) Image: *Barbara*

(d) Image: *Mandrill*

Figure 4.11: Blur-JPEG plots of AS_{FD} versus RS_{FD} for every stage-1 (AS_{DR}) distortion level corresponding to four pristine reference images. Dotted lines represent AS_{DR} scores.



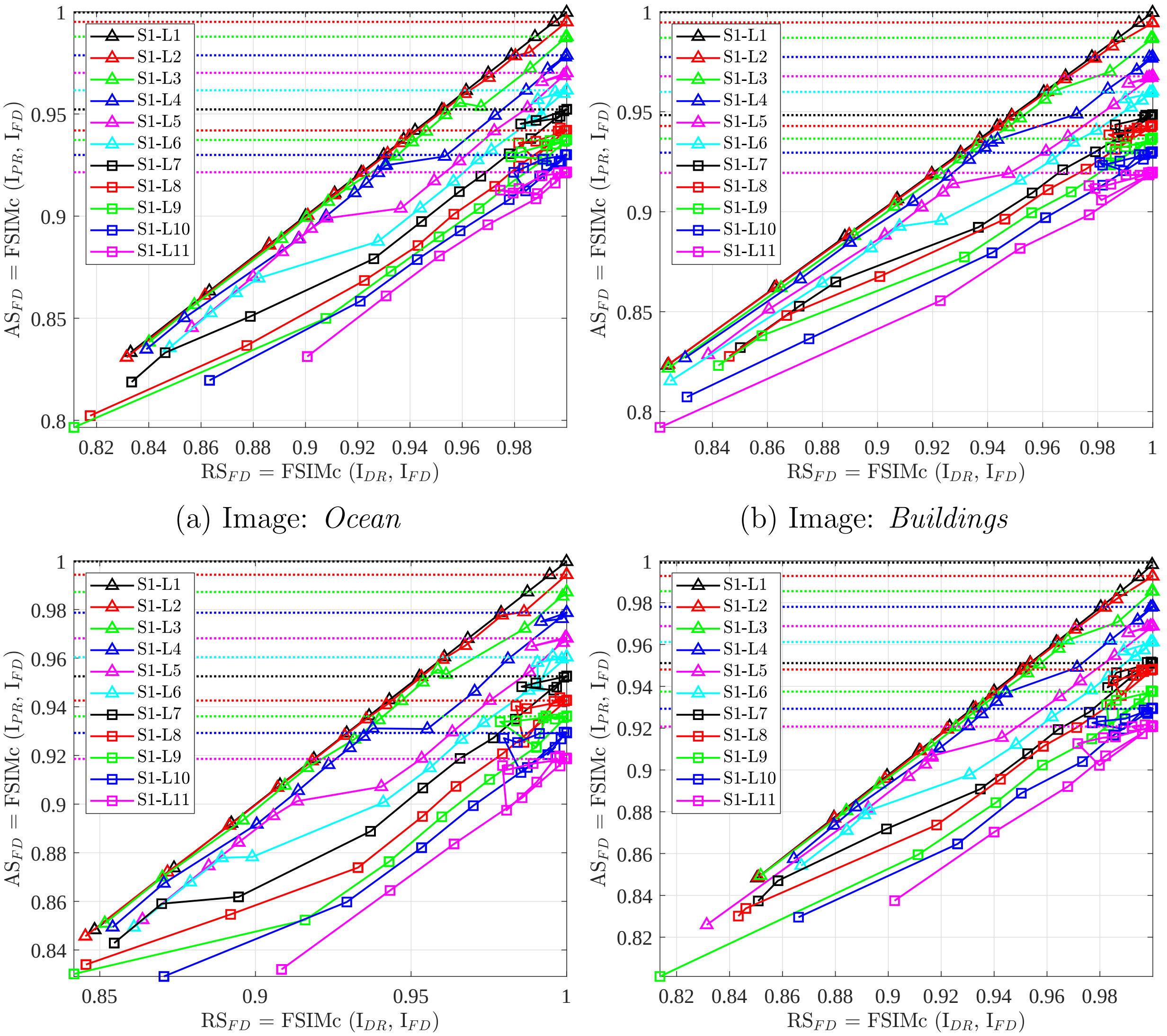
(a) Image: *Ocean*

(b) Image: *Buildings*

(c) Image: *Barbara*

(d) Image: *Mandrill*

Figure 4.12: Blur-Noise plots of AS_{FD} versus RS_{FD} for every stage-1 (AS_{DR}) distortion level corresponding to four pristine reference images. Dotted lines represent AS_{DR} scores.



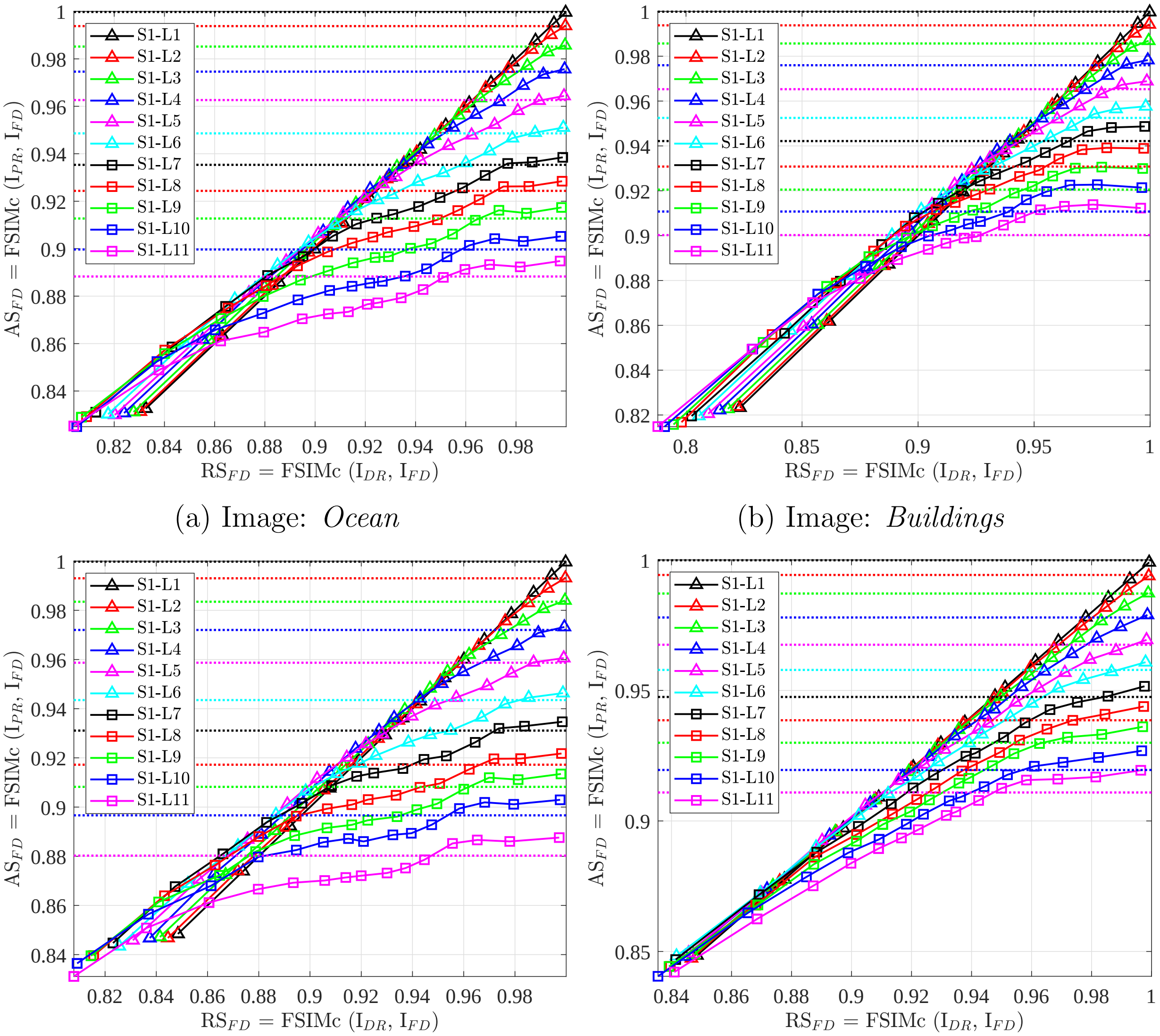
(a) Image: *Ocean*

(b) Image: *Buildings*

(c) Image: *Barbara*

(d) Image: *Mandrill*

Figure 4.13: JPEG-JPEG plots of AS_{FD} versus RS_{FD} for every stage-1 (AS_{DR}) distortion level corresponding to four pristine reference images. Dotted lines represent AS_{DR} scores.



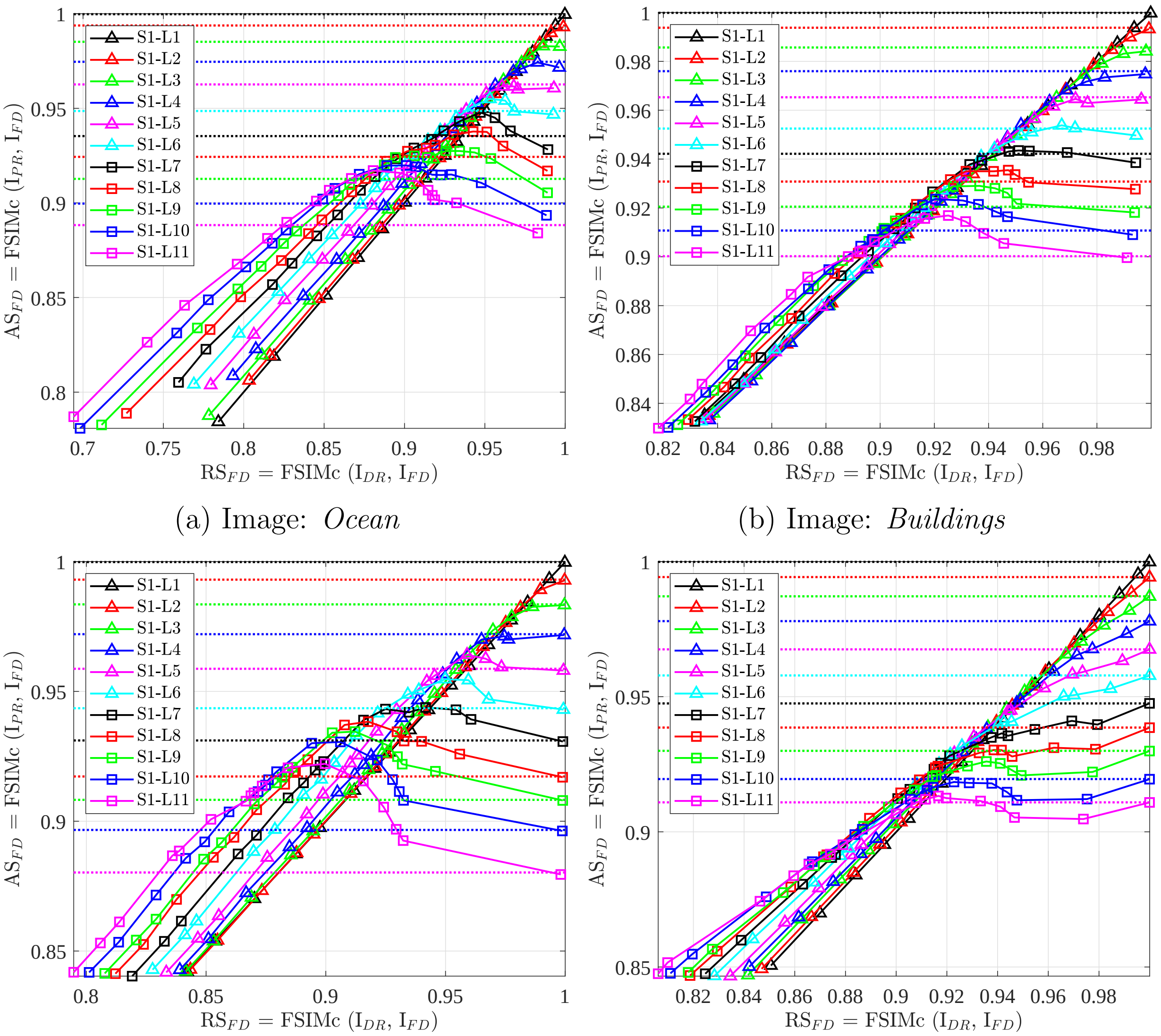
(a) Image: *Ocean*

(b) Image: *Buildings*

(c) Image: *Barbara*

(d) Image: *Mandrill*

Figure 4.14: Noise-JPEG plots of AS_{FD} versus RS_{FD} for every stage-1 (AS_{DR}) distortion level corresponding to four pristine reference images. Dotted lines represent AS_{DR} scores.



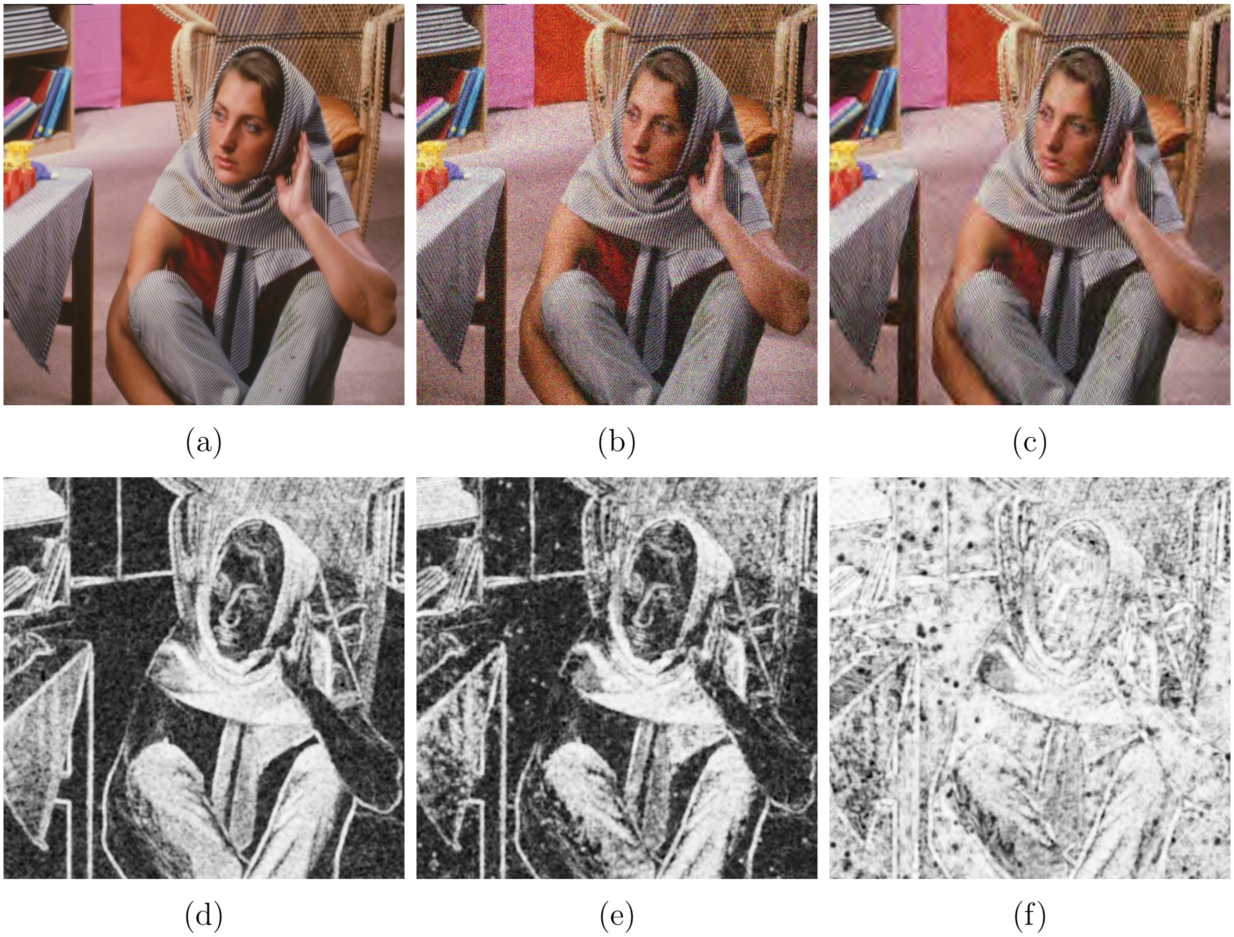
(a) Image: *Ocean*

(b) Image: *Buildings*

(c) Image: *Barbara*

(d) Image: *Mandrill*

Figure 4.15: Noise-JPEG2000 plots of $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ for every stage-1 ($\mathrm{AS}_{\mathrm{DR}}$) distortion level corresponding to four pristine reference images. Dotted lines represent $\mathrm{AS}_{\mathrm{DR}}$ scores.


(a) (b) (c)

(d) (e) (f)

Figure 4.16: Another example of the *Noise-JPEG2000* distortion combination where the final distorted image quality is better than the degraded reference. (a) Pristine reference *Barbara* image. (b) Degraded reference *Barbara* image obtained by contaminating the image in (a) with white Gaussian noise (level 11). (c) Final distorted *Barbara* image obtained by compressing the image in (b) by using JPEG2000 compression (level 6). (d) Absolute quality map of the degraded reference image in (b) with respect to the pristine reference image in (a). (e) Relative quality map of the final distorted image in (c) with respect to the degraded reference image in (b). (f) Absolute quality map of the final distorted image in (c) with respect to the pristine reference image in (a).

11 different DR images. Plots for the Blur-Noise, JPEG-JPEG, Noise-JPEG, and Noise-JPEG2000 distortion combinations are similarly constructed and are depicted in Figures 4.12, 4.13, 4.14, and 4.15, respectively. The following observations can be made:

- For all five distortion combinations, Figures 4.11 to 4.15 show that:
  - At minimum stage-1 distortion, i.e., at $\text{AS}_{\text{DR}} \simeq 1$, $\text{AS}_{\text{FD}} \simeq \text{RS}_{\text{FD}}$. This can be seen by observing Stage-1 Level-1 (S1-L1) curves in all these figures. The DR image at S1-L1 is as good as the PR image because S1-L1 has almost no distortion.
  - As the stage-1 distortion increases, the prediction from RS_{FD} to AS_{FD} becomes more unreliable.
  - It can also be seen that $\text{AS}_{\text{FD}} \simeq \text{AS}_{\text{DR}}$ at minimum stage-2 distortion (Stage-2 Level-1). This is not surprising since at S2-L1, stage-2 is not adding any further distortion to the DR image.
- The plots of the Blur-JPEG and JPEG-JPEG distortion combinations depicted in Figures 4.11 and 4.13, respectively, show that:
  - As stage-1 distortion increases, the curves move away from the S1-L1 curve. All such curves are below the S1-L1 curve, meaning that RS_{FD} is assigning relatively higher quality scores to the FD images compared to the absolute AS_{FD} scores.
  - The curve representing Stage-1 Level-11 (S1-L11) distortion, which is the maximum stage-1 distortion, is furthest from the S1-L1 curve.
  - The AS_{FD} versus RS_{FD} curves for all AS_{DR} scores depict a linear behavior and are almost parallel to each other. This behavior is especially true for the Blur-JPEG case, but can also be roughly seen in the JPEG-JPEG case. This indicates that the constituent distortions in these combinations behave independent of each other and do not lead to complex joint effects.
- The plots for the Blur-Noise distortion combination depicted in Fig. 4.12 show that they also follow the behavior discussed for the Blur-JPEG and JPEG-JPEG

combinations above, with one major difference. As the level of stage-2 distortions increases, the AS_{FD} versus RS_{FD} curves for all AS_{DR} scores begin to converge. This shows that for the Blur-Noise case, as the magnitude of the stage-2 distortion, i.e., Gaussian noise, increases, it overshadows the stage-1 distortion (Gaussian blur), to become the dominant distortion.

- The plots for the Noise-JPEG and Noise-JPEG2000 distortion combinations depicted in Figures 4.14 and 4.15, respectively, show that:

  - The AS_{FD} versus RS_{FD} curves begin exhibiting nonlinear behavior as stage-1 distortion level increases (i.e., as AS_{DR} scores decrease) and are also not parallel to each other. Such behavior is especially true for the Noise-JPEG2000 case. This points to significant joint effects of the two constituent distortions in the combination.

  - It can be seen that as stage-1 distortion levels increase, some portion of the AS_{FD} versus RS_{FD} curves go above their respective AS_{DR} scores, i.e., an overshoot takes place. This behavior is most apparent in the low to mid-level stage-2 distortion levels segment of the AS_{FD} versus RS_{FD} curves corresponding to mid to high level stage-1 distortion levels. While this behavior can be seen in both the Noise-JPEG and Noise-JPEG2000 plots, it is much more pronounced in the latter case. It can also be observed from Figures 4.14 and 4.15, that once the curves drop below their respective AS_{DR} scores, they follow a rather linear pattern. The overshoot phenomenon above AS_{DR} score levels for some AS_{FD} versus RS_{FD} curves indicates that the corresponding FD images have better perceptual quality than their respective DR images, which is a surprising finding. JPEG compression is known to cause blurring and blocking artifacts, while JPEG2000 compression is known to cause blurring and ringing artifacts. When these compression techniques are applied to noisy degraded references, the compression induced blurring has a denoising effect. For certain combinations of stage-1 noise and stage-2 JPEG/JPEG2000 compression levels, this denoising effect may be such that it reduces the amount of noise while producing perceptually low amount of compression artifacts, thereby making the perceptual quality of the FD image better than that of the DR image from which it is derived. However, when the amount of compression becomes too high, the compression artifacts overrun the benefit of noise removal and the curves dip below the AS_{DR} values.

– We visually demonstrate the above point in Fig. 4.16, where the PR *Barbara* image of Fig. 4.16 (a) is distorted at Gaussian noise level-11 to generate the DR image of Fig. 4.16 (b), which is then further distorted at JPEG2000 compression level-6 to generate the FD image of Fig. 4.16 (c). The quality map of the DR image with respect to the PR image is shown in Fig. 4.16 (d), while the quality maps of the FD image with respect to the DR and PR images are shown in Figures 4.16 (e) and 4.16 (f), respectively. The AS_{DR} FSIMc score of the DR image is 0.8802, while the RS_{FD} and AS_{FD} FSIMc scores of the FD image are 0.8997 and 0.9221, respectively. This clearly shows that the stage-1 and stage-2 distortion levels for this example are such that the FD image has better perceptual quality than the DR image. The absolute quality map of the FD image, shown in Fig. 4.16 (f) further attests to this denoising effect of compression following noise. The FD image in this example can also be located on the AS_{FD} versus RS_{FD} curve corresponding to the Stage-1 Level-11 (S1-L11) in Fig. 4.15 (c), where it is the sixth point from the right on the curve, and lies above the corresponding AS_{DR} score.

– For the Blur-JPEG and JPEG-JPEG cases, we noted earlier that the AS_{FD} versus RS_{FD} curves follow a linear behavior and a curve corresponding to a higher stage-1 distortion is always below a curve corresponding to a lower stage-1 distortion for all stage-2 distortion levels. However, in the case of Noise-JPEG2000 and to some extent also for Noise-JPEG, we note that the relationship between AS_{FD} and RS_{FD} is not linear and a curve corresponding to a higher stage-1 distortion is not always below a curve corresponding to a lower stage-1 distortion. It can be observed from Figures 4.14 and 4.15 for both Noise-JPEG and Noise-JPEG2000 (especially for the latter) that at low stage-2 distortions (towards the right side of the plots), a curve corresponding to a higher stage-1 distortion is below a curve corresponding to a lower stage-1 distortion, however, as stage-

2 distortion levels increase, a point comes when a crossover takes place. This is again due to the denoising effect that compression has on noisy degraded references. For a certain range of stage-2 distortion levels, the application of compression actually improves the perceptual quality of the noisy degraded reference images. Images corresponding to higher stage-1 distortion levels have more noise and they benefit more from the additional denoising effect of higher stage-2 distortion levels which leads to their respective $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curves crossing above curves that correspond to lower stage-1 distortion levels.

- The above discussion has revealed that among the five distortion combinations being analyzed, the constituent distortions in Noise-JPEG and Noise-JPEG2000 lead to complex joint effects. Thus, creating DR IQA models for these combinations will be most challenging, as we shall see later in this chapter.

- While we have presented the above analysis by using only four PR images, similar analysis on hundreds of PR images was carried out and it led to similar observations, indicating that the above-mentioned observations can be generalized.

4.3.2 DR IQA Framework

It was discussed at the beginning of Section 4.3, that three images exist at different locations of a two-stage distortion pipeline. These include the PR image at the original source, the DR image at the output of the first distortion stage, and the FD image at the output of the second distortion stage. For the purposes of DR IQA, the DR and FD images are considered available for the task of quality assessment of the FD image. Based on the availability of the PR image, the following two DR IQA frameworks are possible.

DR IQA Framework: Scenario 1

In this first scenario, we assume that the PR image is available early on in the media distribution system and it is possible to ascertain the quality of the DR image with respect to the PR image by using an FR method. We have already defined such an absolute

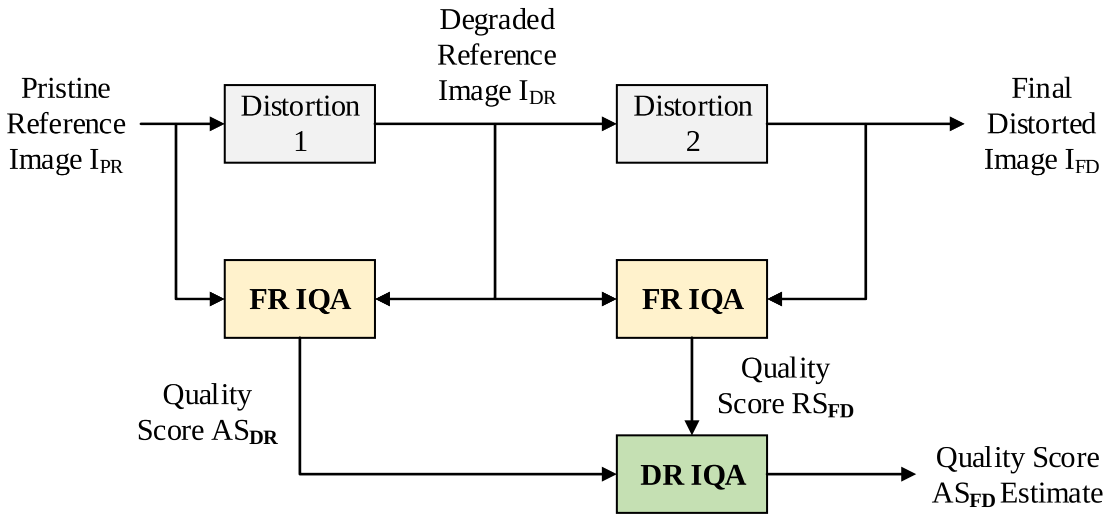


Figure 4.17: General framework of Degraded Reference (DR) IQA Scenario 1 models.

quality score as $\mathrm{AS_{DR}}$ in Equation 4.2. It is important to note that Scenario 1 is practically applicable only when $\mathrm{AS_{DR}}$ is pre-computed at the first distortion stage that leads to the DR image and is then transmitted with the DR image to the second distortion stage. This is because we cannot assume the availability of the PR image at subsequent stages of the media delivery system. For example, Scenario 1 may be implemented in practical image distribution systems that involve two compression stages, though additional protocols need to be used to transfer $\mathrm{AS_{DR}}$ scores as side information, thus requiring minor changes to the distribution system. Since the DR and FD images are available at the second distortion stage, an FR method can be used to determine the relative quality of the FD image with respect to the DR image, which we defined as $\mathrm{RS_{FD}}$ in Equation 4.4. We had also defined $\mathrm{AS_{FD}}$ in Equation 4.3 as the absolute FR quality score between the PR and FD images. In this first scenario, the goal of DR IQA is to predict $\mathrm{AS_{FD}}$ by using both $\mathrm{AS_{DR}}$ and $\mathrm{RS_{FD}}$, i.e.,

$$\widehat{\mathrm{AS_{FD}}} = f(\mathrm{AS_{DR}}, \mathrm{RS_{FD}}), \tag{4.5}$$

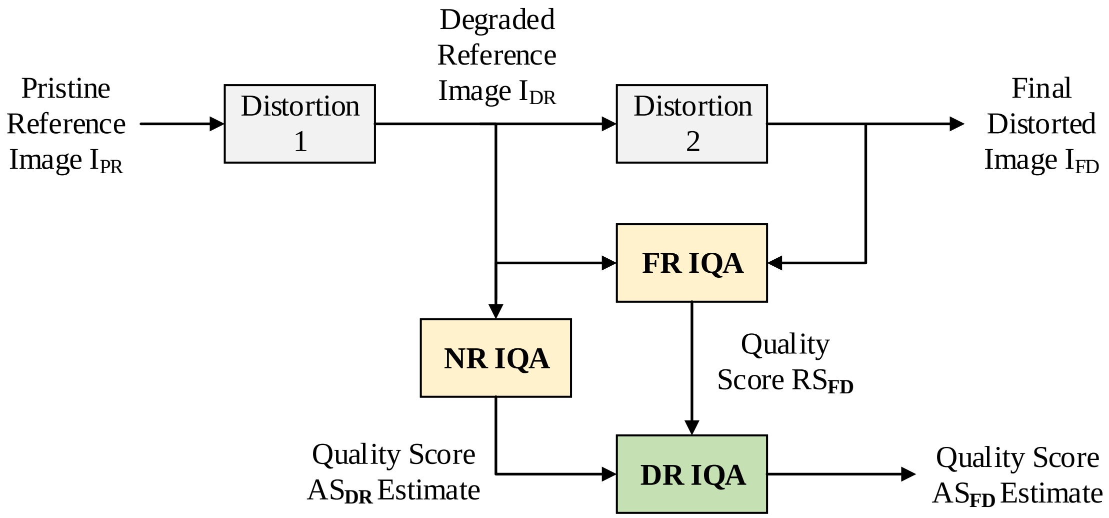


Figure 4.18: General framework of the practical DR IQA Scenario 2 models.

where $\widehat{\mathrm{AS}_{\mathrm{FD}}}$ is the estimated or predicted value of $\mathrm{AS}_{\mathrm{FD}}$. The general framework of the Scenario 1 based DR IQA models is shown in Fig. 4.17.

DR IQA Framework: Scenario 2

As discussed earlier in this chapter, in practical media distribution systems, the PR images are not accessible. In fact, we had defined DR IQA to be a paradigm that ascertains the quality of the FD image by only utilizing the DR image as it does not have access to the PR image. Thus, in this more practical second scenario, the FR computed score of the DR image with respect to PR, i.e., $\mathrm{AS}_{\mathrm{DR}}$ is not available. However, since the DR image is available, its quality may be estimated by using an NR IQA algorithm. i.e.,

$$\widehat{\mathrm{AS}_{\mathrm{DR}}} = \mathrm{NR}(\mathrm{I}_{\mathrm{DR}}), \tag{4.6}$$

where I_{DR} is the degraded reference image, NR is a trusted NR IQA method, and $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ is the quality of the DR image as estimated by the NR method. The relative quality score of the FD image with respect to the DR image, i.e., $\mathrm{RS}_{\mathrm{FD}}$ can still be determined by using an FR method. Thus, in this second scenario, the goal of DR IQA is to predict $\mathrm{AS}_{\mathrm{FD}}$ by using the NR predicted $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$, i.e.,

$$\widetilde{\mathrm{AS}_{\mathrm{FD}}} = f(\widehat{\mathrm{AS}_{\mathrm{DR}}}, \mathrm{RS}_{\mathrm{FD}}), \tag{4.7}$$

where $\widetilde{\mathrm{AS}_{\mathrm{FD}}}$ is the estimated or predicted value of $\mathrm{AS}_{\mathrm{FD}}$ when DR IQA uses the NR-predicted value of $\mathrm{AS}_{\mathrm{DR}}$, i.e., $\widehat{\mathrm{AS}_{\mathrm{DR}}}$. The general framework of the Scenario 2 based DR IQA models is shown in Fig. 4.18.

While the application of Scenario 1 based DR IQA framework requires making minor changes to the media distribution system, no such changes are required for applying Scenario 2 to such systems. All that is required is to add probes at the second distortion stage to sample the DR and FD images. Such ease of implementation makes the Scenario 2 based DR IQA framework readily applicable to pre-existing media distribution systems.

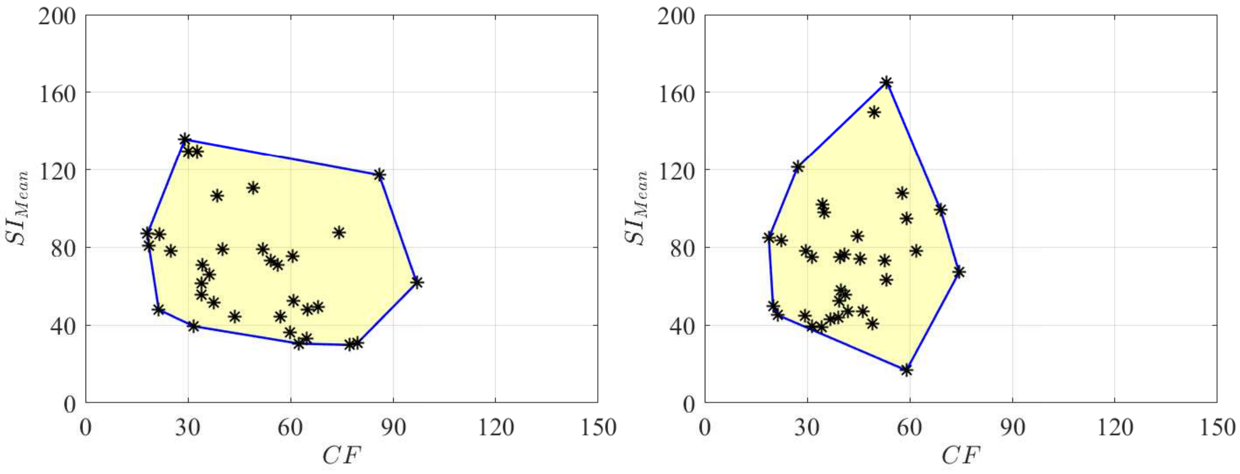
We will use both Scenario 1 and Scenario 2 based DR IQA frameworks to develop DR IQA models in Section 4.5.

4.4 DR IQA Databases Construction

The Waterloo Exp-II database, constructed earlier in Chapter 3 (Section 3.3) has 3,570 pristine reference images, 39,270 singly distorted images each for noise, blur and JPEG compression, and 667,590 multiply distorted images each for the distortion combinations of Blur-JPEG, Blur-Noise, JPEG-JPEG, Noise-JPEG, and Noise-JPEG2000. The singly distorted and multiply distorted images in this dataset are essentially the degraded references and final distorted images in a two-stage distortion pipeline. Therefore, the Waterloo Exp-II database can be used for the purpose of DR IQA model development and testing. While the enormous size of this dataset proves highly beneficial in the development of DNN based models, as we discussed in Chapter 3, other machine learning tools such as SVR [149] running on regular CPU based computers take a lot of time to learn models when they use such a large amount of training data. Instead of extracting smaller subsets from the Waterloo Exp-II dataset, we construct two new relatively small-scale datasets for the purpose of DR IQA model development which we shall release publicly to the research community at large as these will aid in learning models using tools such as SVR. We call these databases DR IQA database Version 1 (V1) and DR IQA database Version 2 (V2). The purpose for developing two such datasets is that one can be used for model training, while the other can be used for model validation if a machine learning based training process is used. The much larger Waterloo Exp-II dataset can then be used for model testing. We ensure that DR IQA database V1, DR IQA database V2, and the Waterloo Exp-II dataset do not have any overlap in content, thereby providing completely disjoint sets of data for model training, validation, and testing. To ensure that the dataset construction process for the two new DR IQA datasets is exactly the same as the Waterloo Exp-II database so that the latter can be used for model testing, we construct these datasets by following the same procedure as earlier described in Section 3.3 for the Waterloo Exp-II database. We only briefly describe the construction of the DR IQA datasets in the following two subsections. For a detailed description of the dataset construction mechanism, refer to Section 3.3.

4.4.1 Reference Content

A total of 68 pristine quality reference images were taken from the following sources: IQA databases CSIQ [26, 63], IVC [30], LIVE R2 [24, 42], TID2013 [19, 62], Toyoma [29] and some pristine images were extracted from raw videos available at CDVL [226]. These images were divided into two disjoint groups of 34 images each, with one group forming the pristine image set of DR IQA database V1 and the other forming the pristine image set of DR IQA database V2. To quantitatively describe the reference image content, we plot it in the 2D SI versus CF space [88], as was done earlier in Chapters 2 and 3. For a detailed description of the SI versus CF space and its use for reference content analysis, refer to Sections 2.2.4 and 3.3.1. The SI versus CF plots of the reference image content of DR IQA databases V1 and V2 are shown in Figures 4.19 (a) and 4.19 (b), respectively. The plots in Fig. 4.19 can be directly compared with those of nine subject-rated IQA datasets given in Fig. 2.1, as all of them have the same scale. It can be seen that the reference content coverage in the DR IQA databases V1 and V2 is at a level similar to what is found in most IQA datasets. A comparison Fig. 4.19 with Fig. 3.1 again demonstrates the enormity of reference content in the Waterloo Exp-II database.



(a) DR IQA Database V1

(b) DR IQA Database V2

Figure 4.19: Spatial Information (SI_{Mean}) versus Colorfulness (CF) plots of the reference images of DR IQA databases V1 and V2. The blue lines represent the convex hull.

4.4.2 Distorted Content and Quality Annotation

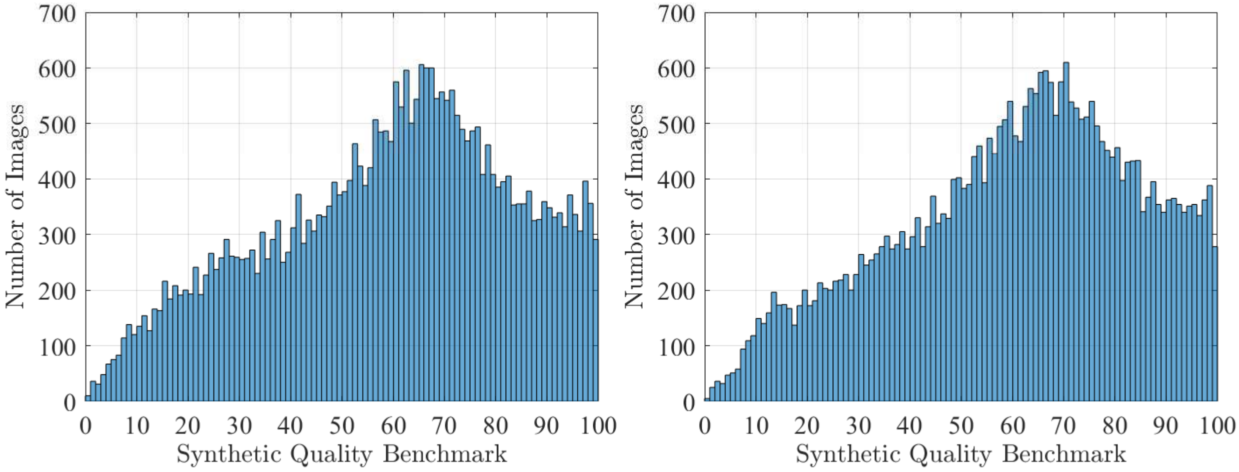
It was discussed in Chapters 2 and 3 that most IQA datasets do not cover the entire quality spectrum adequately which is because they use fixed distortion parameters to create simulated distorted content thereby neglecting the masking effects of the content itself. To address this issue, we had used content adaptive distortion thresholds while creating the Waterloo Exp-II database, as described in Section 3.3.2. We take the same approach to generate the distorted content for the two DR IQA databases. First, we use a wide range of distortion parameters to create 15,000 Gaussian noise images, 10,000 Gaussian blur images, 101 JPEG compressed images, and 20,000 JPEG2000 compressed images for each pristine reference image of the DR IQA databases. Next, we compute the FR SSIMplus [61] scores for all these distorted images and determine distortion parameters for each distortion type that lead to SSIMplus scores closest to the target scores for 17 distortion levels as listed in Table 3.2. This leads to content adaptive distortion parameters for each pristine reference image for each of the four base distortion types.

To align the DR IQA databases with the Waterloo Exploration-II database, we include the same distortion types in the former as in the latter. Details about the distortion types in the Waterloo Exploration-II database, along with their limitations, have been provided in Section 3.3.2. Thus, we include singly distorted degraded reference images in the DR IQA databases belonging to the three distortion categories of Gaussian white noise, Gaussian blur, and JPEG compression. We also want to restrict the degraded references in the *fair* to *excellent* perceptual quality range to better mimic practical media distribution systems. To accomplish this, we distort the pristine reference images with the above-mentioned distortion types by using their respective content adaptive distortion parameters belonging to only distortion Levels 1 to 11 (see Table 3.2). This leads to the creation of 374 degraded references each for noise, blur and JPEG compression in each of the two DR IQA databases. Next, we create multiply distorted images or final distorted images by distorting the DR images and create five distortion combinations. Specifically, for each DR IQA database, each blur DR image is distorted at 17 levels of JPEG compression and 17 levels of Gaussian noise, by using the parent PR image's content adaptive distortion parameters, to create 6,358 Blur-JPEG and 6,358 Blur-Noise FD images. Similarly 6,358 JPEG-JPEG FD images are generated by distorting each JPEG DR image at 17 levels of JPEG compression. Finally, we distort each noise DR image at 17 levels of JPEG compression and 17 levels of JPEG2000 compression to generate 6,358 Noise-JPEG and 6,358 Noise-JPEG2000 FD images for each DR IQA database. By using all 17 levels of distortion to create the FD images, we ensure that they belong to the entire *bad* to *excellent* quality spectrum. Table 4.6 outlines the composition of the two DR IQA databases.

A major limitation of contemporary IQA datasets of multiply distorted content, such as LIVE MD [31], MDIVL [34], MDID [33], MDID2013 [32], and LIVE WCmp [2], is that they do not offer sufficient levels of distortion per distortion stage. For example, LIVE MD [31] offers only three levels of distortion per distortion stage. Such sparse nature of these datasets makes it difficult to analyze how different constituent distortions in a

Table 4.6: Composition of DR IQA databases V1 and V2.

Reference Images in each Database (Pristine Quality)	Stage-1 Distorted Images in each Database (Singly Distorted DRs)		Stage-2 Distorted Images in each Database (Multiply Distorted FDs)	
Number of Images	Distortion	Number of Images	Distortion Combination	Number of Images
34	Blur	374	Blur-JPEG	6,358
			Blur-Noise	6,358
	JPEG	374	JPEG-JPEG	6,358
	Noise	374	Noise-JPEG	6,358
			Noise-JP2K	6,358
	Total	1,122	Total	31,790
	Overall 32,912 Distorted Images in each Database			



(a) DR IQA Database V1

(b) DR IQA Database V2

Figure 4.20: SQB histograms of the DR IQA databases V1 and V2.

multiply distorted image are jointly effecting the image content. By having 11 stage-1 and 17 stage-2 distortion levels, we have ensured that both the DR IQA databases (and also the Waterloo Exp-II database) have adequate density of distortion levels per distortion stage. This has allowed us to comprehensively study the behavior of different multiple distortion combinations, as we have already demonstrated in Section 4.3.1 where the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ plots of Figures 4.11 to 4.15 proved invaluable in our multiple distortions behavior analysis. All the PR, DR, and FD images used to generate these plots are part of either DR IQA database V1 or DR IQA database V2. Contemporary multiply distorted IQA datasets mentioned above do not allow for such comprehensive analysis, highlighting the need for the creation of DR IQA databases V1 and V2. Not only do these new datasets allow for comprehensive analysis, they have also allowed us to design and train effective DR IQA models as we shall see later in this chapter.

While DR IQA databases V1 and V2 are much smaller in scale than the Waterloo Exp-II database, with 32,912 distorted images each, they are still much larger than all subject-rated IQA datasets listed in Table 3.1. Conducting subjective testing for two datasets with a total of 65,824 distorted images is extremely difficult. Therefore, we annotate DR IQA databases V1 and V2 with the synthetic quality benchmark (SQB) that was developed in Section 3.4 for the Waterloo Exp-II database. In fact, the SQB labels for all distorted images in the Waterloo Exp-II database, DR IQA database V1, DR IQA database V2, and nine other subject-rated datasets were generated together (see Section 3.4.2 and Table 3.4). To observe how well the DR IQA databases V1 and V2 cover the perceptual quality spectrum, we plot their SQB histograms in Fig. 4.20. The SQB has a quality range of 0 to 100, where 100 is representative of the best while 0 represents the worst quality. It can be seen from Fig. 4.20 that both DR IQA databases have more than at least 100 annotated images for each integer quality value above 10, thereby providing satisfactory representation of each quality value. It can also be seen that the quality range of 50 to 100 has the most images, thereby ensuring that the higher quality range, which is difficult to assess for objective IQA methods [19], is sufficiently represented.

4.5 DR IQA Model Design

In this section, we will develop three DR IQA models. The first two are parametric models based on empirical observations of distortion behaviors, and the third is based on learning through SVR. For convenience, we name them Model 1, Model 2, and Model 3, respectively. Each model has 35 parameter settings corresponding to the combination of seven distortion combinations (including one for the all distortion combination case) and five different $\mathrm{AS_{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS_{FD}}$ combinations depending upon the choice of FR/NR method for $\mathrm{AS_{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and the choice of FR method for $\mathrm{RS_{FD}}$.

4.5.1 Distortion Behavior based Model 1

Model 1 for DR IQA Framework Scenario 1

We begin by considering Scenario 1 of the DR IQA framework (see Section 4.3.2 and Fig. 4.17) where the PR image is considered available in addition to the DR and FD images. Thus, it is possible to compute $\mathrm{AS_{DR}}$ given by Equation 4.2 and $\mathrm{RS_{FD}}$ given by Equation 4.4 by using FR methods. We choose to use FSIMc [14] as it was found to be one of the best performing methods in our comprehensive review of FR methods in Chapter 2 and is also not a part of the SQB generation mechanism. The sample $\mathrm{AS_{FD}}$ versus $\mathrm{RS_{FD}}$ plots for the five major distortion combinations that utilize FSIMc have already been shown in Figures 4.11 to 4.15. Our goal is to predict $\mathrm{AS_{FD}}$ by utilizing the FSIMc computed $\mathrm{AS_{DR}}$ and $\mathrm{RS_{FD}}$ scores, as stated in Equation 4.5. In this and the next subsection, we use DR IQA databases V1 and V2 in a combined manner, i.e., they are considered as one dataset. They will be considered separately in Section 4.5.3.

Initially, let us consider the case of the Blur-JPEG distortion combination for which $\mathrm{AS_{FD}}$ versus $\mathrm{RS_{FD}}$ plots for four test PR images are shown in Fig. 4.11. Observing the individual curves in the plots of Fig. 4.11, we can see that they follow a rather linear pattern. Similar behavior is observed in the $\mathrm{AS_{FD}}$ versus $\mathrm{RS_{FD}}$ plots of all images of DR IQA databases V1 and V2. Therefore, we use a simple linear model to approximate each curve:

$$\widehat{\mathrm{AS}_{\mathrm{FD}}} = m \cdot \mathrm{RS_{FD}} + (1 - \mathrm{AS_{DR}}) \tag{4.8}$$

where $\widehat{\mathrm{AS}_{\mathrm{FD}}}$ is the predicted value of $\mathrm{AS_{FD}}$ and m is the slope parameter. This model has only one coefficient that needs to be estimated. We applied this model to all the 11 $\mathrm{AS_{FD}}$ versus $\mathrm{RS_{FD}}$ curves corresponding to each of the 68 pristine reference images in DR IQA databases V1 and V2 for the case of Blur-JPEG, and optimized the value of coefficient m

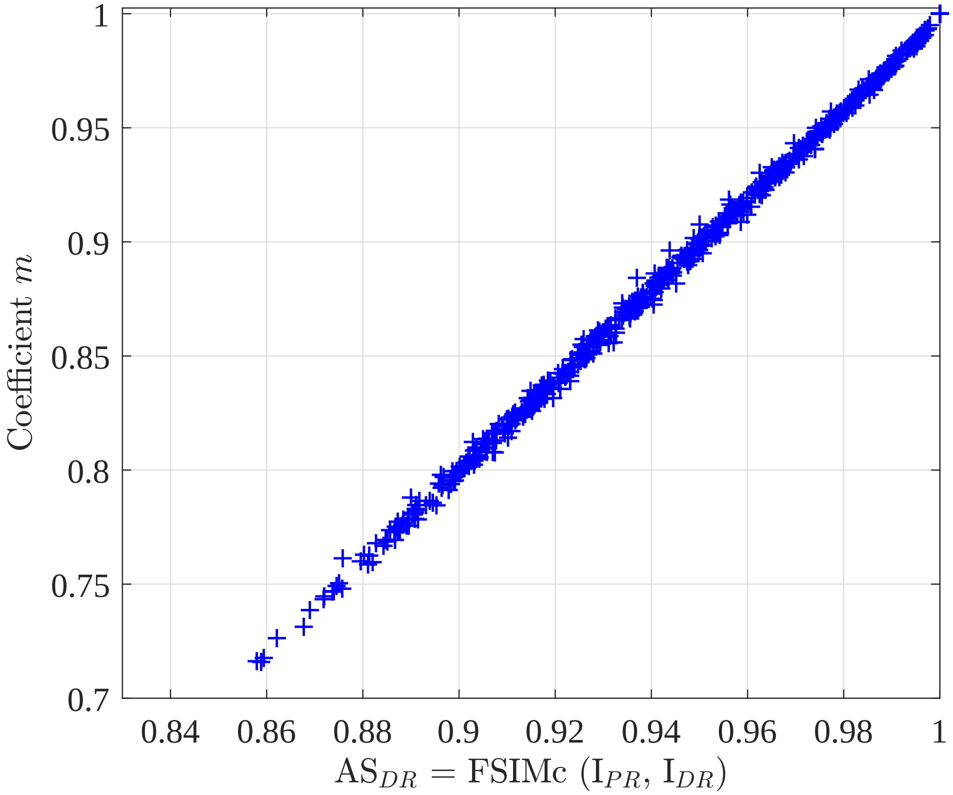


Figure 4.21: Blur-JPEG scatter plot of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$ for the entire DR IQA databases V1 and V2.

in each case by using MATLAB. We plot the 748 estimations of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$ for both DR IQA databases V1 and V2, as shown in Figure 4.21. The important finding here is that the behavior of coefficient m with respect to $\mathrm{AS}_{\mathrm{DR}}$ is highly linear in nature. Thus m can be considered as a function of $\mathrm{AS}_{\mathrm{DR}}$ and approximated well by a simple linear model:

$$\widehat{m} = P_1 \cdot \mathrm{AS}_{\mathrm{DR}} + P_2 \tag{4.9}$$

where $\widehat{m}$ is the predicted value of coefficient m, P_1 is the slope coefficient and P_2 is the intercept coefficient. By using this linear model, we estimate the two coefficients P_1 and P_2 for both DR IQA databases V1 and V2 in a combined manner. We can then rewrite

Equation 4.8 as follows:

$$\widehat{\mathrm{AS}_{\mathrm{FD}}} = \widehat{m} \cdot \mathrm{RS}_{\mathrm{FD}} + (1 - \mathrm{AS}_{\mathrm{DR}}) \tag{4.10}$$

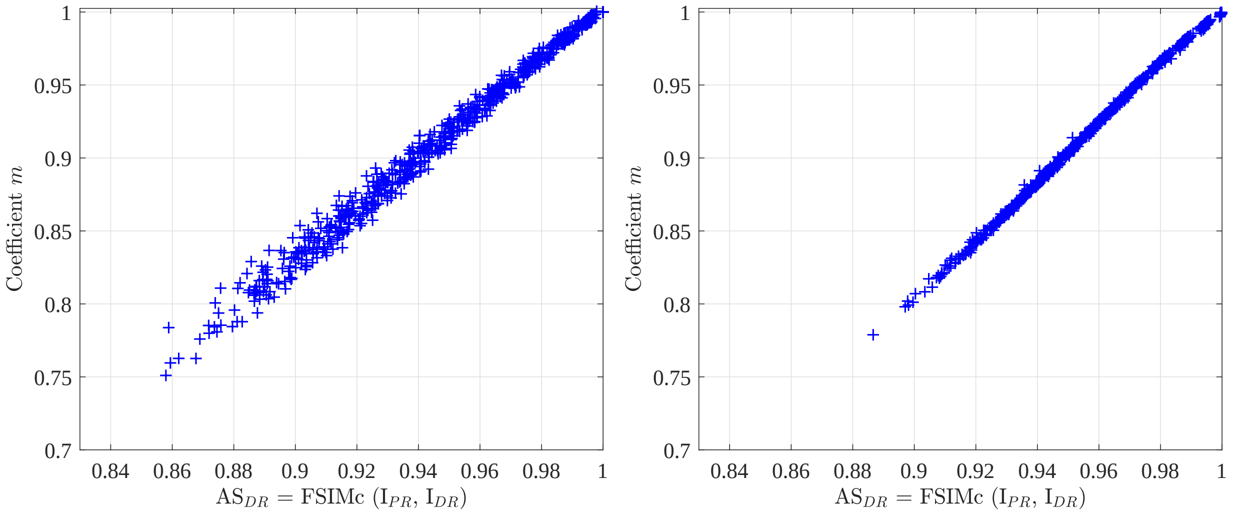
where we have replaced the coefficient m by its predicted value $\widehat{m}$. Plugging in Equation 4.9, we obtain:

$$\widehat{\mathrm{AS}_{\mathrm{FD}}} = (P_1 \cdot \mathrm{AS}_{\mathrm{DR}} + P_2) \cdot \mathrm{RS}_{\mathrm{FD}} + (1 - \mathrm{AS}_{\mathrm{DR}})$$

$$\widehat{\mathrm{AS}_{\mathrm{FD}}} = P_1 \cdot \mathrm{AS}_{\mathrm{DR}} \cdot \mathrm{RS}_{\mathrm{FD}} + P_2 \cdot \mathrm{RS}_{\mathrm{FD}} - \mathrm{AS}_{\mathrm{DR}} + 1 \tag{4.11}$$

Equation 4.11 represents a quality model for the prediction of $\mathrm{AS}_{\mathrm{FD}}$ given that $\mathrm{AS}_{\mathrm{DR}}$ and $\mathrm{RS}_{\mathrm{FD}}$ scores from an FR method are available along with estimated values of coefficients P_1 and P_2. We will refer to this equation as **Model 1**. The vital point to note here is that by following a *2-tier modeling approach*, we were able to narrow down the number of parameters to be estimated to just two (P_1 and P_2) for the entire dataset. Thus, if P_1 and P_2 are known, then Model 1 can be used as a quality prediction model in the realm of degraded reference image quality assessment. This Scenario 1 based DR IQA model will be later referred to as FSIMc-FSIMc as it uses FSIMc to compute both $\mathrm{AS}_{\mathrm{DR}}$ and $\mathrm{RS}_{\mathrm{FD}}$ scores.

While the model of Equation 4.11 has been developed for the distortion combination of Blur-JPEG, it can also be used for other distortion combinations. By observing the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ plots for the Blur-Noise and JPEG-JPEG distortion combinations, depicted respectively in Figures 4.12 and 4.13, it can be seen that the individual curves in these plots also follow a linear pattern. Thus, each curve can be approximated by using the simple linear model of Equation 4.8. The 748 estimations of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$, for both DR IQA databases V1 and V2, are plotted in Fig. 4.22 (a) and in Fig. 4.22 (b) for the combinations of Blur-Noise and JPEG-JPEG, respectively. Like the Blur-JPEG case (Fig. 4.21), the behavior of coefficient m is linear with respect to $\mathrm{AS}_{\mathrm{DR}}$ for both the Blur-Noise and JPEG-JPEG combinations and it can be approximated by the linear model



(a) Blur-Noise

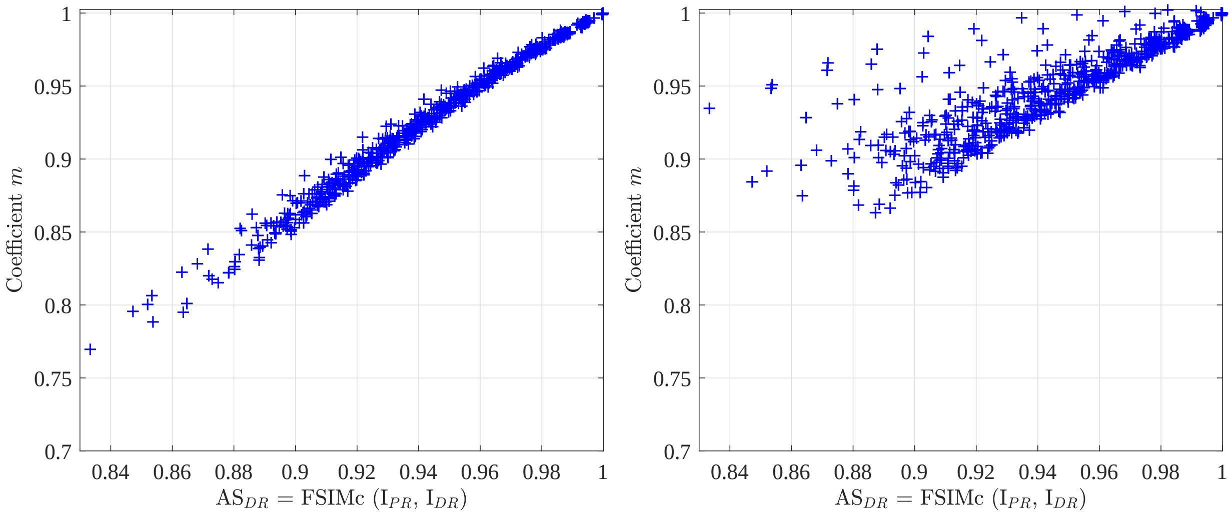
(b) JPEG-JPEG

Figure 4.22: Scatter plots of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$ for the entire DR IQA databases V1 and V2 for the distortion combinations of: (a) Blur-Noise, and (b) JPEG-JPEG. The same scale as Fig. 4.21 is used to enable direct comparison.

of Equation 4.9, which again leads to the DR IQA model of Equation 4.11, albeit with parameters P_1 and P_2 specific to each distortion combination.

The $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ plots for the distortion combinations of Noise-JPEG and Noise-JPEG2000, depicted respectively in Figures 4.14 and 4.15, differ from those of the Blur-JPEG, Blur-Noise, and JPEG-JPEG combinations. These differences have been discussed earlier in Section 4.3.1. Suffice it to say that the curves in the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ plots for Noise-JPEG and Noise-JPEG2000 do not follow a completely linear pattern. A simple 2-tier linear model, based on Equation 4.8, leads to higher approximation errors for the cases of Noise-JPEG and Noise-JPEG2000. Using a cubic polynomial based 2-tier model achieves good individual approximations, but we need to estimate four coefficients for each individual $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curve. These coefficients vary significantly across image content and this limits the generalization capability of such a model.

Careful inspection of the trend followed by the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curves in the case of Noise-JPEG2000 (Fig. 4.15) shows that they are composed of two distinct regions. In the



(a) Noise-JPEG

(b) Noise-JPEG2000

Figure 4.23: Scatter plots of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$ for the entire DR IQA databases V1 and V2 for the distortion combinations of: (a) Noise-JPEG, and (b) Noise-JPEG2000. The same scale as Figures 4.21 and 4.22 is used to enable direct comparison.

first region, the curve is either around $\mathrm{AS}_{\mathrm{DR}}$ or overshoots it and later comes back to it. This region starts from the minimum stage-2 distortion level and extends to a higher stage-2 distortion level depending upon the stage-1 distortion level. In the second region, the curve departs the $\mathrm{AS}_{\mathrm{DR}}$ value and follows a linear pattern which is approximately parallel to the diagonal. This region starts from some lower to mid-level stage-2 distortion and extends up to the maximum stage-2 distortion level. The trend followed by the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curves in the case of Noise-JPEG (Fig. 4.14) is similar to that of Noise-JPEG2000, except that these curves do not overshoot $\mathrm{AS}_{\mathrm{DR}}$ by as much as the Noise-JPEG2000 case.

Therefore, we opt to use a piecewise linear model composed of two pieces which lie in the first and second regions, respectively. We approximate each curve in the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ plots of the Noise-JPEG and Noise-JPEG2000 combinations by using the linear model of Equation 4.8. The 748 estimations of coefficient m versus $\mathrm{AS}_{\mathrm{DR}}$, for both DR IQA databases V1 and V2, are plotted in Fig. 4.23 (a) and Fig. 4.23 (b) for the combinations of Noise-JPEG and Noise-JPEG2000, respectively. The behavior of coefficient m with respect to $\mathrm{AS}_{\mathrm{DR}}$ can be approximated by the linear model of Equation 4.9 for both Noise-JPEG and Noise-JPEG2000, where more error will be incurred for the latter case. This will again lead to the DR IQA model of Equation 4.11, with coefficients P_1 and P_2 specific to each distortion combination. However, this is not the final step for the cases of Noise-JPEG and Noise-JPEG2000. The 2-tier model embodied by Equation 4.11 is applicable to the second region only. For the first region, we directly use $\mathrm{AS}_{\mathrm{DR}}$ to predict $\mathrm{AS}_{\mathrm{FD}}$. A straightforward approach to combine the prediction models for these two regions is as follows:

```
**if** $\widehat{\mathrm{AS}_{\mathrm{FD}}} > \mathrm{AS}_{\mathrm{DR}}$ **then**
    $\widehat{\mathrm{AS}_{\mathrm{FD}}} \leftarrow \mathrm{AS}_{\mathrm{DR}}$
**else**
    $\widehat{\mathrm{AS}_{\mathrm{FD}}} \leftarrow \widehat{\mathrm{AS}_{\mathrm{FD}}}$
**end if**
```

This simple solution allows us to implement a piecewise linear version of **Model 1** for the cases of Noise-JPEG and Noise-JPEG2000. It will not be able to cater to the overshoot above $\mathrm{AS}_{\mathrm{DR}}$ observed in the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curves of Noise-JPEG and Noise-JPEG2000, and thus will not perform as well as it can for the other three distortion combinations. Since the $\mathrm{AS}_{\mathrm{FD}}$ versus $\mathrm{RS}_{\mathrm{FD}}$ curves for the cases of Blur-JPEG, Blur-Noise, and JPEG-JPEG do not exhibit the overshoot above $\mathrm{AS}_{\mathrm{DR}}$, we can apply this modified Model 1 to these distortion combinations as well since it will always stay in the linear mode of the second region for these combinations. This further simplifies the application of Model 1 to different distortion combinations.

Apart from developing Model 1 for the five distortion combinations separately, we also develop this model for two other cases: 1) For NBJ-JPG, where the distortion combinations of Noise-JPEG, Blur-JPEG, and JPEG-JPEG are considered together. This distortion combination is being considered so that comparisons can be made with 2stepQA [1, 2], which is designed for the case where the second distortion stage is JPEG compression. 2) For the *all distortions* case where all five distortion combinations are considered together. The 2-tier modified piecewise linear model based on Equation 4.11 is developed for both these distortion combinations as well.

Model 1 for DR IQA Framework Scenario 2

In practice, the Scenario 1 based DR IQA framework is not applicable when PR images are unavailable and thus FSIMc generated AS_{DR} scores cannot be computed. Instead, only the DR and FD images are available, which means that an FR method can only be used to generate RS_{FD} scores. Thus, only Scenario 2 of the DR IQA framework (see Section 4.3.2 and Fig. 4.18) is applicable where an NR method is used to compute an estimation of AS_{DR}, i.e., $\widehat{\text{AS}_{\text{DR}}}$, which together with the FR computed RS_{FD}, can be used to predict AS_{FD}, as stated in Equation 4.7. We use three NR IQA algorithms, CORNIA [141], dipIQ [36], and NIQE [3] to predict the quality of DR images, i.e., $\widehat{\text{AS}_{\text{DR}}}$. CORNIA and dipIQ are selected as they were found to be the top performers in our comprehensive performance evaluation of NR methods in Chapter 2. We also use NIQE as it has been used in the 2stepQA model [1, 2]. We develop three separate Scenario 2 based DR IQA Model 1 versions by using $\widehat{\text{AS}_{\text{DR}}}$ from CORNIA, dipIQ, and NIQE, and RS_{FD} from the FR method FSIMc [14], thereby leading to the following $\widehat{\text{AS}_{\text{DR}}}$-$\text{RS}_{\text{FD}}$ combinations: CORNIA-FSIMc, dipIQ-FSIMc, and NIQE-FSIMc. We also develop a fourth $\widehat{\text{AS}_{\text{DR}}}$-$\text{RS}_{\text{FD}}$ combination that uses the FR method MSSSIM [4] with NIQE, i.e., NIQE-MSSSIM, to make direct comparisons with 2stepQA [1, 2] as it also combines NIQE and MSSSIM.

Instead of completely redeveloping Model 1 (given in Equation 4.11) for the Scenario 2 based DR IQA framework, we learn a nonlinear mapping from CORNIA, dipIQ, and NIQE to FSIMc for the CORNIA-FSIMc, dipIQ-FSIMc, and NIQE-FSIMc combinations respectively. We also learn a nonlinear mapping from NIQE to MSSSIM for the NIQE-MSSSIM combination. We adopt the five-parameter modified logistic function used in [24] and given in Equation 4.12 to perform the nonlinear mapping from NR to FR scores.

$$F(N) = \beta_1 \left[\frac{1}{2} - \frac{1}{1 + e^{\{\beta_2 (N - \beta_3)\}}} \right] + \beta_4 N + \beta_5 \tag{4.12}$$

where N denotes the NR (CORNIA, dipIQ, or NIQE) computed quality scores, F denotes the FR (FSIMc or MSSSIM) predicted scores after the mapping step, and β_1, β_2, β_3, β_4, and β_5 are mapping coefficients that are found numerically in MATLAB to maximize the correlation between FR and NR scores. These mapping coefficients are determined, by utilizing both DR IQA databases V1 and V2 in a combined manner, for the stage-1 distortions of Gaussian noise, Gaussian blur, JPEG compression, and all three of them considered together. Not only are these coefficients used during the model development phase, they are also utilized later in the testing phase when the Waterloo Exp-II, LIVE WCmp [2], LIVE MD [31], and MDIVL [34] databases are used to evaluate the performance of the DR IQA models (i.e., they are not determined again for the testing databases). The NR-predicted and FR-mapped $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ is given to Model 1 of Equation 4.11 which becomes:

$$\widetilde{\mathrm{AS}_{\mathrm{FD}}} = P_1 \cdot \widehat{\mathrm{AS}_{\mathrm{DR}}} \cdot \mathrm{RS}_{\mathrm{FD}} + P_2 \cdot \mathrm{RS}_{\mathrm{FD}} - \widehat{\mathrm{AS}_{\mathrm{DR}}} + 1 \tag{4.13}$$

where $\widetilde{\mathrm{AS}_{\mathrm{FD}}}$ is the predicted value of $\mathrm{AS}_{\mathrm{FD}}$ when Model 1 uses an NR-predicted value of $\mathrm{AS}_{\mathrm{DR}}$, i.e., $\widehat{\mathrm{AS}_{\mathrm{DR}}}$.

In all, 35 parameter settings of Model 1 are developed. This is because we have five different $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$ combinations which are:

1. Scenario 1: FSIMc-FSIMc
2. Scenario 2: CORNIA-FSIMc
3. Scenario 2: dipIQ-FSIMc
4. Scenario 2: NIQE-FSIMc
5. Scenario 2: NIQE-MSSSIM

For each of the above $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$ combinations, we develop models for seven multiple distortion combinations which are:

1. Blur-JPEG
2. Blur-Noise
3. JPEG-JPEG

4. Noise-JPEG

5. Noise-JPEG2000

6. Noise/Blur/JPEG-JPEG

7. All five individual distortion combinations considered together

We will evaluate the performance of all parameter settings developed under the umbrella of Model 1 in Section 4.6.

4.5.2 Distortion Behavior based Model 2

We had developed Model 1 in Section 4.5.1 in light of the observations and insights gained through the multiple distortions behavior analysis of Section 4.3.1. Model 1 uses a 2-tier modeling approach to predict $\mathrm{AS_{FD}}$. Motivated by the polynomial form of Model 1, we also develop a direct six-parameter polynomial model, called **Model 2**, given by Equation 4.14 for DR IQA framework Scenario 1.

$$\widehat{\mathrm{AS_{FD}}} = a \cdot \mathrm{AS_{DR}}^2 + b \cdot \mathrm{RS_{FD}}^2 + c \cdot \mathrm{AS_{DR}} + d \cdot \mathrm{RS_{FD}} + e \cdot \mathrm{AS_{DR}} \cdot \mathrm{RS_{FD}} + f, \tag{4.14}$$

where a, b, c, d, e and f are model coefficients, $\widehat{\mathrm{AS_{FD}}}$ is the predicted value of $\mathrm{AS_{FD}}$ and we assume that $\mathrm{AS_{DR}}$ is being computed by an FR method. Model coefficients are estimated directly by using MATLAB and using both DR IQA databases V1 and V2 in a combined manner. It can be seen that Model 2 reduces to Model 1 when: $a = 0$, $b = 0$, $c = -1$, $d = P_2$, $e = P_1$ and $f = 1$.

For the case of DR IQA framework Scenario 2, when NR methods are used to predict $\mathrm{AS_{DR}}$, i.e., they provide $\widehat{\mathrm{AS_{DR}}}$, Model 2 takes the form:

$$\widetilde{\mathrm{AS_{FD}}} = a \cdot \widehat{\mathrm{AS_{DR}}}^2 + b \cdot \mathrm{RS_{FD}}^2 + c \cdot \widehat{\mathrm{AS_{DR}}} + d \cdot \mathrm{RS_{FD}} + e \cdot \widehat{\mathrm{AS_{DR}}} \cdot \mathrm{RS_{FD}} + f, \tag{4.15}$$

where $\widetilde{\mathrm{AS}_{\mathrm{FD}}}$ is the predicted value of $\mathrm{AS}_{\mathrm{FD}}$ when $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ is used by Model 2. The NR (CORNIA, dipIQ, NIQE) predicted DR image quality scores are mapped to respective FR (FSIMc or MSSSIM) scores by using the nonlinear mapping function of Equation 4.12, as described in Section 4.5.1.

Since there are five different $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$ combinations, each with its own set of seven distortion combinations, 35 different parameter settings are developed under the umbrella of Model 2. For details of these combinations, refer to Section 4.5.1.

4.5.3 SVR based Model 3

In addition to the empirical distortion behavior based Models 1 and 2, we use Support Vector Regression (SVR) [149, 153], to automatically learn the quality prediction functions of Equations 4.5 and 4.7 for DR IQA framework Scenarios 1 and 2, respectively. We refer to the model so developed as **Model 3**, which is again an umbrella for 35 different SVR-based models depending upon the five different $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$ combinations each with its own set of seven distortion combinations. Since distortion behavior based DR IQA modeling, presented in Sections 4.5.1 and 4.5.2, has not been done before, we develop SVR-based models to create an additional reference point to see whether the distortion behavior based models are performing well or if better models can be learned by using machine learning tools. These models will also act as DR IQA models in their own right.

We develop Model 3 by using nu-SVR that employs the radial basis function (RBF) kernel [134, 153, 227] and four control parameters which include gamma, cost, nu, and epsilon [134, 227]. For each of the 35 models, the predictors are the FR FSIMc/MSSSIM $\mathrm{RS}_{\mathrm{FD}}$ scores and either the FR FSIMc $\mathrm{AS}_{\mathrm{DR}}$ scores or the NR CORNIA/dipIQ/NIQE $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ scores. The training targets are the $\mathrm{AS}_{\mathrm{FD}}$ scores given by the SQB of the FD images. We use DR IQA database V1 for model training and DR IQA database V2 for model validation. The finalized models are later tested on the Waterloo Exp-II, LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases (see Section 4.6). Before model training, we ensure that the data has been scaled properly as recommended in [134]. During training, we determine the best possible SVR control parameters for a particular model through an extensive grid search by training the model on DR IQA database V1 hundreds and at times thousands of times using different combinations of control parameters, and then selecting the parameters that lead to the best model performance, both in terms of PLCC and SRCC, on the validation data (i.e., DR IQA database V2). Since model training by using a large grid is quite time consuming, we use a two-tier grid search. First a coarse-level grid search is performed that identifies the region of the grid that should be focused on. This is followed by a fine-level grid search to finalize the SVR parameters. The finalized SVR control parameters are used to train the final model on DR IQA database V1.

4.6 Performance Evaluation of DR IQA Models

We evaluate the performance of the DR IQA models by using the same test databases and evaluation criteria as were used for baseline performance evaluation in Section 4.2 and described in Section 4.2.1. The test datasets, which include the Waterloo Exp-II, LIVE MD [31], MDIVL [34], and LIVE WCmp [2], have no overlap with the datasets used for model development in Section 4.5 (i.e., DR IQA databases V1 and V2). The performance of DR IQA Models 1, 2, and 3, in terms of PLCC and SRCC, is given in Tables 4.7, 4.8, and 4.9, respectively. The PLCC and SRCC values are computed by considering model outcomes against SQB for the Waterloo Exp-II database and against MOS/DMOS for the LIVE MD, MDIVL, and LIVE WCmp databases.

4.6.1 Comparison with Baseline Models

Comparison with FR-based Baseline Models

The performance of the FR based baseline models, depicted in Fig. 4.3, was discussed in Section 4.2.2 and specifically presented in Table 4.1. When this table is compared with Tables 4.7, 4.8, and 4.9, it can be observed that DR IQA Models 1, 2, and 3 (and their underlying models) outperform the FR based baseline models on all four test datasets. The gains in performance exhibited by the DR IQA models are mostly quite comprehensive. For example, the performance of the FR based baseline models is quite poor on both the constituent distortion combinations of the LIVE MD database and on this dataset as a

Table 4.7: Performance of Distortion Behavior based DR IQA Model 1.

Database	Correlation Metric	Predictors		Distortion Combination and Model Type						
		S1	S2	B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	FSIMc	FSIMc	0.9117	0.9104	0.9221	0.8699	0.8583	0.8593	0.8270
		CORNIA	FSIMc	0.9077	0.9208	0.9101	0.8553	0.8344	0.8562	0.8278
		dipIQ	FSIMc	0.9071	0.9207	0.9227	0.9012	0.8627	0.8688	0.8360
		NIQE	FSIMc	0.8904	0.9129	0.8879	0.8634	0.8389	0.8415	0.8182
		NIQE	MSSSIM	0.9304	0.9681	0.9174	0.8724	0.8103	0.7604	0.7552
	SRCC	FSIMc	FSIMc	0.9084	0.9086	0.9166	0.8661	0.8615	0.8705	0.8281
		CORNIA	FSIMc	0.9057	0.9207	0.8922	0.8486	0.8225	0.8607	0.8225
		dipIQ	FSIMc	0.9059	0.9205	0.9117	0.9009	0.8650	0.8799	0.8360
		NIQE	FSIMc	0.8889	0.9130	0.8622	0.8555	0.8250	0.8481	0.8133
		NIQE	MSSSIM	0.9302	0.9689	0.8837	0.8585	0.7629	0.7692	0.7323
LIVE MD[b]	PLCC	FSIMc	FSIMc	0.7573	0.8079	–	–	–	–	0.7491
		CORNIA	FSIMc	0.7793	0.7026	–	–	–	–	0.7789
		dipIQ	FSIMc	0.7795	0.7707	–	–	–	–	0.7597
		NIQE	FSIMc	0.7828	0.7097	–	–	–	–	0.7521
		NIQE	MSSSIM	0.7763	0.6767	–	–	–	–	0.6994
	SRCC	FSIMc	FSIMc	0.7062	0.7994	–	–	–	–	0.7201
		CORNIA	FSIMc	0.7605	0.6013	–	–	–	–	0.7688
		dipIQ	FSIMc	0.7293	0.7004	–	–	–	–	0.7024
		NIQE	FSIMc	0.7642	0.6176	–	–	–	–	0.7591
		NIQE	MSSSIM	0.7571	0.5445	–	–	–	–	0.6249
MDIVL[b,c]	PLCC	FSIMc	FSIMc	0.8958	–	–	0.9202	–	0.9046	0.9010
		CORNIA	FSIMc	0.9186	–	–	0.8537	–	0.8960	0.8906
		dipIQ	FSIMc	0.9157	–	–	0.9012	–	0.9067	0.9030
		NIQE	FSIMc	0.8594	–	–	0.7808	–	0.8381	0.8272
		NIQE	MSSSIM	0.8521	–	–	0.7475	–	0.8203	0.8030
	SRCC	FSIMc	FSIMc	0.8427	–	–	0.8864	–	0.8606	0.8546
		CORNIA	FSIMc	0.8939	–	–	0.8597	–	0.8933	0.8941
		dipIQ	FSIMc	0.8937	–	–	0.8568	–	0.8834	0.8701
		NIQE	FSIMc	0.8401	–	–	0.7514	–	0.8290	0.8155
		NIQE	MSSSIM	0.8322	–	–	0.6997	–	0.7659	0.7450
LIVE WCmp[b,d]	PLCC	CORNIA	FSIMc	0.9097	0.9151	0.9096	0.9151	0.9058	0.9141	0.9162
		dipIQ	FSIMc	0.9081	0.9080	0.9081	0.9065	0.9002	0.9080	0.9072
		NIQE	FSIMc	0.9278	0.9291	0.9278	0.9228	0.9012	0.9292	0.9259
		NIQE	MSSSIM	0.9261	0.9262	0.9260	0.9063	0.8796	0.8893	0.8594
	SRCC	CORNIA	FSIMc	0.9151	0.9181	0.9149	0.9170	0.9070	0.9178	0.9186
		dipIQ	FSIMc	0.9098	0.9094	0.9098	0.9060	0.8997	0.9096	0.9073
		NIQE	FSIMc	0.9284	0.9295	0.9285	0.9238	0.9015	0.9295	0.9264
		NIQE	MSSSIM	0.9284	0.9283	0.9282	0.9083	0.8822	0.8929	0.8639

[a] PLCC and SRCC are computed with respect to SQB. [b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The NBJ-JPG and All Data Model 1 versions are applied to the entire MDIVL database.

[d] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its images cannot be placed into particular distortion combinations. The various Model 1 versions, trained for the seven distortion combinations, are applied to the entire dataset and results have been reported in respective columns accordingly.

Table 4.8: Performance of Distortion Behavior based DR IQA Model 2.

Database	Correlation	Predictors		Distortion Combination and Model Type						
	Metric	S1	S2	B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	FSIMc	FSIMc	0.9135	0.9003	0.9206	0.8751	0.8857	0.8567	0.8296
		CORNIA	FSIMc	0.9090	0.9117	0.9085	0.8654	0.8432	0.8550	0.8288
		dipIQ	FSIMc	0.9078	0.9116	0.9228	0.9075	0.8751	0.8685	0.8416
		NIQE	FSIMc	0.8911	0.9042	0.8854	0.8723	0.8606	0.8414	0.8248
		NIQE	MSSSIM	0.9336	0.9686	0.9182	0.8726	0.8490	0.8132	0.7980
	SRCC	FSIMc	FSIMc	0.9104	0.8975	0.9157	0.8701	0.8685	0.8680	0.8279
		CORNIA	FSIMc	0.9075	0.9115	0.8897	0.8581	0.8369	0.8580	0.8225
		dipIQ	FSIMc	0.9067	0.9112	0.9116	0.9077	0.8737	0.8791	0.8393
		NIQE	FSIMc	0.8898	0.9046	0.8563	0.8613	0.8525	0.8468	0.8173
		NIQE	MSSSIM	0.9343	0.9693	0.8837	0.8432	0.8357	0.8152	0.7850
LIVE MD[b]	PLCC	FSIMc	FSIMc	0.7575	0.7911	–	–	–	–	0.7628
		CORNIA	FSIMc	0.7745	0.7679	–	–	–	–	0.7901
		dipIQ	FSIMc	0.7797	0.7963	–	–	–	–	0.7772
		NIQE	FSIMc	0.7825	0.7736	–	–	–	–	0.7615
		NIQE	MSSSIM	0.7071	0.7089	–	–	–	–	0.6347
	SRCC	FSIMc	FSIMc	0.7069	0.7865	–	–	–	–	0.7376
		CORNIA	FSIMc	0.7546	0.7162	–	–	–	–	0.7795
		dipIQ	FSIMc	0.7294	0.7625	–	–	–	–	0.7351
		NIQE	FSIMc	0.7646	0.7338	–	–	–	–	0.7378
		NIQE	MSSSIM	0.6761	0.6176	–	–	–	–	0.387
MDIVL[b,c]	PLCC	FSIMc	FSIMc	0.8970	–	–	0.9221	–	0.9008	0.8990
		CORNIA	FSIMc	0.9147	–	–	0.8793	–	0.8953	0.8958
		dipIQ	FSIMc	0.9164	–	–	0.9076	–	0.8906	0.8863
		NIQE	FSIMc	0.8586	–	–	0.7988	–	0.8202	0.8167
		NIQE	MSSSIM	0.8572	–	–	0.7633	–	0.7874	0.7550
	SRCC	FSIMc	FSIMc	0.8464	–	–	0.8889	–	0.8486	0.8426
		CORNIA	FSIMc	0.8894	–	–	0.8882	–	0.8873	0.8889
		dipIQ	FSIMc	0.8946	–	–	0.8691	–	0.8642	0.8613
		NIQE	FSIMc	0.8396	–	–	0.7802	–	0.8035	0.8023
		NIQE	MSSSIM	0.8413	–	–	0.7143	–	0.7595	0.7239
LIVE WCmp[b,d]	PLCC	CORNIA	FSIMc	0.9084	0.9109	0.9093	0.9122	0.9035	0.9141	0.9140
		dipIQ	FSIMc	0.9081	0.9079	0.9081	0.9058	0.9058	0.9079	0.9077
		NIQE	FSIMc	0.9277	0.9271	0.9265	0.9201	0.9100	0.9255	0.9233
		NIQE	MSSSIM	0.9255	0.9258	0.9242	0.9067	0.8886	0.9187	0.9134
	SRCC	CORNIA	FSIMc	0.9139	0.9145	0.9143	0.9121	0.9026	0.9168	0.9154
		dipIQ	FSIMc	0.9097	0.9093	0.9095	0.9055	0.9056	0.9092	0.9087
		NIQE	FSIMc	0.9286	0.9264	0.9266	0.9182	0.9063	0.9247	0.9214
		NIQE	MSSSIM	0.9275	0.9279	0.9258	0.9052	0.8880	0.9190	0.9148

[a] PLCC and SRCC are computed with respect to SQB. [b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The NBJ-JPG and All Data Model 2 versions are applied to the entire MDIVL database.

[d] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its images cannot be placed into particular distortion combinations. The various Model 2 versions, trained for the seven distortion combinations, are applied to the entire dataset and results have been reported in respective columns accordingly.

Table 4.9: Performance of SVR based DR IQA Model 3.

Database	Correlation Metric	Predictors		Distortion Combination and Model Type						
		S1	S2	B-JPG	B-N	JPG-JPG	N-JPG	N-JP2	NBJ-JPG	All Data
Waterloo Exp-II[a]	PLCC	FSIMc	FSIMc	0.9287	0.9104	0.9195	0.8877	0.9074	0.8629	0.8416
		CORNIA	FSIMc	0.9215	0.9180	0.9062	0.8547	0.8449	0.8643	0.8389
		dipIQ	FSIMc	0.9228	0.9195	0.9224	0.9112	0.8825	0.8690	0.8448
		NIQE	FSIMc	0.9017	0.9089	0.8853	0.8809	0.8660	0.8422	0.8317
		NIQE	MSSSIM	0.9383	0.9671	0.9159	0.9327	0.8746	0.8172	0.7952
	SRCC	FSIMc	FSIMc	0.9286	0.9079	0.9155	0.8794	0.8957	0.8753	0.8389
		CORNIA	FSIMc	0.9228	0.9181	0.8861	0.8511	0.8386	0.8734	0.8391
		dipIQ	FSIMc	0.9252	0.9194	0.9109	0.9118	0.8809	0.8800	0.8434
		NIQE	FSIMc	0.9038	0.9095	0.8558	0.8712	0.8601	0.8474	0.8230
		NIQE	MSSSIM	0.9387	0.9689	0.8815	0.9276	0.8670	0.8255	0.7852
LIVE MD[b]	PLCC	FSIM	FSIMc	0.7539	0.8082	–	–	–	–	0.7329
		CORNIA	FSIMc	0.7769	0.8175	–	–	–	–	0.7032
		dipIQ	FSIMc	0.7371	0.7791	–	–	–	–	0.7839
		NIQE	FSIMc	0.7712	0.7641	–	–	–	–	0.7602
		NIQE	MSSSIM	0.7349	0.7270	–	–	–	–	0.6468
	SRCC	FSIMc	FSIMc	0.7154	0.7983	–	–	–	–	0.7015
		CORNIA	FSIMc	0.7613	0.8022	–	–	–	–	0.6533
		dipIQ	FSIMc	0.6890	0.7105	–	–	–	–	0.7545
		NIQE	FSIMc	0.7478	0.7229	–	–	–	–	0.7427
		NIQE	MSSSIM	0.7110	0.6575	–	–	–	–	0.4111
MDIVL[b,c]	PLCC	FSIMc	FSIMc	0.8975	–	–	0.9227	–	0.9063	0.9048
		CORNIA	FSIMc	0.9397	–	–	0.8767	–	0.9085	0.9001
		dipIQ	FSIMc	0.9203	–	–	0.9052	–	0.8999	0.8950
		NIQE	FSIMc	0.8578	–	–	0.8182	–	0.8296	0.8046
		NIQE	MSSSIM	0.8671	–	–	0.8161	–	0.7737	0.7799
	SRCC	FSIMc	FSIMc	0.8266	–	–	0.9020	–	0.8758	0.8671
		CORNIA	FSIMc	0.9145	–	–	0.8871	–	0.9020	0.8910
		dipIQ	FSIMc	0.8731	–	–	0.8776	–	0.8759	0.8726
		NIQE	FSIMc	0.8295	–	–	0.8048	–	0.8211	0.7872
		NIQE	MSSSIM	0.8537	–	–	0.8092	–	0.7239	0.7247
LIVE WCmp[b,d]	PLCC	CORNIA	FSIMc	0.9117	0.9156	0.9166	0.9087	0.9023	0.9166	0.9133
		dipIQ	FSIMc	0.9086	0.9083	0.9065	0.9010	0.9010	0.9067	0.9069
		NIQE	FSIMc	0.9239	0.9237	0.9207	0.9169	0.9056	0.9176	0.9202
		NIQE	MSSSIM	0.9247	0.9217	0.9188	0.8609	0.8514	0.9152	0.9131
	SRCC	CORNIA	FSIMc	0.9150	0.9182	0.9184	0.9087	0.9000	0.9156	0.9145
		dipIQ	FSIMc	0.9104	0.9098	0.9059	0.8939	0.8941	0.9061	0.9045
		NIQE	FSIMc	0.9220	0.9219	0.9159	0.9155	0.9005	0.9143	0.9184
		NIQE	MSSSIM	0.9246	0.9235	0.9158	0.8622	0.8546	0.9126	0.9120

[a] PLCC and SRCC are computed with respect to SQB. [b] PLCC and SRCC are computed with respect to MOS/DMOS.

[c] The NBJ-JPG and All Data Model 3 versions are applied to the entire MDIVL database.

[d] The LIVE WCmp database has images that have authentic distortions followed by JPEG compression. Therefore, its images cannot be placed into particular distortion combinations. The various Model 3 versions, trained for the seven distortion combinations, are applied to the entire dataset and results have been reported in respective columns accordingly.

whole, and also on the B-JPG combination of the MDIVL database and on this dataset as a whole. However, the DR IQA models perform well on these datasets, exhibiting tremendous gains compared to the baseline. The only exceptions are the cases of B-N and N-JPG in the Waterloo Exp-II database and the case of N-JPG in the MDIVL database, where the MSSSIM based baseline models do better than some DR IQA models, yet it cannot outperform all of them. The superior performance of the DR IQA models relative to the FR based baseline of Fig. 4.3 demonstrates the shortcomings of the FR paradigm in the absence of PR images at the final destination and at the same time it establishes the superiority of the DR IQA framework for this case, especially its Scenario 2 (see Fig. 4.18).

In Section 4.2.2, we had also evaluated the performance of FR methods when they are used to determine the absolute quality scores of FD images with respect to their PR images (see Table 4.2). Essentially, we had determined $\mathrm{AS}_{\mathrm{FD}}$ scores using FR methods. Given that the performance of FR methods is well established when PR images are available, and that it is the goal of the DR IQA methods to predict $\mathrm{AS}_{\mathrm{FD}}$, it is vital to compare how well the DR IQA models do against FR computed $\mathrm{AS}_{\mathrm{FD}}$ scores. By comparing Tables 4.7, 4.8, and 4.9 with Table 4.2, we can make the following observations: 1) Generally, DR IQA Models 1, 2, and 3, perform better than FR predicted $\mathrm{AS}_{\mathrm{FD}}$ on the B-JPG and B-N combinations of the LIVE MD database, while performing at par with its all data case. 2) The DR IQA models also perform at par with the FR predicted $\mathrm{AS}_{\mathrm{FD}}$ for the B-JPG, N-JPG, and all data cases of the MDIVL database, where some DR IQA models also outperform the FR methods. 3) On the B-JPG, B-N, JPG-JPG, and N-JPG, combinations of the Waterloo Exp-II database, DR IQA Models 1, 2, and 3, mostly perform at par with the FR predicted $\mathrm{AS}_{\mathrm{FD}}$ scores. The performance demonstrated by the DR IQA models in points 1, 2, and 3 so far, relative to FR-predicted $\mathrm{AS}_{\mathrm{FD}}$, is no small achievement given that FR performance is usually considered as an upper bound in IQA if FR methods have access to the PR images (which is the case here). 4) While the DR IQA Models 1, 2, and 3, perform satisfactorily on the N-JP2, NBJ-JPG, and all data cases of the Waterloo Exp-II database, they do not approach the superior performance exhibited by the FR-predicted $\mathrm{AS}_{\mathrm{FD}}$. This highlights the difficult nature of the N-JP2 case, as can be seen in the distortion behavior plot of Fig. 4.15. It also highlights the difficult nature of the NBJ-JPG and all data cases, where multiple distortion combinations are considered together, making the task of IQA all the more difficult. It should be noted that this analysis is not possible for the LIVE WCmp database as it lacks PR images and FR-predicted $\mathrm{AS}_{\mathrm{FD}}$ scores cannot be determined. However, DR IQA Models 1, 2, and 3 perform quite well on this database as can be seen from Tables 4.7, 4.8, and 4.9, respectively.

Comparison with NR-based Baseline Models

The performance of the NR based baseline models, depicted in Fig. 4.4, was discussed in Section 4.2.3 and specifically presented in Table 4.3. When this table is compared with Tables 4.7, 4.8, and 4.9, it can be observed that most underlying models of DR IQA Models 1, 2, and 3 comprehensively outperform the NR based baseline models on all four test datasets. While there are some exceptions, for example, CORNIA [141] does well on the B-N and B-JPG cases of the LIVE MD and MDIVL databases, respectively, it does inadequately on other distortion combinations and datasets. On the other hand, most DR IQA models offer good stable performance across the wide ranging test data. While dipIQ [36] performs quite well across the Waterloo Exp-II dataset, as we noted in Section 4.2.3, this dataset is favored towards dipIQ because both dipIQ and SQB (used to annotate the Waterloo Exp-II database) follow a ranking based design philosophy. Even then, the DR IQA models perform better than dipIQ, with the exception of the all data case of the Waterloo Exp-II database. As discussed earlier, a simple counterargument to the very premise of developing DR IQA as a new paradigm is that NR methods should be used to directly evaluate the quality of the FD images. Here, by demonstrating the superiority of DR IQA based methods over the NR based baseline models, we have shown that if additional information is available in the form of DR images, then incorporating such information in the task of quality assessment of FD images can lead to much better performance instead of using NR methods directly on FD images. The superior performance of DR IQA framework Scenario 2 based models, compared to the NR based baseline models, has also shown that in the absence of PR images, NR methods can be effectively used to compute $\widehat{\mathrm{AS}_{\mathrm{DR}}}$ scores for DR images, which together with the FR-computed $\mathrm{RS}_{\mathrm{FD}}$ scores between the DR and FD images, can lead to effective DR IQA models, again highlighting the utility of using additional information provided by DR images even if it is through their NR-predicted quality.

Comparison with 2stepQA Baseline Model

The performance of the 2stepQA model [1,2] was discussed in Section 4.2.4 and specifically presented in Table 4.4. We regard 2stepQA as another baseline model, in fact it is the most relevant baseline model since it utilizes both the DR and FD images to perform the quality assessment of FD images. By comparing Table 4.4 with Tables 4.7, 4.8, and 4.9, the following observations can be made: 1) Since 2stepQA uses NIQE to estimate the quality of DR images, MSSSIM to estimate the quality of the FD image with respect to the DR image, and then combines them to yield the final quality score, we included the NIQE-MSSSIM combination in DR IQA Models 1, 2, and 3, for direct comparison. It can be seen that DR IQA Models 1, 2, and 3, that utilize NIQE and MSSSIM, perform better than 2stepQA on the JPG-JPG, N-JPG, N-JP2, and all data cases of the Waterloo Exp-II database while performing at par with it for the other three distortion combinations. The NIQE-MSSSIM based DR IQA models perform better than 2stepQA on the LIVE MD database while the opposite is true for the MDIVL database. Finally, the said DR IQA models perform at par with 2stepQA on the LIVE WCmp database, which is quite surprising since the DR images of this dataset are authentically distorted whereas the DR images in the training datasets of the DR IQA models have only noise, blur or JPEG compression. Even then, the DR IQA models do a fine job on this database. 2) When 2stepQA is compared with the NIQE-FSIMc version of the DR IQA Models 1, 2, and 3, it can be seen that the DR IQA models perform a bit better on the LIVE WCmp database, a bit worse on the B-JPG and N-JPG combinations of MDIVL but they perform better than 2stepQA over the entire MDIVL database, and better than 2stepQA on the LIVE MD database. On the Waterloo Exp-II database, the 2stepQA model performs better on the B-JPG, B-N, and the JPG-JPG combinations, but it is comprehensively outperformed by the DR IQA models on the N-JP2, NBJ-JPG, and all data combinations, while also being slightly outperformed on the N-JPG combination. 3) However, the most important finding here is that the CORNIA-FSIMc and dipIQ-FSIMc based DR IQA Models 1, 2, and 3, comprehensively outperform 2stepQA in many distortion combinations and datasets (with the exception of a few cases). For example, 2stepQA does poorly for the most difficult all data case of Waterloo Exp-II, LIVE MD, and MDIVL databases. By contrast, the DR IQA based models do quite well in this case and comprehensively outperform 2stepQA. We can attribute the overall better performance of the DR IQA based models over the 2stepQA model to two factors: 1) The modeling approach taken in 2stepQA is rather ad hoc, where an NR and FR method have been arbitrarily combined, whereas the DR IQA models have either been developed based on empirical study of the behavior of multiple simultaneous distortions or through SVR (which also incorporates distortion behavior in the training process). 2) The choice of NR and FR methods to be combined is important. Since CORNIA [141] and dipIQ [36] perform better than NIQE [3], some DR models that utilize CORNIA and dipIQ perform better than those that use NIQE.

Validation of using SQB based Training Data

Finally, the superior performance of the DR IQA models with respect to both the FR and NR based baselines, and relative to the 2stepQA model also validates the use of our SQB annotated training/validation datasets, i.e., DR IQA databases V1 and V2. Since these datasets do not have subjective ratings, they have been annotated by the automatically generated synthetic quality benchmark (SQB) scores described in Section 3.4. By testing the DR IQA models on LIVE MD [31], MDIVL [34], and LIVE WCmp [2], which are subject-rated datasets, and finding them to perform better than the baselines, we have again shown the validity of using SQB as an alternative IQA data annotation mechanism.

4.6.2 Inter-Model Comparisons

We perform three kinds of inter-model comparisons: 1) Approach-based comparisons, 2) DR IQA framework-based comparisons, and 3) Distortion combination-based comparisons.

Approach-based Comparisons

First, we compare the three fundamental DR IQA modeling approaches developed in Sections 4.5.1, 4.5.2, and 4.5.3, and referred to as Model 1, 2, and 3, respectively. Each approach has 35 underlying models and they will be considered in a corresponding manner. The simplest of these approaches is Model 1 (Section 4.5.1) which was developed by directly observing the multiple distortion behavior plots exemplified by Figures 4.11 to 4.15 and led to the development of simple two-parameter models. Model 2 (Section 4.5.2) is also distortion behavior based and follows directly from Model 1. It consists of six-parameter based models. The most complex of the three models is Model 3 (Section 4.5.3) which used a sophisticated machine learning tool (SVR) to automatically learn DR IQA models and its development was also computationally expensive.

Intuitively, one would expect that Model 3 will excel in performance compared to Models 1 and 2, because it uses SVR to automatically learn quality models. However, as a comparison of Tables 4.7, 4.8, and 4.9 shows, this is not necessarily the case. Generally, all three modeling approaches offer quite similar performance. The following can be observed: 1) On the Waterloo Exp-2 database, Model 1 offers the best performance for the B-N and JPG-JPG combinations, while Model 3 performs the best in the other five distortion combinations. This is expected, especially in the case of the N-JPG and N-JP2 combinations as their distortion behavior plots were the most complex (see Fig. 4.14 and 4.15 respectively) and the piecewise linear nature of Model 1 with its inability to model the overshoot phenomenon witnessed for these combinations meant that it incurred approximation errors for these combinations. For N-JPG and N-JP2 Model 2 performs better than Model 1, as can be expected as the former is capable of approximating the nonlinear nature of these distortions. However, surprisingly, the performance of all three models is quite close for the most complex distortion combinations of NBJ-JPG and all data. 2) On the LIVE MD database [31], all three modeling approaches offer more or less similar performance on the B-JPG combination. Models 2 and 3 offer similar performance on the B-N combination while Model 1 lags behind in this case. All three models offer mixed performance for the all data case of this database. 3) On the MDIVL database [34], Model 3 performs better on the individual distortion combinations while Model 1 performs better in the difficult NBJ-JPG and all data cases. 4) On the LIVE WCmp database [2], Model 1 offers the best performance, followed by Model 2, and then by Model 3. This is a significant finding because the DR images of the LIVE WCmp database are authentically distorted, and such distortions are not found in the training data. Thus, by performing better than SVR-based

Model 3, Models 1 and 2 show that they have better generalization ability.

While we have mentioned one model or another as offering better performance than the others in the above analysis, Tables 4.7, 4.8, and 4.9 show that although one model may not perform better than another, it also does not lag far behind in performance, with the only exception being the performance of Model 1 on the B-N combination of LIVE MD. Since Model 3 uses SVR, which is a sophisticated machine learning tool, to automatically learn quality models, the fact that Models 1 and 2 not only perform at par with Model 3, but they even outperform Model 3 in quite a number of cases establishes the validity of our distortion behavior based approach to develop DR IQA models. The simplicity of Models 1 and 2, especially the former since it is just a two-parameter model, adds to the overall strength of the distortion behavior based modeling approach.

DR IQA Framework-based Comparisons

Depending upon the choice of FR/NR methods used for $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and the choice of FR methods used for $\mathrm{RS}_{\mathrm{FD}}$, five different combinations are possible. These are mentioned in the format of $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$-$\mathrm{RS}_{\mathrm{FD}}$ as: 1) FSIMc-FSIMc, 2) CORNIA-FSIMc, 3) dipIQ-FSIMc, 4) NIQE-FSIMc, and 5) NIQE-MSSSIM. Of these five, the first belongs to DR IQA framework Scenario 1 which considers the PR image to be available for $\mathrm{AS}_{\mathrm{DR}}$ computation, while the rest belong to the more practical Scenario 2 where the PR image is considered unavailable.

Since, FR methods outperform NR ones and are more reliable, it is natural to think that the Scenario 1 combination of FSIMc-FSIMc should outperform Scenario 2 based methods because it uses an FR method to determine both $\mathrm{AS}_{\mathrm{DR}}$ and $\mathrm{RS}_{\mathrm{FD}}$. However, by observing Tables 4.7, 4.8, and 4.9 for Models 1, 2, and 3, respectively, we can see that this is not the case as in most cases Scenario 2 based DR IQA models perform at par with the Scenario 1 based FSIMc-FSIMc, and in some cases they even outperform it. This is a very important finding as it demonstrates that the more practical Scenario 2 based DR IQA models, that work in the absence of the PR image, are able to perform as well as the Scenario 1 based model which uses the PR image. Thus, the lack of having access to the PR image does not constrain the performance of DR IQA models, as long as appropriate NR and FR methods are used to compute $\widehat{\mathrm{AS}}_{\mathrm{DR}}$ and $\mathrm{RS}_{\mathrm{FD}}$, respectively. This highlights that DR IQA models can be applied to practical multiple distortions based media distribution systems, such as the one shown in Fig. 4.2, and have the promise to yield good performance, which cannot be said for FR methods because of lack of access to the PR images and for NR methods due to their performance issues.

Distortion Combination-based Comparisons

In our current work, we are dealing with the following seven multiple distortion combinations: 1) B-JPG, 2) B-N, 3) JPG-JPG, 4) N-JPG, 5) N-JP2, 6) NBJ-JPG, and 7) the first five distortion combinations together (the all data case). Tables 4.7, 4.8, and 4.9, respectively show that DR IQA Models 1, 2, and 3, perform quite consistently across the LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases. These datasets also do not have all seven of the above-mentioned distortion combinations. Thus, for the distortion combination-based analysis, we will focus on the Waterloo Exp-II database which has all the distortion combinations. It can be seen from the respective tables that DR IQA Models 1, 2, and 3, perform quite well for the B-JPG, B-N, and JPG-JPG distortion combinations. Figures 4.11, 4.12, and 4.13, show that the distortion behavior for these multiple distortion cases is quite straightforward and hence it can be modeled effectively. The above-mentioned tables show that for the cases of N-JPG, N-JP2, NBJ-JPG, and all data, the performance of DR IQA Models 1, 2, and 3, is satisfactory when considered independently but is lowered quite considerably when considered in comparison with the cases of B-JPG, B-N, and JPG-JPG. This is understandable because as Figures 4.14 and 4.15 show, the distortion behavior for the cases of N-JPG and N-JP2 is nonlinear in nature and thus more difficult to capture. It not only makes the development of DR IQA models for N-JPG and N-JP2 difficult, but also hardens the task of model development for the NBJ-JPG and all data cases (which include these combinations). The fact that SVR based Model 3 can also not do much better than Models 1 and 2, shows the difficult nature of model development for these cases. This area is therefore calling for future research to come up with innovative means to enhance the performance of DR IQA models for these difficult distortion combinations.

4.7 Practical Application

The DR IQA paradigm has its roots in the practical limitations of the contemporary paradigms of FR/RR/NR IQA. Although this new paradigm can be regarded as being in its infancy and the models developed in this chapter are one of the first attempts to explore it, nevertheless, they are practically applicable even in their current form. Among the different practical multiple distortion scenarios discussed in Section 4.1, many involve images that are first afflicted by noise and/or blur due to imperfect capture conditions followed by compression due to storage or media distribution requirements. For example, social media platforms such as Facebook or video sharing platforms such as YouTube compress content which may be contaminated by distortions such as noise or blur. Thus, the distortion combination specific DR IQA models developed for the cases of B-JPG, JPG-JPG, and N-JPG can be deployed if it is known that the second distortion is JPEG compression and information is available about the first distortion. However, in a more realistic scenario, information about the first distortion may not be available. In such a case the NBJ-JPG DR IQA model can be applied. Since the DR IQA models have outperformed the NR IQA based baseline models, which are the only other practically feasible options in the absence of the pristine images, the DR IQA models can enhance the perceptual quality prediction capabilities of media distribution chains that currently rely on NR IQA models. As noted in Section 4.6, Scenario 2 based DR IQA models perform as well as Scenario 1 based models, and we know from Section 4.3.2 that they do not require any modifications in the media delivery chains. Thus, Scenario 2 based DR IQA models can be readily deployed in currently functioning media delivery systems.

4.8 Summary

In this chapter, we attempt to tackle the challenging practical problem of IQA of images undergoing multiple stages of distortions where earlier degraded versions of the final distorted visual content are also available. We demonstrate that FR methods are unable to perform well in such a setting because pristine reference images are generally not available and NR methods suffer from performance issues. To study this challenging area, we conduct a first-of-its-kind comprehensive multiple distortions behavior analysis specifically for the case of a two-stage distortion pipeline where five different practically prevalent simultaneous distortion combinations are considered. Next, we introduce a new IQA paradigm, which we call degraded reference (DR) IQA, that is applicable to the real-world visual content distribution systems for which FR methods are inapplicable and NR methods struggle for performance. Specifically, we introduce two DR IQA framework scenarios, where the pristine reference images are considered available in the first and unavailable in the second. We also construct two new DR IQA databases (V1 and V2) that are composed of pristine references, singly distorted degraded references, and final distorted images with multiple distortions. Overall, these databases have more than 30,000 images each and use the SQB mechanism (developed in Chapter 3) for quality annotation. By using the lessons learned from the multiple distortions behavior analysis and the DR IQA databases, we are able to develop two novel DR IQA modeling approaches that evaluate the quality of a final distorted image by also utilizing the degraded references. We also develop SVR-based models as an additional comparison point to determine the efficacy of our distortion behavior based DR IQA models. We extensively test the performance of the DR IQA models and also some baseline models on four multiply distorted databases, which include the Waterloo Exp-II database and the subject-rated LIVE MD [31], MDIVL [34], and LIVE WCmp [2] databases. These test databases have no overlap with the DR IQA databases V1 and V2 used for model development. This testing demonstrates the superior performance of DR IQA models when compared with existing FR and NR IQA paradigms, thereby establishing DR IQA as a major IQA paradigm in its own right.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this thesis, we have endeavored to address two major challenges affecting the development of practically applicable IQA algorithms: 1) The shortage of large-scale annotated data, and 2) The design of IQA algorithms for multiply distorted content in the presence of degraded references.

Until now, only small-scale annotated datasets exist in IQA due to the constraints of subjective testing. Thus, machine learning based models trained on such limited data suffer from overfitting issues and cannot be generalized. This also means that the true potential of approaches such as deep neural networks, which require large-scale training data, has not been harnessed in developing IQA models, especially BIQA models. Researchers have tried to enhance the performance of DNN based BIQA methods by focusing on the modeling part and relying on data augmentation, but have achieved only limited success. Efforts to fill the void of large-scale annotated training data have remained missing thus far.

To address the data shortage challenge, we construct the very large-scale Waterloo Exploration-II database. This dataset consists of 3,570 pristine and around 3.45 million distorted images, making it the largest IQA database. The distorted content is created in two stages, where the first stage distorts the pristine reference images using three distortion types at 11 distortion levels leading to singly distorted content. The second stage distorts the singly distorted content using three distortion types at 17 distortion levels leading to five multiply distorted combinations. We adopt content adaptive distortion parameters to ensure that the masking effect of content is taken into account, so that the dataset has images covering the entire quality spectrum. To annotate this dataset, we develop a synthetic quality benchmark (SQB) mechanism that automatically assigns perceptual quality labels to images. SQB uses reciprocal rank fusion (RRF) [23] to fuse four state-of-the-art FR methods in a training-free manner. It then uses mapping coefficients derived from a subject-rated database to map the outcome of RRF to a perceptually meaningful scale. Extensive tests reveal that SQB outperforms state-of-the-art individual and fused FR methods, thus justifying its use as an alternative of subjective testing to annotate IQA data. To test the validity of our approach, we train a DNN based BIQA method, called EONSS, on the SQB-annotated Waterloo Exploration-II database. Extensive testing reveals that EONSS not only comprehensively outperforms existing DNN based BIQA methods but that it also performs better than the very state-of-the-art in BIQA, both in terms of prediction performance and computation time. Chapter 3 discusses our approach to address the data shortage challenge in detail.

As a prerequisite requirement of developing SQB, we conducted a comprehensive performance evaluation survey of state-of-the-art individual and fused FR methods. We evaluated the performance of 43 FR and seven fused FR (22 versions) methods on nine subject-rated databases including both singly and multiply distorted datasets. Among the fused FR methods evaluated was RAS [41] that relies on RRF [23]. To extensively evaluate RAS, we performed an exhaustive search that included testing 737,280 FR fusion combinations. This comprehensive review showed that the RRF based RAS outperforms all other individual and fused FR methods, thus, allowing us to identify the RRF [23] as a basis for our SQB mechanism and also the four FR methods to fuse. We also evaluated the performance of 14 NR methods in this review. To-date this is the largest performance evaluation survey carried out in the area of IQA and is discussed in detail in Chapter 2.

In practical media distribution systems, visual content undergoes a number of degradations between the source and the final destination, making it multiply distorted. Access to the pristine reference version of such content is either extremely rare or altogether nonexistent, which makes the application of FR and RR IQA paradigms infeasible. Contemporary research efforts are focused on using NR IQA or BIQA methods to evaluate the quality of multiply distorted images, but have shown limited success due to the difficult nature of the NR IQA paradigm. Apart from the final multiply distorted images, access to its earlier degraded versions, which we called degraded references, is also usually available in media delivery systems and may prove to be beneficial for the task of quality assessment of the final content. However, due to their respective design philosophies, the three major IQA paradigms are unable to use this additional information.

To address this multi-stage distortion challenge, where degraded references are also available, we propose a new IQA paradigm called degraded reference (DR) IQA. The goal of DR IQA is to evaluate the quality of the final multiply distorted images by also considering their degraded references, but in the absence of pristine references at the end user level. We consider two scenarios for the DR IQA framework, where the pristine references are considered available in the early part of the media distribution system in the first scenario, while absolutely no access to such images is available in the second scenario. For the first time in IQA, we study the behavior of five combinations of multiple simultaneous distortions in detail, which include: 1) Blur-JPEG, 2) Blur-Noise, 3) JPEG-JPEG, 4) Noise-JPEG, and 5) Noise-JPEG2000. We construct two datasets, called DR IQA databases V1 and V2, for the development of DR IQA models. These datasets are formed on the same pattern as the Waterloo Exploration-II database and are annotated by using SQB. The singly distorted content in these datasets constitutes the degraded references. By considering the lessons learned from the multiple distortions behavior analysis and using the DR IQA databases, we develop two novel DR IQA models with 35 parameter settings each, where both DR IQA framework scenarios are considered. We also develop an SVR-based DR IQA model to give us an additional reference point. Extensive testing of the DR IQA models and some baseline models, on an independent set of test datasets, reveals the superior performance of the DR IQA models. The significant performance gains with respect to using NR IQA methods to directly evaluate the quality of final distorted images, amply demonstrates that the use of additional information, available in the form of degraded references, has a highly beneficial impact on enhancing the perceptual quality prediction performance of IQA models, thereby establishing DR IQA as a new IQA paradigm in its own right. Chapter 4 discusses our work on DR IQA in detail.

5.2 Future Work

The work presented in this thesis can be regarded as establishing the foundations of two new research directions in IQA, and can be expanded in many different ways.

Large-scale IQA Database of Even Higher Diversity: Tables 3.8 and 3.9 show that the weighted average PLCC and SRCC of EONSS on nine subject-rated databases is 0.6933 and 0.6509, respectively for the *all distortions* category, while Tables 3.10 and 3.11 show that these values increase to 0.8430 and 0.8205, respectively for the *subset distortions* category. Thus, there is a significant gap in EONSS performance between these two categories. As noted earlier, the distortions contained in the *subset distortions* category are more aligned with those in the Waterloo Exploration-II database, which is used to train EONSS. However, the *all distortions* category includes many distortions that are not part of this database (refer to Section 3.5.2 and Table 2.2 to see a list of distortions that are in the test set but not contained in the Waterloo Exploration-II database). Since, EONSS performs remarkably well for the *subset distortions* category, even though it is a BIQA method, we conclude that the drop in EONSS performance in the *all distortions* category can be attributed to the absence of many distortion types in the Waterloo Exploration-II database, and if this gap is filled, then the performance of EONSS, and any other DNN based IQA models trained on this dataset, should go up. Thus, Waterloo Exploration-II database should be extended and diversified by including a host of distortion types not found in the current version. This process can begin by including distortion types that are found in available subject-rated datasets (see Table 2.2 for a list of such distortions), so that the same test set as in Tables 3.8 and 3.9 can be used, thereby enabling direct comparisons between the current and subsequent versions of EONSS. Since Waterloo Exploration-II database primarily includes multiply distorted images, the new dataset should include a large number of either singly distorted images belonging to diverse distortion types or multiply distorted images where the first distortion can belong to a new distortion type followed by JPEG compression. We believe that such a new dataset would lead to even more powerful DNN-based IQA models.

DNN-based FR IQA: In this thesis, we trained the DNN-based BIQA method EONSS on the Waterloo Exploration-II database as a means of validating our approach to address the data shortage challenge in IQA. Extensive tests revealed that EONSS outperforms the very state-of-the-art in BIQA, both in terms of perceptual quality prediction performance and execution time. Since, the Waterloo Exploration-II database has both pristine reference and distorted images, it can be used to train FR-based DNN models. While a few such models exist in the IQA literature, they have also been trained on small-scale IQA databases and suffer from the associated issues. Thus, it will be interesting to see how well an FR-based DNN model, trained on the large-scale Waterloo Exploration-II database, performs with respect to the very state-of-the-art in FR IQA, such as IWS-SIM [13] (which was found to be one of the best FR methods in Chapter 2).

Unified Framework for FR/DR/NR IQA: FR and NR IQA are independent paradigms that are unable to adapt based on the availability of additional information about a distorted image. Not only is DR IQA a major new paradigm, it also provides an opportunity to unify the major paradigms of IQA into one larger framework. Given a singly or multiply distorted image and its pristine or degraded reference, such a framework should assess the quality of the reference image and it should operate as an FR method if the quality of the reference is found to be pristine, operate as an NR method if the quality of the reference is found to be so distorted that it is of no help or if the reference image is missing, and operate as a DR method if the reference image is distorted but still useful for the task of quality assessment of the final distorted image. It should be noted that the idea for a unified framework for different IQA paradigms has been proposed earlier in the context of RR IQA [11], where the RR method could operate as an FR, RR, or NR method given the data rate available to transmit RR features. Yet, not much work has been done on such a framework, which can be attributed to the lack of powerful learning tools in the past. However, with the availability of computational resources and large-scale annotated training data, learning such an adaptable framework becomes a feasible task. One potential approach is end-to-end learning utilizing DNNs. Given the availability of the large-scale Waterloo Exploration-II database, which has degraded references of varying quality, it is possible to train such a DNN-based model to accomplish this task. We believe that the practical applicability of such models makes this an exciting research direction.

Further Development of two-stage DR IQA: Tables 4.7, 4.8, and 4.9, show that the DR IQA models developed for a two-stage distortion pipeline in this thesis do considerably well for the B-JPG, B-N, and JPG-JPG distortion combinations. While their performance is satisfactory for the more difficult N-JPG, N-JP2, NBJ-JPG, and all data cases, there is significant room for improvement, especially for the all data case. Since we have employed both distortion behavior based and SVR-based approaches to combine $\mathrm{AS}_{\mathrm{DR}}/\widehat{\mathrm{AS}_{\mathrm{DR}}}$ and $\mathrm{RS}_{\mathrm{FD}}$ scores generated by different IQA methods, alternative DR IQA design philosophies that not only use objective quality scores of the degraded reference and final distorted images, but also additional features need to be investigated. Future DR IQA models are desired to be general-purpose, i.e., not specific to a particular distortion combination, but applicable to a wide variety of multiple distortion combinations. One possibility is to use NSS [123] based features as earlier works that developed wavelet-domain based NSS models for RR IQA [121, 228] have shown that different distortion types affect the wavelet coefficient distributions in uniquely different manners, thereby affording an opportunity for IQA. Similarly, in [140], it was shown that NSS-based MSCN coefficients are also uniquely affected by different distortion types. However, these earlier studies have focused on the single distortion case, and the impact of multiple distortions on NSS features has not been studied in detail. Thus, future studies should investigate the impact of multiple distortions on NSS features and develop new general-purpose DR IQA models that either rely solely on NSS features or are a hybrid of NSS features and objective scores of the degraded reference and final distorted images. Another possibility is to learn general-purpose DR IQA models in a truly end-to-end manner by employing DNNs. The large-scale Waterloo Exploration-II database developed in this work can provide an adequate amount of training data for such an endeavor.

Development of DR IQA for Other Application Scenarios: It has been discussed in Section 4.3 that since the interaction of even two simultaneous distortions has not been analyzed in depth, a logical point to start work on DR IQA is with a two-stage distortion pipeline and simulated distortion images, which is what has been done in this thesis and we have established a baseline against which future models can be compared. However, in practical media distribution systems, images may undergo more than two distortion stages and even the source image may be authentically distorted. Thus, future DR IQA models should be able to handle images undergoing more than two distortion stages. It has been shown in this thesis that, in a two-stage distortion pipeline, using the DR image is beneficial in determining the quality of an FD image. However, in a distortion pipeline consisting of more than two stages, multiple DR images may be available. Thus, the construction of DR IQA models that make use of multiple DR images needs to explored so that it can be determined if using multiple DR images is beneficial to determine the quality of the FD image. The case of the source image being authentically distorted should also be explored further. While we showed that the DR models developed in this thesis perform well on the LIVE Wild Compressed database [2], which consists of authentically distorted DR images, further work needs to be done as this is a small-scale dataset (80 DR and 320 FD images).

Construction of a Large-Scale VQA Database: The work in this thesis has focused on *image quality assessment* (IQA). Similar approaches may be extended to video quality assessment (VQA). The subjective testing of videos takes even more time than such testing of images, which means that contemporary annotated VQA databases are even smaller in size than their IQA counterparts. The availability of large-scale VQA datasets will enhance the development of machine learning based models in that area, much like the Waterloo Exploration-II database has positively impacted the creation of DNN based IQA methods. Thus, a new large-scale VQA database, with thousands of pristine and hundreds of thousands of both singly and multiply distorted videos, belonging to a diverse set of distortion types, should be constructed. It can then be synthetically annotated much like SQB is used to annotate the Waterloo Exploration-II database. As a first step, the SQB version developed in this thesis can be applied in a frame-by-frame manner to videos and the average SQB value of all video frames can be used to annotate it. Since SQB in its current shape will only be able to cater to the spatial aspect of video, its video-specific versions should be developed that also take into account the temporal aspect. The development of such a synthetically annotated dataset would enable the development of powerful DNN based VQA models, especially *blind* VQA models, whose availability is highly desirable given the very high global usage of videos [8].

Degraded Reference Video Quality Assessment (DR VQA): The DR IQA paradigm developed in this thesis is directly extendable to DR VQA. In practical video distribution systems, such as streaming media platforms, multiple rounds of distortions are commonplace. This usually involves an original content which is either pristine (in case of high end production houses) or afflicted with noise, blur, compression, or color distortions (in case of amateur video), undergoing subsequent rounds of compression during distribution. Thus, DR VQA focusing on distortion combinations of *Noise-Compression*, *Blur-Compression*, and *Compression-Compression*, can be pursued. With a number of video coding methods available, different kinds of compression techniques can lead to even more practically occurring multiple distortion combinations. Development of practically applicable DR VQA models can lead to the distribution of video at defined quality thresholds in a bandwidth efficient manner.

References

[1] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, “Predicting the Quality of Images Compressed After Distortion in Two Steps,” in *Proc. SPIE Opt. Eng. Appl.*, vol. 10752, San Diego, CA, USA, Sept. 2018, pp. 107 520K:1–107 520K:8.

[2] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, “Predicting the Quality of Images Compressed After Distortion in Two Steps,” *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5757–5770, Dec. 2019.

[3] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “Completely Blind” Image Quality Analyzer,” *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[4] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2003, pp. 1398–1402.

[5] X. Liu, M. Pedersen, and J. Y. Hardeberg, “CID:IQ – A New Image Quality Database,” in *Proc. Int. Conf. Image, Signal Process. (ICISP)*, Cherbourg, France, July 2014, pp. 193–202.

[6] J. Ballé, V. Laparra, and E. P. Simoncelli, “End-to-end optimized image compression,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017.

[7] “YouTube for Press.” [Online]. Available: https://www.youtube.com/about/press/

[8] Cisco and/or its affiliates, “Cisco Visual Networking Index: Forecast and Trends, 2017–2022,” *White Paper, Cisco Public Information*, 2019.

[9] Z. Wang and A. C. Bovik, "Mean Squared Error: Love It or Leave It? A new look at Signal Fidelity Measures," *IEEE Signal Process. Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[10] "The 67th engineering Emmy awards." [Online]. Available: https://www.emmys.com/news/press-releases/honorees-announced-67th-engineering-emmy-awards

[11] Z. Wang and A. C. Bovik, "Modern Image Quality Assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, 2006, Morgan & Claypool Publishers.

[12] Z. Wang and A. C. Bovik, "Reduced- and No-Reference Image Quality Assessment," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 29–40, Nov. 2011.

[13] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[15] L. Zhang, Y. Shen, and H. Li, "VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Oct. 2014.

[16] A. Balanov, A. Schwartz, Y. Moshe, and N. Peleg, "Image quality assessment based on DCT subband similarity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sept. 2015, pp. 2105–2109.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[18] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, June 2009, pp. 248–255.

[19] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process.: Image Commun.*, vol. 30, pp. 57–77, Jan. 2015.

[20] T. Liu, H. Liu, S. Pei, and K. Liu, "A High-Definition Diversity-Scene Database for Image Quality Assessment," *IEEE Access*, vol. 6, pp. 45 427–45 438, 2018.

[21] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo Exploration Database: New Challenges for Image Quality Assessment Models," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[22] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A Large-scale Artificially Distorted IQA Database," in *Proc. Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Berlin, Germany, June 2019, pp. 1–3.

[23] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Boston, MA, USA, July 2009, pp. 758–759.

[24] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[25] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "TID2008–A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[26] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 011 006:1–011 006:21, Jan. 2010.

[27] "A57 Image Quality Database," Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=26.

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

[29] Y. Horita, K. Shibata, and K. Yoshikazu, "MICT Image Quality Evaluation Database," http://mict.eng.u-toyama.ac.jp/mictdb.html.

[30] P. L. Callet and F. Autrusseau, "Subjective quality assessment IRCCyN/IVC database," 2005, Available: http://www2.irccyn.ec-nantes.fr/ivcdb/.

[31] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. Asilomar Conf. Signals, Syst., Comput. (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2012, pp. 1693–1697.

[32] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Hybrid No-Reference Quality Metric for Singly and Multiply Distorted Images," *IEEE Trans. Broadcast.*, vol. 60, no. 3, pp. 555–567, Sept. 2014.

[33] W. Sun, F. Zhou, and Q. Liao, "MDID: A multiply distorted image database for image quality assessment," *Pattern Recognit.*, vol. 61, pp. 153–168, Jan. 2017.

[34] S. Corchs and F. Gasparini, "A Multidistortion Database for Image Quality," in *Proc. Int. Workshop Comput. Color Imag. (CCIW)*, Milan, Italy, Mar. 2017, pp. 95–104.

[35] W. Xue, L. Zhang, and X. Mou, "Learning without Human Scores for Blind Image Quality Assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, June 2013, pp. 995–1002.

[36] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, Aug. 2017.

[37] Y. Zhang and D. M. Chandler, “Opinion-Unaware Blind Quality Assessment of Multiply and Singly Distorted Images via Distortion Parameter Estimation,” *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5433–5448, Nov. 2018.

[38] J. Kim and S. Lee, “Deep blind image quality assessment by employing FR-IQA,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sept. 2017, pp. 3180–3184.

[39] J. Kim and S. Lee, “Fully Deep Blind Image Quality Predictor,” *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 206–220, Feb. 2017.

[40] S. Athar, A. Rehman, and Z. Wang, “Quality assessment of images undergoing multiple distortion stages,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sept. 2017, pp. 3175–3179.

[41] P. Ye, J. Kumar, and D. Doermann, “Beyond Human Opinion Scores: Blind Image Quality Assessment based on Synthetic Scores,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, June 2014, pp. 4241–4248.

[42] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “LIVE Image Quality Assessment Database Release 2,” Available: http://live.ece.utexas.edu/research/Quality/subjective.htm.

[43] M. Pedersen and J. Y. Hardeberg, “Survey of full-reference image quality metrics,” Høgskolen i Gjøviks Rapportserie 5, Norwegian Color Res. Lab., Gjøvik Univ. College Norway, Tech. Rep., June 2009.

[44] W. Lin and C.-C. J. Kuo, “Perceptual visual quality metrics: A survey,” *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297 – 312, May 2011.

[45] M. Pedersen and J. Y. Hardeberg, “Full-Reference Image Quality Metrics: Classification and Evaluation,” *Found. Trends Comput. Graph. Vis.*, vol. 7, no. 1, pp. 1–80, Mar. 2012.

[46] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Sept. 2012, pp. 1477–1480.

[47] P. Mohammadi, A. Ebrahimi-Moghadam, and S. Shirani, "Subjective and Objective Quality Assessment of Image: A Survey," *arXiv preprint arXiv:1406.7799*, June 2014.

[48] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb. 2013.

[49] L. He, F. Gao, W. Hou, and L. Hao, "Objective image quality assessment: a survey," *Int. J. Comput. Math.*, vol. 91, no. 11, pp. 2374–2388, 2014.

[50] M. Pedersen, "Evaluation of 60 full-reference image quality metrics on the CID:IQ," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sept. 2015, pp. 1588–1592.

[51] R. A. Manap and L. Shao, "Non-distortion-specific no-reference image quality assessment: A survey," *Inf. Sci.*, vol. 301, pp. 141 – 160, Apr. 2015.

[52] S. Xu, S. Jiang, and W. Min, "No-reference/Blind Image Quality Assessment: A Survey," *IETE Tech. Rev.*, vol. 34, no. 3, pp. 223–245, 2017.

[53] Y. Niu, Y. Zhong, W. Guo, Y. Shi, and P. Chen, "2D and 3D Image Quality Assessment: A Survey of Metrics and Challenges," *IEEE Access*, vol. 7, pp. 782–801, 2019.

[54] A. Zarić, N. Tatalović, N. Brajković, H. Hlevnjak, M. Lončarić, E. Dumić, and S. Grgić, "VCL@FER Image Quality Assessment Database," *AUTOMATIKA*, vol. 53, no. 4, pp. 344–354, 2012.

[55] H. Yeganeh, M. Rostami, and Z. Wang, "Objective Quality Assessment of Interpolated Natural Images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4651–4663, Nov. 2015.

[56] R. M. Nasiri and Z. Wang, “Perceptual aliasing factors and the impact of frame rate on video quality,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, Sept. 2017, pp. 3475–3479.

[57] R. M. Nasiri, Z. Duanmu, and Z. Wang, “Temporal Motion Smoothness and the Impact of Frame Rate Variation on Video Quality,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 1418–1422.

[58] K. Ma, K. Zeng, and Z. Wang, “Perceptual Quality Assessment for Multi-Exposure Image Fusion,” *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[59] R. Hassen, Z. Wang, and M. M. A. Salama, “Objective Quality Assessment for Multiexposure Multifocus Image Fusion,” *IEEE Trans. Image Process.*, vol. 24, no. 9, pp. 2712–2724, Sept. 2015.

[60] K. Ma, T. Zhao, K. Zeng, and Z. Wang, “Objective Quality Assessment for Color-to-Gray Image Conversion,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4673–4685, Dec. 2015.

[61] A. Rehman, K. Zeng, and Z. Wang, “Display Device-Adapted Video Quality-of-Experience Assessment,” in *Proc. SPIE Electron. Imag.*, vol. 9394, San Francisco, CA, USA, Mar. 2015, pp. 939 406:1–939 406:11.

[62] “Tampere Image Database 2013 (TID2013) Version 1.0,” Available: http://www.ponomarenko.info/tid2013.htm.

[63] E. C. Larson and D. M. Chandler, “Computational and Subjective Image Quality (CSIQ) database,” Available: http://vision.eng.shizuoka.ac.jp/mod/page/view.php?id=23.

[64] “VCL@FER image quality assessment database,” Available: http://www.vcl.fer.hr/quality/vclfer.html.

[65] “Colourlab Image Database: Image Quality (CID:IQ),” Available: https://www.ntnu.edu/web/colourlab/software.

[66] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "LIVE Multiply Distorted Image Quality Database," Available: http://live.ece.utexas.edu/research/Quality/live_multidistortedimage.html.

[67] "Multiply Distorted Image Database (MDID)," Available: http://www.sz.tsinghua.edu.cn/labs/vipl/mdid.html.

[68] S. Corchs, F. Gasparini, and R. Schettini, "Noisy images-JPEG compressed: subjective and objective image quality evaluation," in *Proc. SPIE Electron. Imag.*, vol. 9016, San Francisco, CA, USA, Feb. 2014, pp. 90 160V:1–90 160V:9.

[69] "Multiply Distorted Database MD-IVL," Available: http://www.mmsp.unimib.it/image-quality/.

[70] Rec. ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[71] Rec. ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," Apr. 2008.

[72] Video Quality Experts Group and others, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, Phase II," 2003.

[73] S. Tourancheau, F. Autrusseau, Z. M. P. Sazzad, and Y. Horita, "Impact of subjective dataset on the performance of image quality metrics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, San Diego, CA, USA, Oct. 2008, pp. 365–368.

[74] International Commission on Illumination (CIE), "Guidelines for the evaluation of gamut mapping algorithms," 2004.

[75] D. M. Chandler and S. S. Hemami, "VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sept. 2007.

[76] "Tampere Image Database 2008 (TID2008) Version 1.0," Available: http://www.ponomarenko.info/tid2008.htm.

[77] A. Ciancio, A. L. N. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-Reference Blur Assessment of Digital Pictures Based on Multifeature Classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[78] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen, "CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 390–402, Jan. 2015.

[79] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

[80] "Amazon Mechanical Turk," https://www.mturk.com/.

[81] H. Lin, V. Hosu, and D. Saupe, "KonIQ-10K: Towards an ecologically valid and large-scale IQA database," *arXiv preprint arXiv:1803.08489*, 2018.

[82] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The New Data in Multimedia Research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[83] "Figure Eight," https://www.figure-eight.com/.

[84] H. Yang, Y. Fang, and W. Lin, "Perceptual Quality Assessment of Screen Content Images," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4408–4421, Nov. 2015.

[85] J. Kumar, P. Ye, and D. Doermann, "A Dataset for Quality Assessment of Camera Captured Document Images," in *Proc. Int. Workshop Camera-Based Document Anal. Recognit. (CBDAR)*, Washington, DC, USA, Aug. 2013, pp. 113–125.

[86] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Real-Time No-Reference Image Quality Assessment Based on Filter Learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Portland, OR, USA, June 2013, pp. 987–994.

[87] S. Winkler, “Image and Video Quality Resources,” Available: https://stefan.winkler.site/resources.html.

[88] S. Winkler, “Analysis of Public Image and Video Databases for Quality Assessment,” *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[89] H. Yu and S. Winkler, “Image complexity and spatial information,” in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Klagenfurt am Wörthersee, Austria, July 2013, pp. 12–17.

[90] D. Hasler and S. Suesstrunk, “Measuring colorfulness in natural images,” in *Proc. SPIE Electron. Imag.*, vol. 5007, Santa Clara, CA, USA, June 2003, pp. 87–95.

[91] M. Buczkowski and R. Stasiński, “Effective Coverage as a New Metric for Image Quality Assessment Databases Comparison,” in *Proc. Int. Conf. Syst., Signals, Image Process. (IWSSIP)*, Poznan, Poland, May 2017, pp. 1–5.

[92] Z. Wang, “Objective Image Quality Assessment: Facing the Real-World Challenges,” in *Proc. IS&T Int. Symp. Electron. Imag.*, vol. 2016, no. 13, San Francisco, CA, USA, Feb. 2016, pp. 1–6.

[93] S. Rezazadeh and S. Coulombe, “A novel discrete wavelet transform framework for full reference image quality assessment,” *Signal, Image, Video Process.*, vol. 7, no. 3, pp. 559–573, May 2013.

[94] S. Li, F. Zhang, L. Ma, and K. N. Ngan, “Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments,” *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[95] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker, “Image-Difference Prediction: From Grayscale to Color,” *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 435–446, Feb. 2013.

[96] E. D. Di Claudio and G. Jacovitti, “A Detail-Based Method for Linear Full Reference Image Quality Prediction,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 179–193, Jan. 2018.

[97] S. Rezazadeh and S. Coulombe, "Low-complexity computation of visual information fidelity in the discrete wavelet domain," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, USA, Mar. 2010, pp. 2438–2441.

[98] X. Zhang, X. Feng, W. Wang, and W. Xue, "Edge Strength Similarity for Image Quality Assessment," *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 319–322, Apr. 2013.

[99] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[100] A. Liu, W. Lin, and M. Narwaria, "Image Quality Assessment Based on Gradient Similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[101] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An Information Fidelity Criterion for Image Quality Assessment Using Natural Scene Statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[102] T. Wang, L. Zhang, H. Jia, B. Li, and H. Shu, "Multiscale contrast similarity deviation: An effective and efficient index for perceptual image quality assessment," *Signal Process.: Image Commun.*, vol. 45, pp. 1–9, July 2016.

[103] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image Quality Assessment Based on a Degradation Model," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[104] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, and M. Carli, "Modified Image Visual Quality Metrics for Contrast Change and Mean Shift Accounting," in *Proc. 11th Int. Conf. Exper. Designing Appl. CAD Syst. Microelectron. (CADSM)*, Polyana-Svalyava, Ukraine, Feb. 2011, pp. 305–311.

[105] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, vol. 4, Scottsdale, AZ, USA, Jan. 2006.

[106] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On Between-Coefficient Contrast Masking of DCT Basis Functions," in *Proc. 3rd Int. Workshop Video Process., Qual. Metrics Consum. Electron. (VPQM)*, vol. 4, Scottsdale, AZ, USA, Jan. 2007.

[107] L. Li, H. Cai, Y. Zhang, W. Lin, A. C. Kot, and X. Sun, "Sparse Representation-Based Image Quality Index With Adaptive Sub-Dictionaries," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3775–3786, Aug. 2016.

[108] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Hong Kong, China, Sept. 2010, pp. 321–324.

[109] H. Chang, H. Yang, Y. Gan, and M. Wang, "Sparse Feature Fidelity for Perceptual Image Quality Assessment," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 4007–4018, Oct. 2013.

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

[111] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[112] Z. Wang and A. C. Bovik, "A Universal Image Quality Index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[113] H. R. Sheikh and A. C. Bovik, "Image Information and Visual Quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[114] H. R. Sheikh and A. C. Bovik, "Pixel domain version of VIF," 2005, Available: http://live.ece.utexas.edu/research/Quality/VIF.htm.

[115] S. Rezazadeh and S. Coulombe, “A novel approach for computing and pooling Structural SIMilarity index in the discrete wavelet domain,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2209–2212.

[116] U. Engelke, H. Kaprykowsky, H. Zepernick, and P. Ndjiki-Nya, “Visual Attention in Quality Assessment,” *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 50–59, Nov. 2011.

[117] A. K. Moorthy and A. C. Bovik, “Visual Importance Pooling for Image Quality Assessment,” *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[118] C. C. Yang and S. H. Kwok, “Efficient gamut clipping for color image processing using LHS and YIQ,” *Opt. Eng.*, vol. 42, no. 3, pp. 701–711, Mar. 2003.

[119] X. Hou and L. Zhang, “Saliency Detection: A Spectral Residual Approach,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Minneapolis, MN, USA, June 2007, pp. 1–8.

[120] L. Zhang, Z. Gu, and H. Li, “SDSP: A novel saliency detection method by combining simple priors,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, VIC, Australia, Sept. 2013, pp. 171–175.

[121] Z. Wang and E. P. Simoncelli, “Reduced-Reference Image Quality Assessment using a Wavelet-Domain Natural Image Statistic Model,” in *Proc. SPIE Electron. Imag.*, vol. 5666, San Jose, CA, USA, Mar. 2005, pp. 149–159.

[122] Q. Li and Z. Wang, “Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation,” *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[123] E. P. Simoncelli and B. A. Olshausen, “Natural Image Statistics and Neural Representation,” *Annu. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, Mar. 2001.

[124] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random Cascades on Wavelet Trees and Their Use in Analyzing and Modeling Natural Images," *Appl. Comput. Harmon. Anal.*, vol. 11, no. 1, pp. 89–123, Jul. 2001.

[125] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: A flexible architecture for multi-scale derivative computation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 3, Washington, DC, USA, Oct. 1995, pp. 444–447.

[126] K. Okarma, "Combined Image Similarity Index," *Opt. Rev.*, vol. 19, no. 5, pp. 349–354, Sept. 2012.

[127] K. Okarma, "Quality Assessment of Images with Multiple Distortions using Combined Metrics," *Elektronika Ir Elektrotechnika*, vol. 20, no. 6, pp. 128–131, 2014.

[128] V. V. Lukin, N. N. Ponomarenko, O. I. Ieremeiev, K. O. Egiazarian, and J. Astola, "Combining full-reference image visual quality metrics by neural network," in *Proc. SPIE Electron. Imag.*, vol. 9394, San Francisco, CA, USA, Mar. 2015, pp. 93 940K:1–93 940K:12.

[129] K. Okarma, "Hybrid Feature Similarity Approach to Full-Reference Image Quality Assessment," in *Proc. Int. Conf. Comput. Vis. Graph. (ICCVG)*, Warsaw, Poland, Sept. 2012, pp. 212–219.

[130] T. Liu and W. Lin and C.-C. J. Kuo, "Image Quality Assessment Using Multi-Method Fusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1793–1807, May 2013.

[131] K. Okarma, "Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Assessment," in *Proc. Int. Conf. AI. Soft Comput. (ICAISC)*, Zakopane, Poland, June 2010, pp. 539–546.

[132] K. Okarma, "Extended Hybrid Image Similarity–Combined Full-Reference Image Quality Metric Linearly Correlated with Subjective Scores," *Elektronika ir Elektrotechnika*, vol. 19, no. 10, pp. 129–132, 2013.

[133] T. Liu, W. Lin, and C.-C. J. Kuo, “A multi-metric fusion approach to visual quality assessment,” in *Proc. Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Mechelen, Belgium, Sept. 2011, pp. 72–77.

[134] C. W. Hsu, C. C. Chang, and C. J. Lin, “A Practical Guide to Support Vector Classification,” 2003 (Last updated: May 2016), Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

[135] N. Ponomarenko, O. Ieremeiev, V. Lukin, L. Jin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, “A New Color Image Database TID2013: Innovations and Results,” in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst. (ACVIS)*, Poznan, Poland, Oct. 2013, pp. 402–413.

[136] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, “A no-reference perceptual blur metric,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 3, Rochester, NY, USA, Sept. 2002, pp. III:57–III:60.

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

[138] H. R. Sheikh, A. C. Bovik, and L. Cormack, “No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000,” *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[139] A. K. Moorthy and A. C. Bovik, “A Two-Step Framework for Constructing Blind Image Quality Indices,” *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[140] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-Reference Image Quality Assessment in the Spatial Domain,” *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[141] P. Ye, J. Kumar, L. Kang, and D. Doermann, “Unsupervised Feature Learning Framework for No-Reference Image Quality Assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, June 2012, pp. 1098–1105.

[142] Q. Li, W. Lin, and Y. Fang, “No-Reference Quality Assessment for Multiply-Distorted Images in Gradient Domain,” *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 541–545, Apr. 2016.

[143] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, “Blind Image Quality Assessment Based on High Order Statistics Aggregation,” *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sept. 2016.

[144] L. Zhang, L. Zhang, and A. C. Bovik, “A Feature-Enriched Completely Blind Image Quality Evaluator,” *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[145] Q. Wu, Z. Wang, and H. Li, “A highly efficient method for blind image quality assessment,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sept. 2015, pp. 339–343.

[146] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, “End-to-End Blind Image Quality Assessment Using Deep Neural Networks,” *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[147] Q. Li, W. Lin, J. Xu, and Y. Fang, “Blind Image Quality Assessment Using Statistical Structural and Luminance Features,” *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2457–2469, Dec. 2016.

[148] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, “Deep Neural Networks for No-Reference and Full-Reference Image Quality Assessment,” *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[149] V. N. Vapnik, *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 2000.

[150] I. Daubechies, *Ten Lectures on Wavelets.* Philadelphia, PA, USA: SIAM, 1992, vol. 61.

[151] K. Sharifi and A. Leon-Garcia, "Estimation of Shape Parameter for Generalized Gaussian Distributions in Subband Decompositions of Video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, Feb. 1995.

[152] N. Lasmar, Y. Stitou, and Y. Berthoumieu, "Multiscale skewed heavy tailed model for texture analysis," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2281–2284.

[153] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, Apr. 2011.

[154] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, July 2002.

[155] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sept. 2016, pp. 3773–3777.

[156] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[157] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Haifa, Israel, June 2010, pp. 807–814.

[158] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional Neural Networks for No-Reference Image Quality Assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, June 2014, pp. 1733–1740.

[159] W. Hou, X. Gao, D. Tao, and X. Li, "Blind Image Quality Assessment via Deep Learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, June 2015.

[160] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sept. 2015, pp. 2791–2795.

[161] J. Fu, H. Wang, and L. Zuo, "Blind image quality assessment for multiply distorted images via convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 1075–1079.

[162] J. Li, L. Zou, J. Yan, D. Deng, T. Qu, and G. Xie, "No-reference image quality assessment using Prewitt magnitude based on convolutional neural networks," *Signal, Image, Video Process.*, vol. 10, no. 4, pp. 609–616, Apr. 2016.

[163] J. Li, J. Yan, D. Deng, W. Shi, and S. Deng, "No-reference image quality assessment based on hybrid model," *Signal, Image, Video Process.*, vol. 11, no. 6, pp. 985–992, Sept. 2017.

[164] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432 – 442, Sept. 2018.

[165] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image, Video Process. (SIViP)*, vol. 12, no. 2, pp. 355–362, Feb. 2018.

[166] H. Talebi and P. Milanfar, "NIMA: Neural Image Assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[167] H. Zeng, L. Zhang, and A. C. Bovik, "Blind Image Quality Assessment with a Probabilistic Quality Representation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 609–613.

[168] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-Based Blind Image Quality Predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.

[169] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Vancouver, BC, Canada, July 2001, pp. 416–423.

[170] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to Rank using Gradient Descent," in *Proc. Int. Conf. Mach. Learn (ICML)*, Bonn, Germany, Aug. 2005, pp. 89–96.

[171] K. Gu, G. Zhai, M. Liu, X. Yang, W. Zhang, X. Sun, W. Chen, and Y. Zuo, "FISBLIM: A FIve-Step BLInd Metric for quality assessment of multiply distorted images," in *Proc. IEEE Workshop Signal Process. Syst. (SiPS)*, Taipei City, Taiwan, Oct. 2013, pp. 241–246.

[172] D. Zoran and Y. Weiss, "Scale invariance and noise in natural images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Kyoto, Japan, Sept. 2009, pp. 2209–2216.

[173] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[174] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A Psychovisual Quality Metric in Free-Energy Principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[175] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman & Hall/CRC, 2011.

[176] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[177] A. K. Moorthy and A. C. Bovik, "Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[178] D. Ghadiyaram and A. C. Bovik, “Feature Maps-Driven No-Reference Image Quality Prediction of Authentically Distorted Images,” in *Proc. SPIE Electron. Imag.*, vol. 9394, San Francisco, CA, USA, Mar. 2015, pp. 93 940J:1–93 940J:14.

[179] D. Ghadiyaram and A. C. Bovik, “Perceptual Quality Prediction on Authentically Distorted Images Using a Bag of Features Approach,” *J. Vis.*, vol. 17, no. 1, pp. 32:1–32:25, Jan. 2017.

[180] H. Hadizadeh and I. V. Bajić, “Color Gaussian Jet Features For No-Reference Quality Assessment of Multiply-Distorted Images,” *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1717–1721, Dec. 2016.

[181] C. Li, Y. Zhang, X. Wu, and Y. Zheng, “A Multi-Scale Learning Local Phase and Amplitude Blind Image Quality Assessment for Multiply Distorted Images,” *IEEE Access*, vol. 6, pp. 64 577–64 586, 2018.

[182] K. Gu, G. Zhai, X. Yang, and W. Zhang, “Using Free Energy Principle For Blind Image Quality Assessment,” *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Jan. 2015.

[183] Q. Wu, H. Li, F. Meng, K. N. Ngan, B. Luo, C. Huang, and B. Zeng, “Blind Image Quality Assessment Based on Multichannel Feature Fusion and Label Transfer,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 425–440, Mar. 2016.

[184] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, “Deep Convolutional Neural Models for Picture-Quality Prediction: Challenges and Solutions to Data-Driven Image Quality Assessment,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[185] K. Ma, Q. Wu, Z. Wang, Z. Duanmu, H. Yong, H. Li, and L. Zhang, “Group MAD Competition – A New Methodology to Compare Objective Image Quality Models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 1664–1673.

[186] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, “Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features,” *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[187] X. Yang, F. Li, and H. Liu, “A Survey of DNN Methods for Blind Image Quality Assessment,” *IEEE Access*, vol. 7, pp. 123 788–123 806, Sept. 2019.

[188] A. Torralba, R. Fergus, and W. T. Freeman, “80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.

[189] “ImageNet,” Available: http://image-net.org/index.

[190] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[191] C. Fellbaum, *WordNet: An Electronic Lexical Database.* MIT press, 1998.

[192] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[193] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sept. 2014, pp. 818–833.

[194] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, Banff, AB, Canada, Apr. 2014.

[195] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional Architecture for Fast Feature Embedding,” in *Proc. ACM Int. Conf. Multimedia (MM)*, Orlando, FL, USA, Nov. 2014, pp. 675–678.

[196] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, June 2015, pp. 1–9.

[197] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[198] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Washington, DC, USA, June 2004, pp. 178–178.

[199] G. Griffin, A. Holub, and P. Perona, "Caltech-256 Object Category Dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. 7694, Apr. 2007.

[200] X. Qin, T. Xiang, Y. Yang, and X. Liao, "Pair-Comparing Based Convolutional Neural Network for Blind Image Quality Assessment," in *Int. Symp. Neural Netw. (ISNN)*, Moscow, Russia, Jul. 2019, pp. 460–468.

[201] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning Deep Features for Scene Recognition using Places Database," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 487–495.

[202] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Columbus, OH, USA, June 2014, pp. 512–519.

[203] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Nottingham, UK, Sept. 2014, pp. 1–12.

[204] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.

[205] C. Zhang and K. Hirakawa, "Blind full reference quality assessment of poisson image denoising," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 2719–2723.

[206] X. Jin and K. Hirakawa, “Approximations to Camera Sensor Noise,” in *Proc. SPIE Electron. Imag.*, vol. 8655, Burlingame, CA, USA, Feb. 2013, pp. 86 550H:1–86 550H:7.

[207] W. Liu, Z. Duanmu, and Z. Wang, “End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks,” in *Proc. ACM Int. Conf. Multimedia*, Seoul, Republic of Korea, Oct. 2018, p. 546–554.

[208] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 1026–1034.

[209] D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[210] “Youtube Help: Recommended upload encoding settings.” [Online]. Available: https://support.google.com/youtube/answer/1722171?hl=en

[211] “Facebook Help Center: How can I make sure that my photos display in the highest possible quality?” [Online]. Available: https://www.facebook.com/help/266520536764594

[212] H. Gudbjartsson and S. Patz, “The Rician Distribution of Noisy MRI Data,” *Magn. Reson. Med.*, vol. 34, no. 6, pp. 910–914, 1995.

[213] H. Lu, X. Li, I. Hsiao, and Z. Liang, “Analytical Noise Treatment for Low-Dose CT Projection Data by Penalized Weighted Least-Square Smoothing in the K-L Domain,” in *Proc. SPIE Med. Imag.*, vol. 4682, San Diego, CA, USA, May 2002, pp. 146–152.

[214] P. Coupe, P. Hellier, C. Kervrann, and C. Barillot, “Nonlocal Means-Based Speckle Filtering for Ultrasound Images,” *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2221–2229, Oct. 2009.

[215] D. A. Koff and H. Shulman, “An Overview of Digital Compression of Medical Images: Can We Use Lossy Image Compression in Radiology?” *J. Can. Assoc. Radiologists*, vol. 57, no. 4, pp. 211–217, Oct. 2006.

[216] D. Koff, P. Bak, P. Brownrigg, D. Hosseinzadeh, A. Khademi, A. Kiss, L. Lepanto, T. Michalak, H. Shulman, and A. Volkening, “Pan-Canadian Evaluation of Irreversible Compression Ratios (“Lossy” Compression) for Development of National Guidelines,” *J. Digit. Imag.*, vol. 22, no. 6, pp. 569–578, Dec. 2009.

[217] R. L. White, “High-Performance Compression of Astronomical Images,” in *NASA. Goddard Space Flight Center, The Space and Earth Science Data Compression Workshop*, Jan. 1993, pp. 117–123.

[218] N. W. Lewis and J. W. Allnatt, “Subjective quality of television pictures with multiple impairments,” *IET Electron. Lett.*, vol. 1, no. 7, pp. 187–188, Sept. 1965.

[219] D. M. Chandler, “Seven Challenges in Image Quality Assessment: Past, Present, and Future Research,” *ISRN Signal Process.*, vol. 2013, Article ID 905685, pp. 1–53, 2013.

[220] Y. Lu, F. Xie, T. Liu, Z. Jiang, and D. Tao, “No Reference Quality Assessment for Multiply-Distorted Images Based on an Improved Bag-of-Words Model,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1811–1815, Oct. 2015.

[221] C. Li, Y. Zhang, X. Wu, W. Fang, and L. Mao, “Blind multiply distorted image quality assessment using relevant perceptual features,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sept. 2015, pp. 4883–4886.

[222] W. Cheng and K. Hirakawa, “Corrupted reference image quality assessment,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Orlando, FL, USA, Sept. 2012, pp. 1485–1488.

[223] C. Zhang, W. Cheng, and K. Hirakawa, “Corrupted Reference Image Quality Assessment of Denoised Images,” *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1732–1747, Apr. 2019.

[224] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, “LIVE Wild Compressed Picture Quality Database,” 2018, Available: http://live.ece.utexas.edu/research/twostep/index.html.

[225] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of Subjective and Objective Quality Assessment of Video,” *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[226] “The Consumer Digital Video Library (CDVL).” [Online]. Available: http://www.cdvl.org/

[227] C. C. Chang and C. J. Lin, “LIBSVM – A Library for Support Vector Machines,” Available: https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[228] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, “Quality-Aware Images,” *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, June 2006.