Blind Image Quality Assessment: Exploiting New Evaluation and Design Methodologies

by

Kede Ma

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, 2017

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	NAME	David J. Fleet
	Title	Professor
Supervisor(s)	NAME	Zhou Wang
	Title	Professor
Internal Member	NAME	En-hui Yang
	Title	Professor
Internal Member	NAME	Patrick Mitran
	Title	Associate Professor
Internal-external Member	NAME	David Clausi
	Title	Professor

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

The great content diversity of real-world digital images poses a grand challenge to automatically and accurately assess their perceptual quality in a timely manner. In this thesis, we focus on blind image quality assessment (BIQA), which predicts image quality with no access to its pristine quality counterpart. We first establish a large-scale IQA database—the Waterloo Exploration Database. It contains $4,744$ pristine natural and $94,880$ distorted images, the largest in the IQA field. Instead of collecting subjective opinions for each image, which is extremely difficult, we present three test criteria for evaluating objective BIQA models: pristine/distorted image discriminability test (D-test), listwise ranking consistency test (L-test), and pairwise preference consistency test (P-test). Moreover, we propose a general psychophysical methodology, which we name the group MAximum Differentiation (gMAD) competition method, for comparing computational models of perceptually discriminable quantities. We apply gMAD to the field of IQA and compare 16 objective IQA models of diverse properties. Careful investigations of selected stimuli shed light on how to improve existing models and how to develop next-generation IQA models. The gMAD framework is extensible, allowing future IQA models to be added to the competition.

We explore novel approaches for BIQA from two different perspectives. First, we show that a vast amount of reliable training data in the form of quality-discriminable image pairs (DIPs) can be obtained automatically at low cost. We extend a pairwise learning-to-rank (L2R) algorithm to learn BIQA models from millions of DIPs. Second, we propose a multi-task deep neural network for BIQA. It consists of two sub-networks—a distortion identification network and a quality prediction network—sharing the early layers. In the first stage, we train the distortion identification sub-network, for which large-scale training samples are readily available. In the second stage, starting from the pre-trained early layers and the outputs of the first sub-network, we train the quality prediction sub-network using a variant of stochastic gradient descent. Extensive experiments on four benchmark IQA databases demonstrate the proposed two approaches outperform state-of-the-art BIQA models. The robustness of learned models is also significantly improved as confirmed by the gMAD competition methodology.

Acknowledgements

I would like to thank all the people who made this thesis possible.

First and foremost, I wish to thank my adviser Professor Zhou Wang. Through more than 1,000 emails and 250 hours of meetings over the five years, Zhou taught me the fundamentals of good research, writing, and presentation. Zhou provided me with the full freedom to explore research topics that both of us would be interested in, and meanwhile incorporated his foresights and detailed suggestions into our solutions. It has been a privilege working with him. I am also honored to have Professor En-hui Yang, Professor Patrick Mitran, Professor David Clausi, and Professor David Fleet on my thesis committee.

Here at the Image and Vision Computing Lab, I've had the opportunity to work with many amazing labmates. These include Shiqi Wang, Yuming Fang, Kai Zeng, Abdul Rehman, Tiesong Zhao, Jiheng Wang, Hojatollah Yeganeh, Qingbo Wu, Wentao Liu, Xiongkuo Min, Shahrukh Athar, and Rasoul Mohammadi Nasiri. I am particularly grateful to Zhengfang Duanmu, who devoted tons of time to my projects and helped me a lot as a collaborator and a friend.

During my Ph.D. studies, I was fortunate to be advised by Professor Lei Zhang and Professor Dacheng Tao as a visiting student. Professor Lei Zhang introduced me to sparse and low-rank models for low-level vision. Professor Dacheng Tao introduced me to statistical learning theory and deep learning. Back then, I had the chance to collaborate with some of their wonderful students, including Hui Li, Hongwei Yong, Kai Zhang, Tongliang Liu, and Huan Fu. I've also had the fortune to learn from many other great students in their groups, including Shuhang Gu, Jun Xu, Sijia Cai, Dongwei Ren, Faqiang Wang, Mu Li, Ruxin Wang, Mingming Gong, Zhe Chen, Xiyu Yu, Maoying Qiao, and Long Lan.

I am grateful to my dear friends who have helped me proofread the thesis and given me so many constructive suggestions, including Dongwei Ren, Xiaoxi Zheng, Hui Li, Wufeng Xue, Tiesong Zhao, Yuming Fang, Ruihe Xiong, and Long Lan.

I also owe sincere gratitude to many people who had a positive impact on my life and made my Ph.D. journey unforgettable: Chong Lou, Qiang Ye, Chao Wu, Chaojie Ou, Wen Wu, Tianrong Rao, Lingxiang Wu, and many anonymous werewolf teammates.

Last but not least, I would like to thank my parents, Qingquan Ma and Wei Wu, for their love and support. I also wish to thank my girlfriend who appeared during my Ph.D. studies and made me more concentrated on my research. Tingting Lu, I love you from the very bottom of my heart.

Table of Contents

List of Tables

List of Figures

List of Acronyms

BIQA blind image quality assessment. 2–7, 9, 14, 16, 17, 19–21, 23, 26, 27, 29, 31, 33, 35, 36, 38–40, 49, 57, 59, 60, 66, 67, 69–74, 80, 83–86, 92, 93, 95, 102, 104–107

BN batch normalization. 88, 89, 103

BVQA blind video quality assessment. 107

CDF cumulative distribution function. 56, 57

D-test pristine/distorted image discriminability test. 3, 6, 27, 35, 36, 38, 40, 68, 71, 94, 95, 105, 106

DIL quality-discriminable image list. 5, 74, 77, 78

dilIQ DIL inferred quality. 5, 79, 80, 105

DIP quality-discriminable image pair. 3, 4, 6, 7, 31, 33, 39, 40, 59–61, 63, 64, 66, 67, 70, 71, 74, 78, 80, 105, 107

dipIQ DIP inferred quality. 4, 5, 19, 60, 64–67, 69–74, 79, 80, 95, 102, 105

DNN deep neural network. 5, 6, 17, 20, 21, 83, 84, 86–88, 91, 92, 96, 105, 107, 135

FR full-reference. 1, 7, 15, 16, 19, 20, 27, 31, 33, 49, 57, 61, 63, 66, 69, 80, 84, 86

GDN generalized divisive normalization. 5, 85, 87–90, 93, 102–104

GGD generalized Gaussian distribution. 17, 19

gMAD group MAximum Differentiation. 4–6, 43, 45–49, 52, 54, 56–58, 60, 66, 73, 86, 97, 99, 104–106, 133, 134, 137, 138, 140

HAS HTTP adaptive streaming. 137

HVS human visual system. 1, 15, 17, 83, 97

IQA image quality assessment. 1–4, 6, 7, 9, 10, 14–16, 19, 20, 23, 26, 27, 31, 33, 43, 44, 46, 49, 50, 53, 54, 56, 57, 60, 61, 63, 66, 67, 69, 80, 84, 86, 92, 93, 105, 106, 133

KRCC Kendall rank-order correlation coefficient. 2, 57

L-test listwise ranking consistency test. 3, 6, 27, 29, 36, 38–40, 68, 71, 94, 102, 105, 106

L2R learning-to-rank. 4–7, 59–61, 63, 65, 66, 73, 74, 80, 105–107

MAD MAximum Differentiation. 4, 43–45, 47, 50

MEON Multi-task End-to-end Optimized deep neural Network. 5, 7, 86, 91–93, 95–97, 99, 102, 104, 105

MOS mean opinion score. 2, 3, 6, 10, 12, 13, 16, 20, 21, 23, 31, 35, 38, 50, 51, 54, 57, 60, 61, 67, 68, 71–74, 83, 84, 89, 92, 95

MSE mean squared error. 1, 135

NR no-reference. 1

NSS natural scene statistics. 17, 38, 57, 95

OA opinion-aware. 60, 72–74

OU opinion-unaware. 4–7, 60, 66, 69–71, 74, 80, 107

P-test pairwise preference consistency test. 3, 6, 27, 31, 40, 61, 68, 71, 94, 105, 106

PLCC Pearson linear correlation coefficient. 2, 68, 70, 71, 94, 95

QoE quality of experience. 46, 106, 137–141

ReLU rectified linear unit. 5, 20, 67, 85, 87, 88, 103

RR reduced-reference. 1, 2

SRCC Spearman's rank-order correlation coefficient. 2, 68, 70, 94, 95, 102

SSIM structural similarity. 15, 49, 51, 52, 57, 69

std standard deviation. 26, 31, 51, 53, 54

SVM support vector machine. 20

SVR support vector regression. 16, 20, 134, 135

Chapter 1

Introduction

1.1 Motivations

With the explosion of digital image data, it becomes increasingly important to automatically and accurately assess image quality in a timely manner, which has a tremendous impact on monitoring, maintaining, and improving the perceived quality in various image acquisition, compression, transmission, processing, and display systems. Over the past twenty decades, there has been a remarkably increasing interest in developing image quality assessment (IQA) methods in both academia and industry [1]. The goal of IQA is to quantify the human perception of image quality, which may be degraded in various ways [120, 152, 153]. Since the human visual system (HVS) is the ultimate receiver in most visual applications, subjective testing is the most reliable way of quantifying image quality, but is also time-consuming, cumbersome, and expensive. Therefore, objective IQA that can automate this process becomes indispensable [75]. Objective IQA models can be broadly classified into full-reference (FR), reduced-reference (RR), and no-reference (NR)/blind methods based on their accessibility to the pristine-quality reference image. Specifically, FR-IQA methods assume full access to the reference image and some of them generate a map to indicate quality variations across spatial locations, as shown in Fig 1.1 [18]. Mean squared error (MSE), the dominant quantitative metric in the field of signal processing,

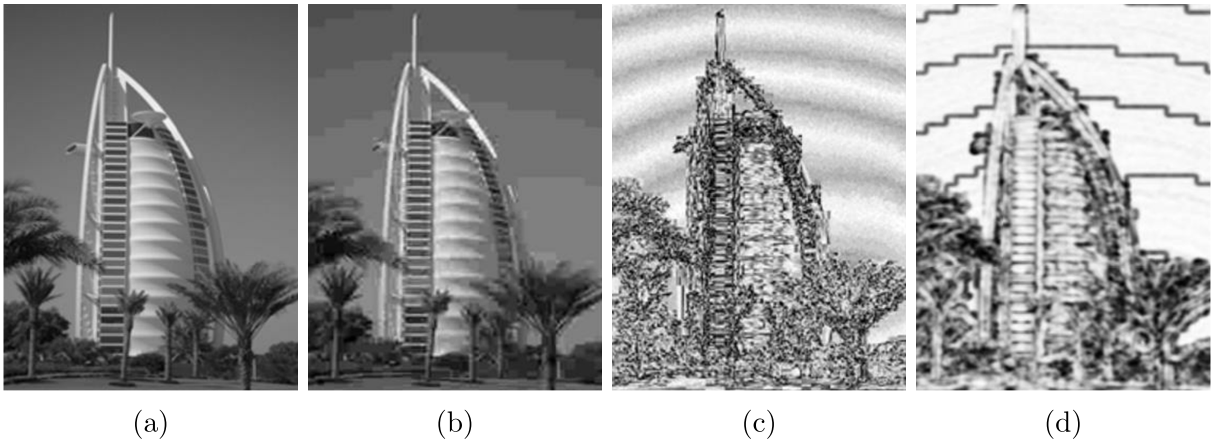


Figure 1.1: Quality maps generated by FR-IQA models. (a) Reference image. (b) JPEG compressed image. (c) Absolute error map of the distorted image (enhanced for visibility). (d) SSIM [157] map of the distorted image (enhanced for visibility). It is not hard to observe that SSIM gives a more reasonable quality map that successfully captures annoying pseudo-contouring effects (in the sky region) and blocking artifacts (along the boundaries of the building) induced by JPEG compression. Image by courtesy of Wang and Bovik [153].

belongs to this category. RR-IQA methods rely on statistical features from the reference image to evaluate the quality of the distorted image [163, 166]. The extracted features are assumed to transmit to the receiver side using an error-free ancillary channel. Blind image quality assessment (BIQA) methods predict image quality without accessing the reference image, making them the most challenging among the three. The focus of this thesis is on BIQA.

With a variety of BIQA models available, how to fairly compare their relative performance becomes a challenge. The conventional approach in the literature is to build an image database, collect subjective opinions for all images, and compute correlations between model responses and the mean opinion scores (MOSs) given by human subjects. A-mong many correlation metrics, Pearson linear correlation coefficient (PLCC), Spearman's rank-order correlation coefficient (SRCC), and Kendall rank-order correlation coefficient (KRCC) are the most widely used [149]. However, collecting MOSs via subjective testing is slow, costly, and cumbersome. In practice, the largest IQA database that is publicly available contains a maximum of $3,000$ subject-rated images [115]. Among those, many are generated from the same source images, differing only in distortion types and levels. In other words, fewer than 30 source images are included. By contrast, a digital image lives in a very high dimensional space, whose dimension is equal to the number of pixels in the image. Therefore, it is extremely difficult to collect sufficient subjective opinions to adequately cover the space. Perhaps more importantly, a few dozen source images are unlikely to provide a sufficient representation of the variations of the real-world image content. In addition, most objective BIQA methods are developed after commonly used IQA databases became publicly available, and often involve machine learning techniques or manual parameter adjustments to boost the performance. All these issues cast challenges on the generalizability of existing BIQA models to real-world scenarios.

We believe that a large-scale database with greater content diversity is critical to evaluate BIQA models. This motivates us to build the Waterloo Exploration Database, or in short the Exploration database, which in its current state contains $4,744$ pristine natural images, spanning a variety of the real-world content. We extend it by adding four distortion types—JPEG compression, JPEG2000 compression, white Gaussian noise contamination, and Gaussian blur—each with five distortion levels, resulting in a total of $99,624$ images. Given such a large number of sample images, it is extremely difficult (if not impossible) to collect MOSs for all images in a well-controlled laboratory environment. Therefore, innovative methods are necessary to make use of the Exploration database for comparing BIQA models. Here we present three test criteria that do not require subjective testing, termed as pristine/distorted image discriminability test (D-test), listwise ranking consistency test (L-test), and pairwise preference consistency test (P-test), respectively. Each test exams the robustness and generalizability of BIQA models from a different perspective. Specifically, D-test quantifies the ability of BIQA models to discriminate pristine from distorted images. L-test checks the consistency of BIQA models under test images differing only in distortion levels. P-test builds upon the notion of the quality-discriminable image pair (DIP), which consists of two images whose perceptual quality is discriminable, and evaluates the preference concordance of BIQA models on DIPs. By applying the three test criteria to the Exploration database, we perform a systematic comparison of 12 existing BIQA models and make a number of useful observations to reveal their weaknesses and to provide insights on how to improve these models.

Moreover, we propose a new psychophysical methodology, which we name group MAximum Differentiation (gMAD) competition for comparing computational models of a perceptually discriminable quantity. Specifically, we adopt the fundamental idea behind the MAximum Differentiation (MAD) competition [164] introduced by Wang and Simoncelli, which attempts to falsify a model, and one that is more difficult to be falsified is considered better. MAD gives us an opportunity to largely reduce the required number of test stimuli because ideally even one counterexample is sufficient to falsify a model. Here we extend MAD [164] in several key ways. When falsifying a model, instead of generating the stimuli from the space of all possible stimuli using computationally expensive optimization algorithms, we structure gMAD to explore a pre-selected set of stimuli, finding a stimulus pair that have maximally different responses of one model while holding the other fixed. This relaxation brings many benefits to model comparisons. First, the stimulus generation is gradient-free, which enables almost all computational models to compete with one another. Second, we can specialize a competition for computational models by testing them on some specific types of stimuli. Third, the gMAD-generated stimuli are more natural and interpretable than those generated by MAD and therefore provide more practical insights on how to improve competing models. gMAD allows multiple models to participate in the competition and aggregates pairwise measurements into a global ranking by maximum likelihood. To help summarize the relative performance, we introduce the notions of aggressiveness and resistance. We apply gMAD to the field of IQA and report the competition results on 16 objective IQA models. The framework is extensible, allowing future IQA models into the competition.

Now, we turn our attention from evaluation methodologies to algorithm designs of BIQA. One of the biggest challenges in learning BIQA models is the conflict between the gigantic image space and the extremely limited reliable ground truth data for training. Here we first show that a vast amount of reliable training data in the form of DIPs can be obtained automatically at low cost by exploiting large-scale databases with diverse image content. We then learn an opinion-unaware (OU) BIQA (meaning that no subjective opinions are used for training) model using RankNet [8], a pairwise learning-to-rank (L2R) algorithm, from millions of DIPs, leading to a DIP inferred quality (dipIQ) index. Extensive experiments on four benchmark IQA databases demonstrate that dipIQ outperforms state-of-the-art OU-BIQA models. The robustness of dipIQ is also significantly improved as confirmed by the gMAD competition methodology. Furthermore, we extend the proposed framework and learn models with ListNet [9] (a listwise L2R algorithm) on quality-discriminable image lists (DILs). The resulting DIL inferred quality (dilIQ) index achieves an additional performance gain.

Finally, we propose a Multi-task End-to-end Optimized deep neural Network (MEON) for BIQA. MEON consists of two sub-networks—a distortion identification network and a quality prediction network—sharing the early layers. Unlike traditional methods used for training multi-task networks, our training process is performed in two stages. In the first stage, we train a distortion type identification sub-network, for which large-scale ground truth training samples are readily available. In the second stage, starting from the pre-trained early layers and the outputs of the first sub-network, we train a quality prediction sub-network using a variant of the stochastic gradient descent method. Unlike most deep neural networks (DNNs), we choose biologically inspired generalized divisive normalization (GDN) instead of rectified linear unit (ReLU) as the activation function. We empirically demonstrate that GDN is effective at reducing model parameters/layers while achieving similar quality prediction performance. With modest model complexity, the proposed MEON index achieves state-of-the-art performance on four publicly available benchmarks. Moreover, we demonstrate the strong competitiveness of MEON against existing BIQA models using the gMAD competition methodology.

1.2 Objectives

The objectives of this thesis are to overcome the fundamental limitations of traditional evaluation and design methodologies of BIQA, especially in their limited capacities at exploiting large-scale image databases, and to develop accurate and reliable BIQA models using innovative techniques.

1.3 Contributions

The main contributions of the thesis are four-fold. First, we establish the Waterloo Exploration Database, which is the largest one in IQA research, and present three test criteria (D-test, P-test, and L-test) for BIQA model comparison without MOSs. Second, we propose the gMAD competition methodology for comparing computational models of perceptually discriminable quantities and apply gMAD to the field of IQA. Third, we introduce the notion of DIPs and propose L2R frameworks for OU-BIQA. Last, we leverage the recent advances in DNNs and propose a multi-task DNN for end-to-end BIQA.

1.4 Thesis Outline

The layout of this thesis is as follows.

Chapter 2 discusses the related work in the literature. It starts with a brief introduction of subjective testing methodologies, followed by a summary of commonly used IQA databases. We then provide an overview of existing IQA models with emphasis on BIQA ones.

Chapter 3 presents in detail the construction of the Waterloo Exploration Database. Three innovative test criteria are proposed. We systematically compare 12 state-of-the-art BIQA models by applying the three test criteria to the Exploration database and make a number of useful observations.

Chapter 4 introduces the gMAD competition, an efficient and practical methodology for comparing multiple computational models of a perceptually discriminable quantity. We apply gMAD to image quality and perform a systematic comparison of 16 IQA models on the Exploration database.[1]

Chapter 5 studies BIQA from an L2R perspective. We first summarize the limitations of most existing BIQA models. We then adopt a pairwise L2R algorithm (RankNet [8]) for

[1] To demonstrate the generality of the gMAD competition methodology, we investigate its usage in two more applications: image aesthetics evaluation [19] and video quality of experience prediction [47], whose details are given in Appendix A.

OU-BIQA. The input to RankNet, namely DIPs, can be automatically generated with the help of the most-trusted FR-IQA models. We extend the proposed pairwise L2R approach for OU-BIQA to a listwise L2R one.

Chapter 6 studies BIQA by leveraging the latest advances in deep learning. We propose a multi-task learning framework for BIQA, namely MEON, by decomposing the BIQA task into two subtasks with dependent loss functions. We end-to-end optimize MEON for both distortion identification and quality prediction.

Finally, Chapter 7 summarizes the thesis and points out promising directions for future endeavors.

Chapter 2

Literature Review

This chapter provides a literature review of previous studies that are closely related to our work, including subjective testing methodologies, commonly used IQA databases, and existing objective IQA models with emphasis on BIQA ones.

2.1 Subjective Testing Methodologies and IQA Databases

Human subjective testing is the first step towards understanding the visual perception of image quality. Depending on how test images are presented to human subjects and how they are instructed to give opinions, subjective testing may be broadly classified into three categories: single-stimulus methods [135], paired comparison methods (also known as two-alternative-forced-choice) [115], and multiple-stimulus methods [89]. For single-stimulus methods, one test image is shown at any time instance and is given ratings to reflect its perceptual quality. For paired comparison methods, a pair of images are shown either simultaneously or consecutively, and the subjects are asked which image is perceived to have better quality. For multiple-stimulus methods, multiple images are shown simultaneously, and the subjects rank or give ratings to all images based on their perceptual quality. Suppose that there are N test images in total. $\mathcal{O}(N)$ evaluations are needed in single-stimulus and multiple-stimulus methods, while $\mathcal{O}(N^2)$ evaluations are required in a full paired comparison experiment. Although paired comparison methods are often preferred to collect reliable subjective opinions, an exhaustive paired comparison requires a very large number of evaluations, which are impractical when N is large. A number of methods have been proposed to improve its efficiency. Four types of balanced subset designs were developed in the 1950s [13], among which the square design method became popular and was later improved by Li *et al.* in measuring visual discomfort of 3DTV [72]. An alternative method is to randomly select a small subset of image pairs for each subject [24], and it has been shown that at least $\mathcal{O}(N \log N)$ distinct pairs are necessary for large random graphs to guarantee the graph connectivity and to achieve a robust global ranking using HodgeRank [55]. In the construction of a subjective IQA database [115], a Swiss competition principle was adopted to decrease the evaluations to $\mathcal{O}(N \log N)$. Recently, an active sampling strategy for subjective testing was proposed [182] with a complexity of $\mathcal{O}(N)$. Either a single-stimulus test or a paired comparison is queried based on already accumulated subjective opinions with the goal of maximizing the expected information gain [182].

Several IQA databases have been widely used in the literature. In 2005, Sheikh *et al.* conducted a subjective user study and created the LIVE [135] database that consists of 29 reference and 779 distorted images with 5 distortion types—JPEG2000 compression, JPEG compression, white Gaussian noise contamination, Gaussian blur, and fast fading transmission error. A single-stimulus continuous-scale method [149] is adopted for testing, where the reference images are also evaluated under the same experimental configuration [134]. MOS scaling and realignment (based on an additional double-stimulus subjective experiment) are performed to align the scores across different distortion sessions. In particular, the scaling compensates for different scales used by different subjects during rating, while the realignment avoids the significant bias of MOSs towards any specific distortion type and/or level.

The TID2008 [114] database contains 24 pristine natural and 1 computer generated images. 18 of them are taken from LIVE [135], differing only in size via cropping. 17 distortion types with four distortion levels are added, resulting in $1,700$ distorted images.

Table 2.1: Distortion types used in the TID2008 database [116]

No.	Distortion type
1	White Gaussian noise
2	White Gaussian noise in color components
3	Spatial correlated noise
4	Masked noise
5	High frequency noise
6	Impulse noise
7	Quantization noise
8	Gaussian blur
9	Image denoising artifacts
10	JPEG compression
11	JPEG2000 compression
12	JPEG transmission errors
13	JPEG2000 transmission errors
14	Non-eccentricity pattern noise
15	Local block-wise distortions
16	Mean shift
17	Contrast change

Table 2.2: Seven added distortion types in the TID2013 database [115]

No.	Distortion type
1	Change of color saturation
2	Multiplicative Gaussian noise
3	Comfort noise
4	Lossy compression of noisy images
5	Image color quantization with dither
6	Chromatic aberrations
7	Sparse sampling and reconstruction

The distortion types are summarized in Table 2.1. The testing methodology is a paired comparison method [20], where the reference image is also shown to the subjects. A Swiss competition principle is used to reduce the number of pairs for subjective testing such that each image appears in at most nine pairs. No explicit MOS scaling and realignment are reported to refine the raw MOSs collected from multiple sessions in three countries. TID2008 was later extended to TID2013 [115] by adding seven new distortion types (shown in Table 2.2) and one additional distortion level, making it the largest public database so far. However, the design philosophies of TID2008 [114] have been questioned since its release [69, 34]. First, the MOS of each image is computed as the number of wins in nine pair comparisons it involves, rather than aggregated using mature global ranking approaches [55, 147, 97]. Second, there are no cross-content pairs evaluated during the experiment nor extra realignment experiments across content. As a result, the MOSs of the test images of different content may not be comparable.

The CSIQ [69] database contains 30 reference and 866 distorted images by adding 6 distortion types with 4 to 5 distortion levels. The distortion types used in CSIQ are JPEG compression, JPEG2000 compression, global contrast decrement, pink Gaussian noise contamination, white Gaussian noise contamination, and Gaussian blur. CSIQ uses a multi-stimulus absolute category method based on a linear displacement of images of the same content across four calibrated LCD monitors placed side by side with equal viewing distance to the observer. MOSs with different content are realigned according to a separate, but identical experiment in which observers place subsets of all the images linearly in space.

Table 2.3: Comparison of existing IQA databases

Database	# of pristine images	# of distorted images	Distortion type
LIVE [135]	29	779	5 / Simulated
TID2008 [114]	25	$1,700$	17 / Simulated
TID2013 [115]	25	$3,000$	25 / Simulated
CSIQ [69]	30	866	6 / Simulated
MD [54]	15	405	2 / Simulated
Challenge [34]	0	$1,162$	Authentic
Exploration	$4,744$	$94,880$	4 / Simulated

Table 2.4: Comparison of subjective testing methodologies used in existing IQA databases

Database	Testing methodology
LIVE [135]	Single-stimulus continuous scale
TID2008 [114]	Paired comparison
TID2013 [115]	Paired comparison
CSIQ [69]	Multi-stimulus absolute category
MD [54]	Single-stimulus continuous scale
Challenge [34]	Single-stimulus continuous scale with crowdsourcing
Exploration	Need-based

The LIVE multiply distorted (MD) database [54] and the LIVE in the wild image quality challenge database [34] (Challenge) focus on images with mixed distortions. LIVE MD Database simulates two multiple distortion scenarios, one for image storage (Gaussian blur followed by JPEG compression) and the other for digital image acquisition (Gaussian blur followed by white Gaussian noise contamination). It contains 15 pristine and 405 distorted images. The test methodology is the same as in LIVE [135]. LIVE Challenge Database takes a step further and directly works with authentically distorted images captured from mobile devices. A total of $1,162$ images are included, whose MOSs are crowdsourced using the Amazon Mechanical Turk platform. Substantial effort has been put to process the noisy raw data and to verify the reliability of the human opinions from such an uncontrolled testing environment. A summary of the aforementioned databases are given in Tables 2.3 and 2.4.

Other widely known but smaller databases include IVC [70], Toyama-MICT [46], Cornell A57 [11], and WIQ [25]. A useful collection of IQA databases has been assembled [169, 168].

A major common issue of all the existing IQA databases is the limited number of source images being used (as a matter of fact, none of the databases includes more than 30 source images), which creates a large gap between the diversity of real-world images and the variation of image content that can be tested using the databases. Therefore, IQA models developed and validated using the databases are inevitably questioned on their generalizability to real-world applications. This is evidenced by the recent test results on the LIVE Challenge Database, a direct collection of images from the Internet, where the performance of the most advanced BIQA models drops significantly [34]. The limitation on the number of source images is largely due to the limited capacity of the affordable subjective testing. For example, testing the $1,700$ distorted images in TID2008 [114] is an expensive and highly time-consuming task, but given the combinations of distortion types and levels that are applied to each source image, eventually, only 25 source images can be included.

The above issue motivates us to build a new database for IQA research, which aims to significantly expand the diversity of image content. Meanwhile, testing all images in the database using conventional subjective testing methodologies becomes extremely difficult, if not impossible. Therefore, innovative approaches on how to use the database to test IQA models have to be developed in order to meet the challenge. These are some of the key issues we would like to address in this thesis.

2.2 Objective IQA Models

Pioneering work on perceptual image processing and IQA dated back to as early as 1970s [94], when Mannos and Sakrison investigated a class of visual fidelity measures in the context of rate-distortion optimization. Since then, more and more researchers began to realize that the widely used mean squared error (MSE) as the dominant quantitative criterion for assessing signal quality and fidelity has a poor correlation with human perception of image quality [35]. Therefore, a number of alternatives that better account for

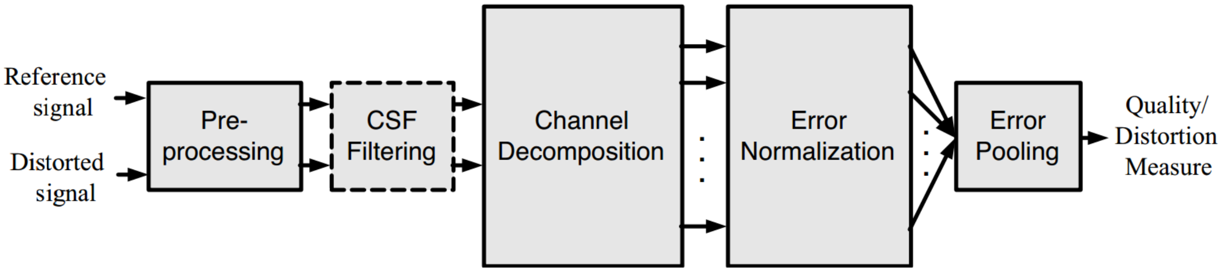


Figure 2.1: A prototypical quality assessment system based on error sensitivity. CSF: Contrast sensitivity function. Image by courtesy of Wang *et al.* [157].

perceived image quality have been proposed. For example, Safranek and Johnston incorporated a perceptually masking model into their image coder [126]. Daly proposed the visible differences predictor [18] that transfers the physical differences into perceptually visible differences by explicitly taking into account three sensitivity variations of the HVS. These are the variations as a function of light level (referred to as the amplitude nonlinearity), spatial frequency (measured by the contrast sensitivity function), and signal content (referred to as masking). Daly's work [18] uses 2D images rather than only the parameters of the imaging systems, resulting in a probability detection map of visible differences rather than a single score to represent the fidelity of the whole image.

The development of the structural similarity (SSIM) index transformed the design paradigm of IQA from computing error visibility (shown in Fig. 2.1) to measuring structural similarity [151, 157]. The underlying assumption is that the HVS is highly adapted to extract structural information from the viewing field and a measure of structural information change can provide a good approximation to perceived image distortions [157]. Fig. 2.2 shows the computation diagram of SSIM, where it decomposes the image signal into the luminance, contrast and structure terms, and measures the distortions separately. It opened the door to a new class of FR-IQA algorithms that consider the HVS as a black box and model it with some holistic assumptions. A variety of research [165, 69, 189, 159, 32, 76, 136, 188, 190, 180, 178, 118, 162] has used SSIM as a basis.

Yet, in most present and emerging practical real-world visual communication environ-

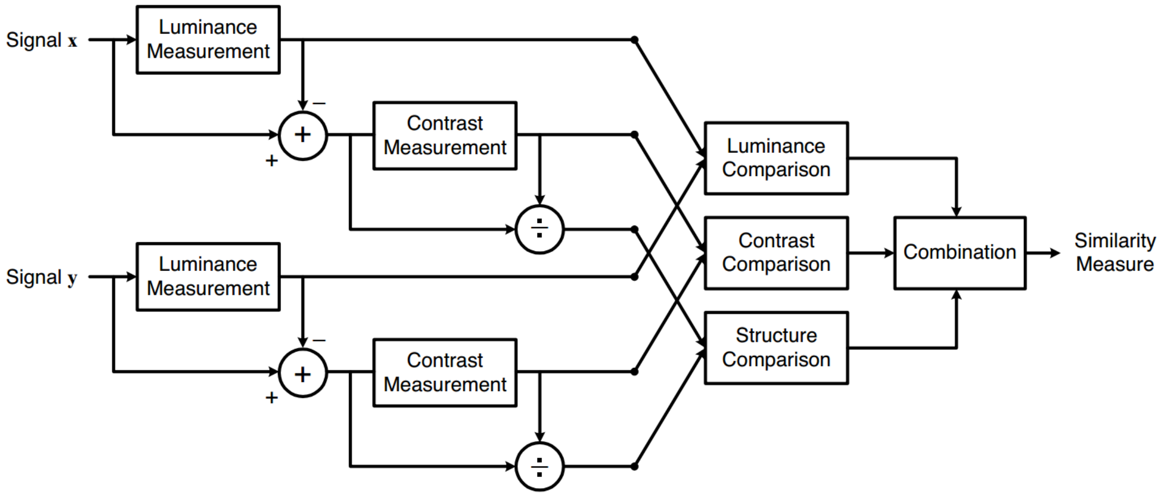


Figure 2.2: Computation diagram of the SSIM [157] index. Image by courtesy of Wang *et al.* [157].

ments, the FR-IQA methods described above are not applicable because the reference images are not accessible at the receiver side (or perhaps do not exist at all[1]) [154]. Therefore, being able to blindly assess image quality is highly desirable. Most existing BIQA models [102, 125, 99, 184, 103, 173, 177, 38, 171, 172] share a common two-stage structure: 1) perception- and/or distortion-relevant features (denoted by $\mathbf{x}$) are extracted from a test image; 2) a non/linear quality prediction function $f(\mathbf{x})$ is learned by mature statistical machine learning tools, such as support vector regression (SVR) [15, 128] and artificial neural network [122] from training images with MOSs.

2.2.1 Feature Extraction

Three types of knowledge can be exploited to produce useful features for BIQA, as shown in Fig. 2.4. The first type is knowledge about the visual world to which we are exposed, which summarizes the statistical regularities of the undistorted images. The second type

[1] It is pragmatical that even under the most ideal and controlled circumstances, a captured optical image will inevitably suffer from some kind of distortions [156, 6].

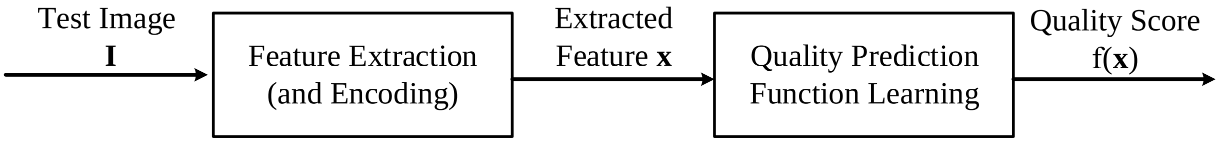


Figure 2.3: Two-stage structure of BIQA models.

is knowledge about degradation, where we can explicitly construct features that are aware of particular artifacts, such as blocking [170, 155, 78], blurring [145, 161, 192], and ringing [108, 133, 77]. The third type is knowledge of the HVS [150, 121, 138], which is based on perceptual models originated from visual physiology and psychophysical studies [28, 33, 43, 50]. The so-called natural scene statistics (NSS) that seek to capture the natural statistical behaviors of images embody the three-fold modeling in a delightful way [154]. NSS can be extracted directly in spatial domain or in transform domains such as DFT [111], DCT [7], and wavelets [137, 92, 93].

In the spatial domain, edges are presumably the most important image features for the HVS. The edge spread can be used to detect blurring [74, 95] and the intensity variance in smooth regions close to edges can indicate ringing artifacts [108]. Step edge detectors that operate at 8×8 block boundaries measure the discontinuities caused by JPEG compression [170, 98]. The sample entropy [131, 16] of intensity histograms is used to identify image anisotropy [71, 26]. The responses of image gradients and Laplacian of Gaussian filters are jointly modeled to describe the destruction of the statistical naturalness of images [177]. In [99], the mean subtracted contrast normalized pixel value statistics are modeled using a generalized Gaussian distribution (GGD), which is inspired by the adaptive gain control mechanism seen in neurons [43]. Such a normalization approach has been used in many BIQA models [101, 100, 185, 183] as a starting point of feature extraction and a preprocessing step for DNN-based BIQA models [59, 60, 62]. There are also local image features based on singular value decompositions of local image gradient matrices [193].

Statistical modeling in wavelet domain bears a natural resemblance to the early visual system [50], and natural images exhibit statistical regularities in wavelet space. Specifically, it is widely acknowledged that the marginal distribution of wavelet coefficients of a natural image (regardless of content) has a sharp peak near zero and heavier than Gaussian tails,

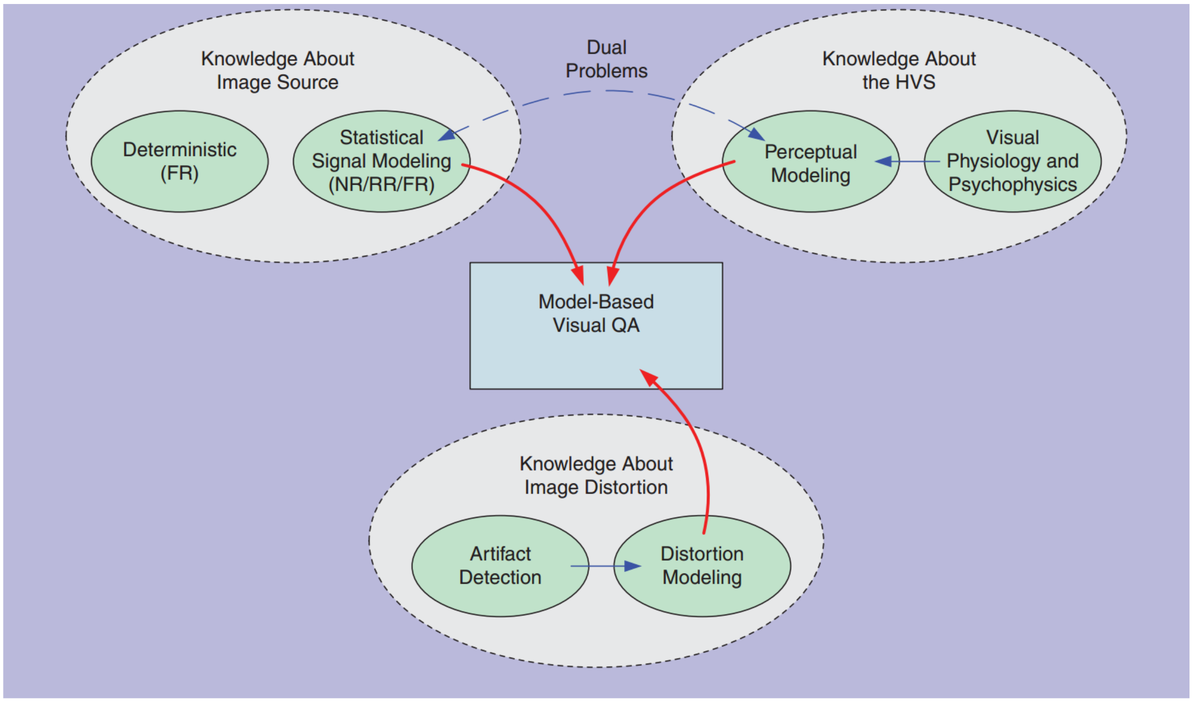


Figure 2.4: Knowledge map that we can exploit to produce useful features for IQA. Image by courtesy of Wang and Bovik [154].

as shown in Fig. 2.5. As a result, the wavelet magnitudes [102, 48] and the wavelet coefficient correlations in the neighborhood [133, 103, 141, 142, 181] can be individually or jointly modeled as image naturalness. The phase information of wavelet coefficients, for example, expressed as the local phase coherence is exploited to describe the perception of blur [40, 161].

In DFT domain, blur kernels can be efficiently estimated [141, 175, 142] to quantify the degree of image blurring. The regular peaks at feature frequencies can be used to identify blocking artifacts [155, 160]. However, it is generally accepted that much of perceptual information in an image signal is stored in the Fourier phase rather than the Fourier amplitude [110, 49]. Phase congruency [67] is such an implementation that identifies perceptually significant image features at spatial locations, where the Fourier components are maximally in-phase with one another [71].

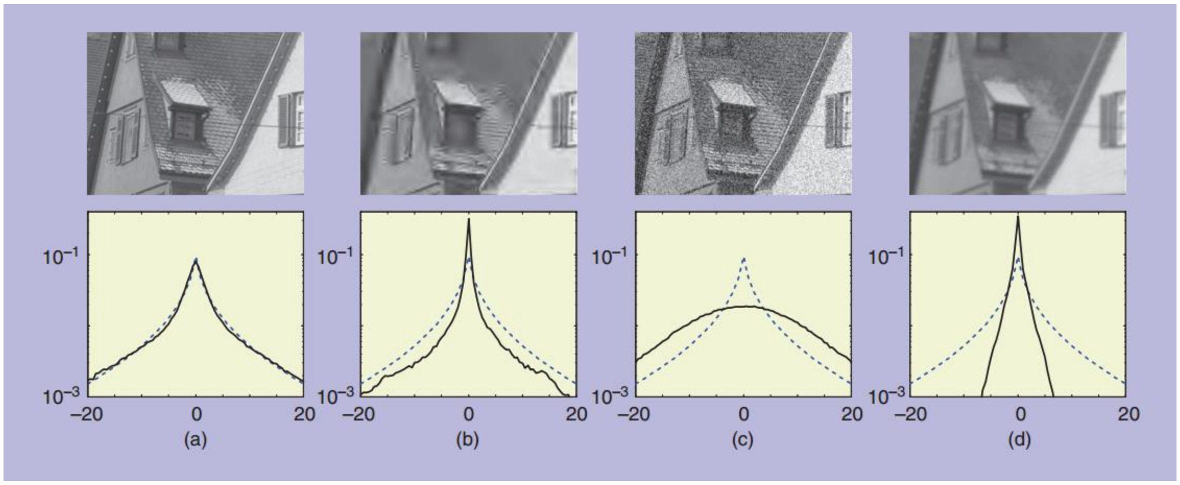


Figure 2.5: Wavelet coefficient histograms (solid curves) of (a) original “buildings” image, (b) compressed by JPEG2000, (c) with white Gaussian noise, and (d) blurred by a Gaussian kernel. The histogram in (a) is well fitted by a GGD model (dashed curves). The shapes of the histograms change in different ways for different distortion types. Image by courtesy of Wang and Bovik [154].

In DCT domain, blocking artifacts may be identified in a shifted 8×8 block [78]. The ratio of AC coefficients to DC components may be interpreted as a measure of local contrast [124]. The kurtosis of AC coefficients may be used to quantify the structure statistics. In addition, AC coefficients may also be jointly modeled using a GGD [125].

Feature learning techniques have been investigated in the context of BIQA. Ye *et al.* learned quality filters on image patches using K-means clustering [52] and used the filter responses as features [184]. Despite its high dimension, the feature representation in [184] has been frequently adopted in later BIQA models such as BLISS [183] and dipIQ [88] (details will be given in Chapter 5). Ye *et al.* then took one step further by adopting supervised filter learning [185]. Xue *et al.* [179] proposed a quality-aware clustering scheme guided by an FR-IQA measure [189].

The dimension of the feature vector $\mathbf{x}$ extracted from a test image may be extremely high, while the number of training samples is typically small. This poses a challenge to traditional machine learning algorithms. As a result, dimensionality reduction techniques, such as principal component analysis [148] may be adopted to reduce $\mathbf{x}$ to a lower dimension.

2.2.2 Model Learning

From a model learning point of view, support vector machine (SVM) and its continuous variant SVR [128] are the most commonly used machine learning tools [102, 103, 181, 184, 185, 177] in the field of BIQA. The capabilities of neural networks to pre-train a model without labels and to easily scale up have also been exploited for this purpose [71, 142, 48]. Example-based methods approximate the quality score of a test image by the weighted average of quality scores of training images, where the weight encodes the perceptual similarity between the test and training images [181, 179, 171]. Saad *et al.* jointly modeled $\mathbf{x}$ and MOS using a multivariate Gaussian distribution and performed prediction by maximizing the conditional probability $\Pr(\mathbf{x}|\text{MOS})$ [124, 125]. Similar probabilistic modeling strategies have been investigated in [101, 187]. Other advanced learning algorithms include topic modeling (*e.g.*, probabilistic latent semantic analysis [45]), Gaussian process [142], and multi-kernel learning [31, 30].

Recently, there has been a growing interest in jointly learning the feature representation and quality prediction function using DNNs. Ma *et al.* [87] proposed a fully convolutional network for local blur mapping. Kang *et al.* [59] implemented a DNN with one convolutional and two fully connected layers for BIQA. To perform both maximum and minimum pooling, ReLU nonlinearity [105] is omitted right after convolution. Bianco *et al.* investigated various design choices of DNN for BIQA [4]. They first adopted DNN features pre-trained on an image classification task as input to learn a quality evaluator using SVR [128]. They then fine-tuned the pre-trained features in a multi-class classification setting by quantizing MOSs into five categories and fed the fine-tuned features to SVR. Nevertheless, their proposal is not end-to-end optimized and involves heavy manual parameter adjustments [4]. Bosse *et al.* [5] significantly increased the depth of their DNN by stacking ten convolutional and two fully connected layers, whose architecture was inspired by the VGG16 network [139] for image classification. They also adapted their network to handle FR-IQA. Kim and Lee [62] first utilized local scores of an FR-IQA algorithm as

Table 2.5: Model sizes of DNN-based BIQA algorithms

BIQA model	Kang14 [59]	Kang15 [60]	DeepBIQ [4]	deepIQA [5]	Kim17 [62]
Size ($\times 10^4$)	72	7.9	5,687	523	739

ground truths to pre-train their network and then fine-tuned it using MOSs. They observed that pre-training with adequate epochs is necessary for the fine-tuning step to converge. We summarize the complexities of DNN-based BIQA models in Table 2.5.

Chapter 3

Waterloo Exploration Database and Testing Criteria

In this chapter, we construct the Waterloo Exploration Database, which is the largest among all IQA databases in the literature. Sourcing MOSs of all images in such a large-scale database using conventional subjective testing methodologies becomes extremely difficult. Therefore, innovative methods on how to use the Exploration database to test BIQA models have to be developed in order to meet the challenge. We propose three test criteria: pristine/distorted image discriminability test (D-test), listwise ranking consistency test (L-test), and pairwise preference consistency test (P-test).

3.1 Constructing the Waterloo Exploration Database

We construct a new image database—the Waterloo Exploration Database—which currently contains $4,744$ pristine natural images that span a great diversity of image content. An important consideration in selecting the images is that they need to be representative of scenes we see in our daily life. Therefore, we resort to the Internet and elaborately select 196 keywords to search for images. The keywords can be broadly classified into 7 categories: human, animal, plant, landscape, cityscape, still-life, and transportation. We

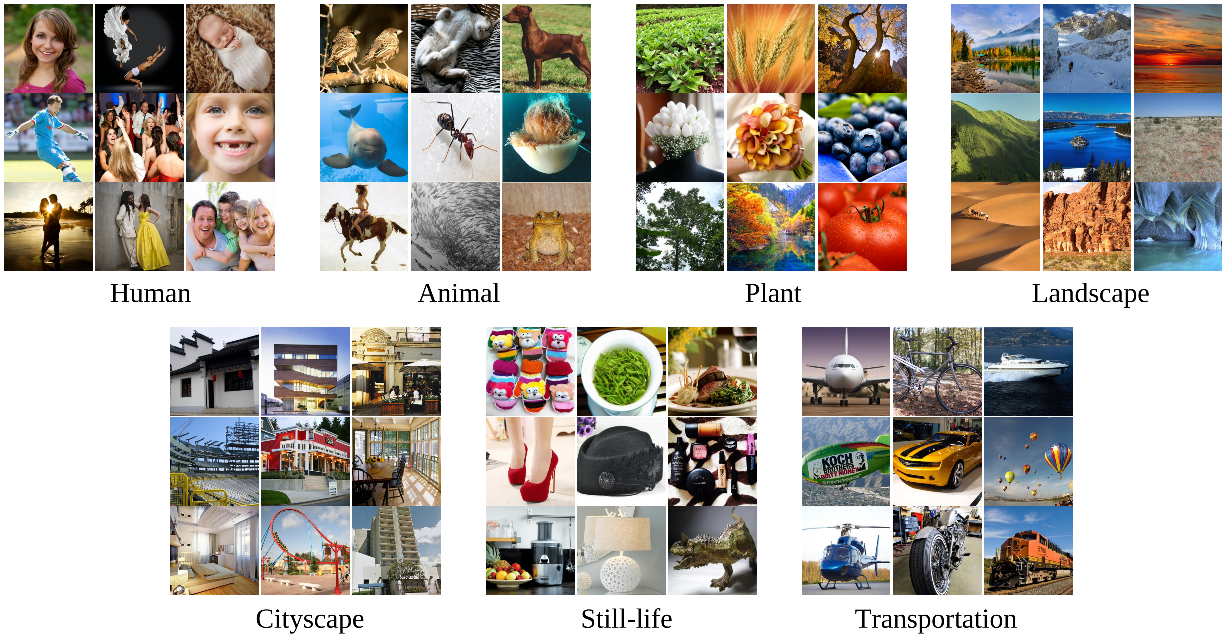
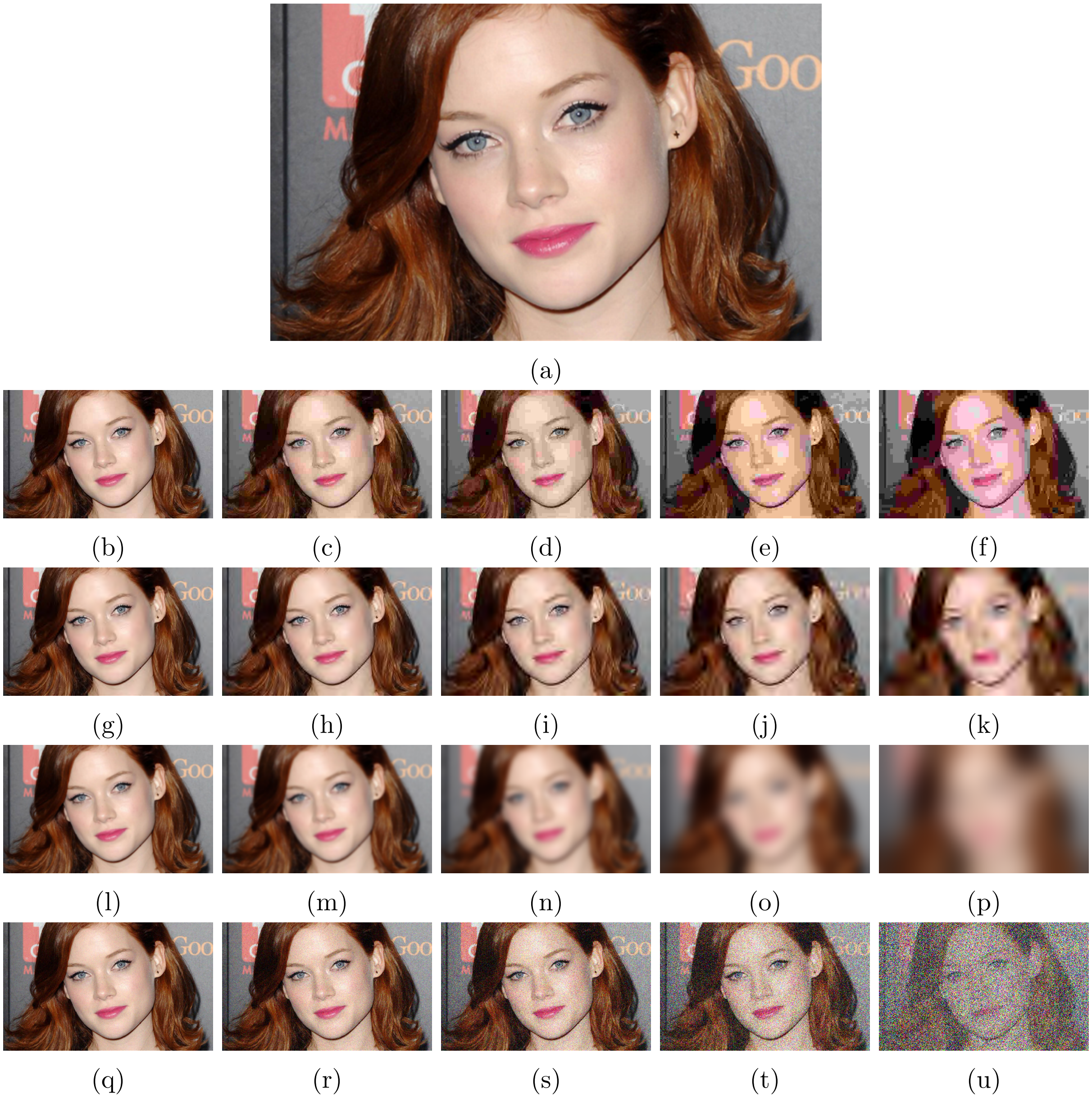


Figure 3.1: Sample source images in the Waterloo Exploration Database. All images are cropped for better visibility.

initially obtain more than $200,000$ images. Many of them contain significant distortions or inappropriate content, and thus a sophisticated manual process is applied to refine the selection. In particular, we remove images that have obvious distortions, including heavy compression artifacts, strong motion blur, out of focus blur, low contrast, underexposure, overexposure, substantial sensor noise, visible watermarks, artificial image borders, and other distortions due to improper operations during acquisition. Images of too small or too large sizes, cartoon and computer generated content, and inappropriate content are excluded. After this step, about $7,000$ images remain. To make sure that the remaining images are of pristine quality, we further carefully inspect each image multiple times by zooming in and remove images with visible compression distortions. Eventually, we end up with $4,744$ high-quality natural images. Sample images grouped into different categories are shown in Fig. 3.1.

Four distortion types with five levels each are chosen to alter the source images. All distorted images are generated using MATLAB functions as follows.



(a)

(b) (c) (d) (e) (f)

(g) (h) (i) (j) (k)

(l) (m) (n) (o) (p)

(q) (r) (s) (t) (u)

Figure 3.2: Sample distorted images in the Exploration database. (a) Source reference image "Actress". (b) JPEG: level 1. (c) JPEG: level 2. (d) JPEG: level 3. (e) JPEG: level 4. (f) JPEG: level 5. (g) JP2K: level 1. (h) JP2K: level 2. (i) JP2K: level 3. (j) JP2K: level 4. (k) JP2K: level 5. (l) BLUR: level 1. (m) BLUR: level 2. (n) BLUR: level 3. (o) BLUR: level 4. (p) BLUR: level 5. (q) WN: level 1. (r) WN: level 2. (s) WN: level 3. (t) WN: level 4. (u) WN: level 5.

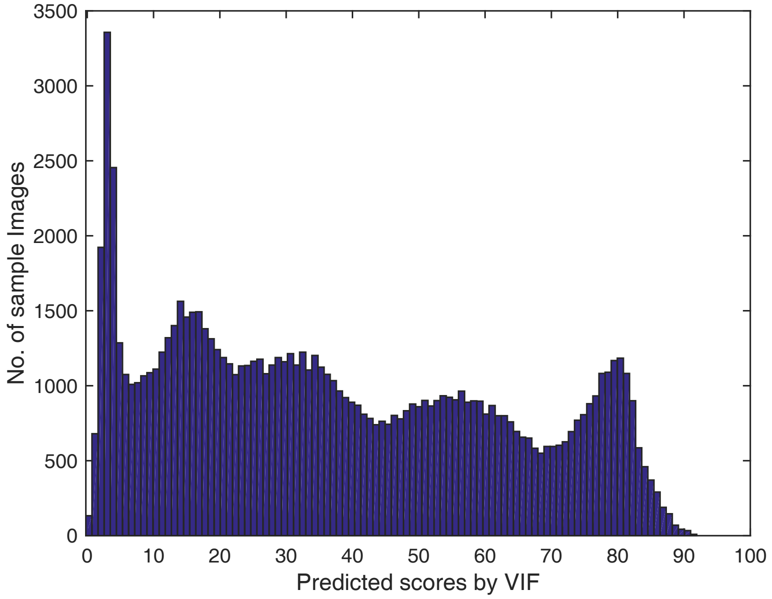


Figure 3.3: Score distribution of the Exploration database predicted by the FR-IQA model, VIF [132].

- JPEG compression (JPEG): The quality factor that parameterizes the DCT quantization matrix is set to be $[43, 12, 7, 4, 0]$ for five levels, respectively.

- JPEG2000 compression (JP2K): The compression ratio is set to be $[52, 150, 343, 600, 1200]$ for five levels, respectively.

- Gaussian blur (BLUR): 2D circularly symmetric Gaussian blur kernels with standard deviations (stds) of $[1.2, 2.5, 6.5, 15.2, 33.2]$ for five levels are used to blur the source images.

- White Gaussian noise contamination (WN): White Gaussian noise is added to the source images, where variances are set to be $[0.001, 0.006, 0.022, 0.088, 1.000]$ for five levels, respectively.

Sample distorted images are shown in Fig. 3.2. Note that these are the most common distortion types in existing IQA databases [134, 116], and many BIQA models are claimed to excel at handling these distortions [102, 103, 125, 99, 184, 101, 179, 177, 127, 187, 173, 38, 171]. Therefore, whether these models perform well on the new Exploration database becomes a strong test on their claimed performance and generalizability in the real-world. The specific parameters that control the distortion levels for each type are chosen so as to uniformly cover the subjective quality scale, predicted by the FR-IQA model VIF [132] as shown in Fig. 3.3. Once determined, the parameters are fixed for all images. In summary, the Exploration database contains a total of $99,624$ images. The number of pristine and distorted images is 150 times and 30 times, respectively, as many as those of the largest existing databases.

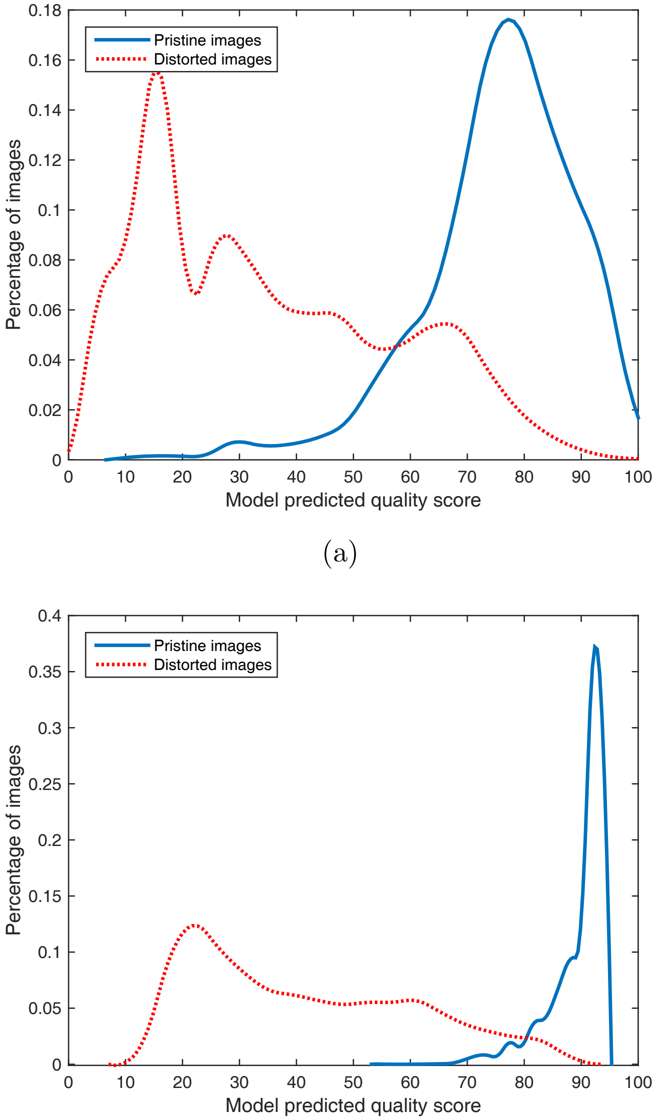
3.2 Evaluating BIQA Models

To make use of the Exploration database in comparing the relative performance of BIQA models, we introduce three test criteria: D-test, L-test, and P-test, which do not require subjective testing.

3.2.1 Pristine/Distorted Image Discriminability Test (D-Test)

Considering pristine and distorted images as two distinct classes, D-test aims to test how well an IQA model separates the two classes. An illustration using the Exploration database is shown in Fig. 3.4, where an IQA model with strong discriminability (*e.g.*, Wang05 [163]) is able to map pristine and distorted images onto easily separable intervals with minimal overlaps, whereas a less competitive model creates two score distributions with large overlaps. We introduce a measure to quantify this discriminability. We first group indices of pristine and distorted images into the sets of I_p and I_d, respectively, and use $|\cdot|$ to denote the cardinality of a set. Let $\hat{q}_i$ represent the predicted quality of the i-th image by a model. We apply a threshold T on $\{\hat{q}_i\}$ to classify the images such that $I'_p = \{i|\hat{q}_i > T\}$ and $I'_d = \{i|\hat{q}_i \leq T\}$. The average correct classification rate is given by

$$\mathrm{ACR}(T) = \frac{1}{2}\left(\frac{|I_p \cap I'_p|}{|I_p|} + \frac{|I_d \cap I'_d|}{|I_d|}\right) . \tag{3.1}$$



(a)

(b)

Figure 3.4: Distributions of IQA model predictions on the Exploration database. An ideal IQA model is expected to have strong discriminability and to create small overlaps between the two distributions. (a) DIIVINE [103]. (b) WANG05 [163].

T should be optimized to yield the maximum correct classification rate, from which we define a discriminability index as

$$D = \max_{T} \mathrm{ACR}(T)\,. \tag{3.2}$$

D lies in $[0, 1]$ with a larger value indicating better separability between pristine and distorted images. (3.2) is a univariate optimization problem that can be solved using a line search method.

3.2.2 Listwise Ranking Consistency Test (L-Test)

The idea behind L-test has been advocated by Winkler [95, 168]. The goal is to evaluate the robustness of BIQA models when rating images with the same content and the same distortion type but different distortion levels. The underlying assumption is that image quality degrades monotonically with the increase of the distortion level for any distortion type. Therefore, an excellent BIQA model should rank the images in the same order. An example on the Exploration database is given in Fig. 3.5, where different models may or may not produce the same quality rankings in consistency with distortion levels. Given a database with N_p pristine images, C distortion types and K distortion levels, we use the average Spearman's rank-order correlation coefficient (SRCC) and Kendall rank-order correlation coefficient (KRCC) to quantify the ranking consistency between distortion levels and model predictions

$$\mathrm{LRC}_s = \frac{1}{N_p C} \sum_{i=1}^{N_p} \sum_{j=1}^{C} \mathrm{SRCC}(\mathbf{l}_{ij}, \hat{\mathbf{q}}_{ij})\,, \tag{3.3}$$

and

$$\mathrm{LRC}_k = \frac{1}{N_p C} \sum_{i=1}^{N_p} \sum_{j=1}^{C} \mathrm{KRCC}(\mathbf{l}_{ij}, \hat{\mathbf{q}}_{ij})\,, \tag{3.4}$$

where $\mathbf{l}_{ij}$ and $\hat{\mathbf{q}}_{ij}$ are both length-K vectors representing distortion levels and the corresponding model responses to the set of images from the same (i-th) source image and the

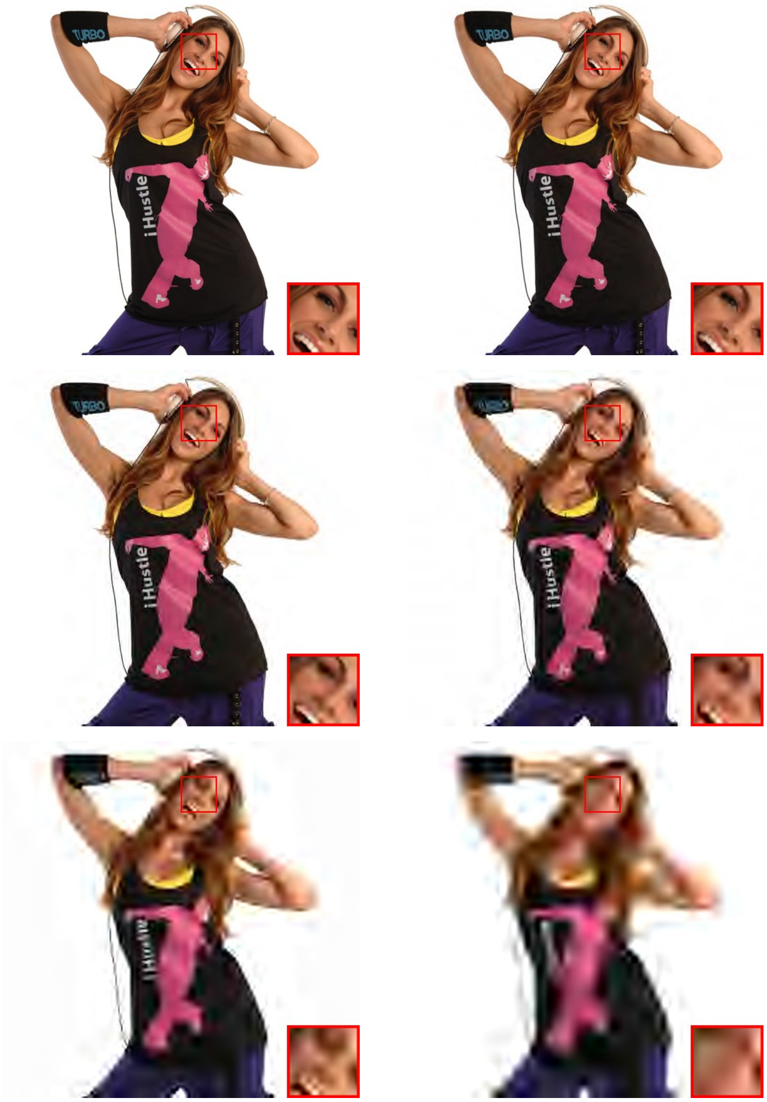


Figure 3.5: L-test of "Hip-hop Girl" images under JPEG2000 compression. Image quality degrades with the distortion level from left to right and from top to bottom. An excellent BIQA model (*e.g.*, ILNIQE [187]) ranks the images in exactly the same order. By contrast, a less competitive model (*e.g.*, QAC [179]) may give different rankings.

same (j-th) distortion type. LRC$_s$ and LRC$_k$ lie in $[0, 1]$ with higher values indicating better listwise ranking consistency.

3.2.3 Pairwise Preference Consistency Test (P-Test)

P-test compares BIQA model predictions on pairs of images whose quality is clearly discriminable. We call such pair of images quality-discriminable image pair (DIP). An ideal BIQA model should consistently predict preferences concordant with DIPs. Paired comparison is a widely used subjective testing methodology in IQA research, as discussed in Chapter 2. Pairwise preference has also been exploited to learn rank BIQA models [176, 30]. Nevertheless, in all previous studies, DIPs that can be used for testing or developing objective models are obtained exclusively from subjective ratings, which largely limits the number of available DIPs and is apparently impractical for large-scale image databases such as the Exploration database.

We propose a novel automatic DIP generation engine by leveraging the quality prediction power of several most-trusted FR-IQA measures in the literature. Specifically, we consider an image pair to be a valid DIP if the absolute differences of the predicted scores from FR models are all larger than a pre-defined threshold T. To explore this idea, we first experiment with the LIVE database [135], from which we extract all possible image pairs whose absolute MOS differences are larger than $T_g = 20$ and consider them as the ground truth DIPs. The legitimacy of $T_g = 20$ on LIVE can be validated from two sources. First, the average std of MOSs on LIVE is around 9 and $T_g = 20$ is right outside the ± 1 std range, which justifies the perceptual quality discriminability of the image pair. Second, from the subjective experiment conducted by Gao *et al.* [30], it can be observed that the consistency between the subjects on the relative quality of one pair from LIVE increases with T_g, and when T_g is larger than 20, the consistency approaches 100%. Using the available MOSs in LIVE [135], we are able to generate $206,717$ ground truth DIPs, termed as the ground truth set. After that, we use our DIP generation engine to generate DIPs on LIVE and observe whether the generated pairs are in the ground truth set. Fig. 3.6 shows the percentage of generated DIPs in the ground truth set as a function of T for different combinations of FR-IQA measures, where three bases FR measures (MS-SSIM [165],

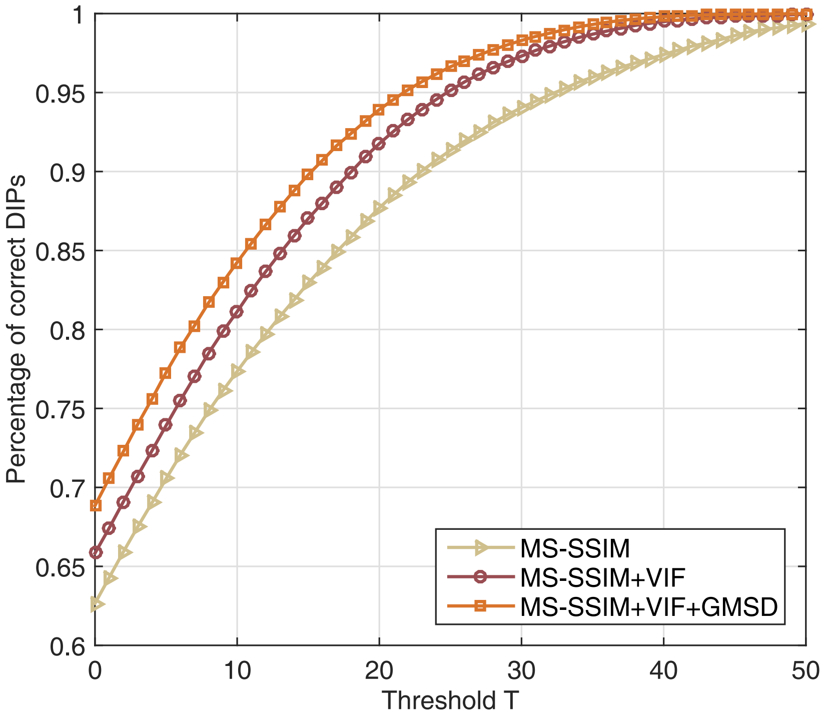


Figure 3.6: The percentage of generated DIPs in the ground truth set of the LIVE database [135] as a function of T for different combinations of base FR-IQA models.

VIF [132], and GMSD [180]) are selected. It can be seen that the percentage increases when more FR-IQA models are involved and is maximized when all 3 FR-IQA models are used. Using all the three models together with $T = 40$, we achieve 99.81% accuracy, which verifies the reliability of our DIP generation engine. This configuration is used as the default setting. Note that the model predictions of the three FR-IQA models should be mapped to the same perceptual scale before DIP generation. Fig. 3.7 shows 3 DIPs generated by the proposed engine on the Exploration database. One can see that the left images of the 3 DIPs have superior perceived quality compared to the right ones.

Given an image database S, our DIP generation engine goes through all possible pairs to create the full DIP set. Suppose the total number of DIPs in the set is Q and the number of concordant pairs of a BIQA model (meaning that the model predicts the correct preference) is Q_c, then a pairwise preference consistency ratio is defined as

$$\mathrm{PCR} = \frac{Q_c}{Q}\,. \tag{3.5}$$

PCR lies in $[0, 1]$ with a higher value indicating better performance.

3.2.4 Discussion

The above test criteria are defined independently of the database being created. Each of them challenges BIQA models from a different perspective. One would not be surprised to see that one model is superb under one criterion but subpar under another (as will be shown in Section 3.3). Meanwhile, all of them benefit from larger databases, where the weaknesses and failure cases of test models have more chances to be detected. These failure cases may provide insights on how to improve BIQA models.

3.3 BIQA Model Comparison

We apply the aforementioned test criteria on the Exploration database and compare the performance of 12 well-known BIQA models. These include 1) BIQI [102], 2) BLINDS-

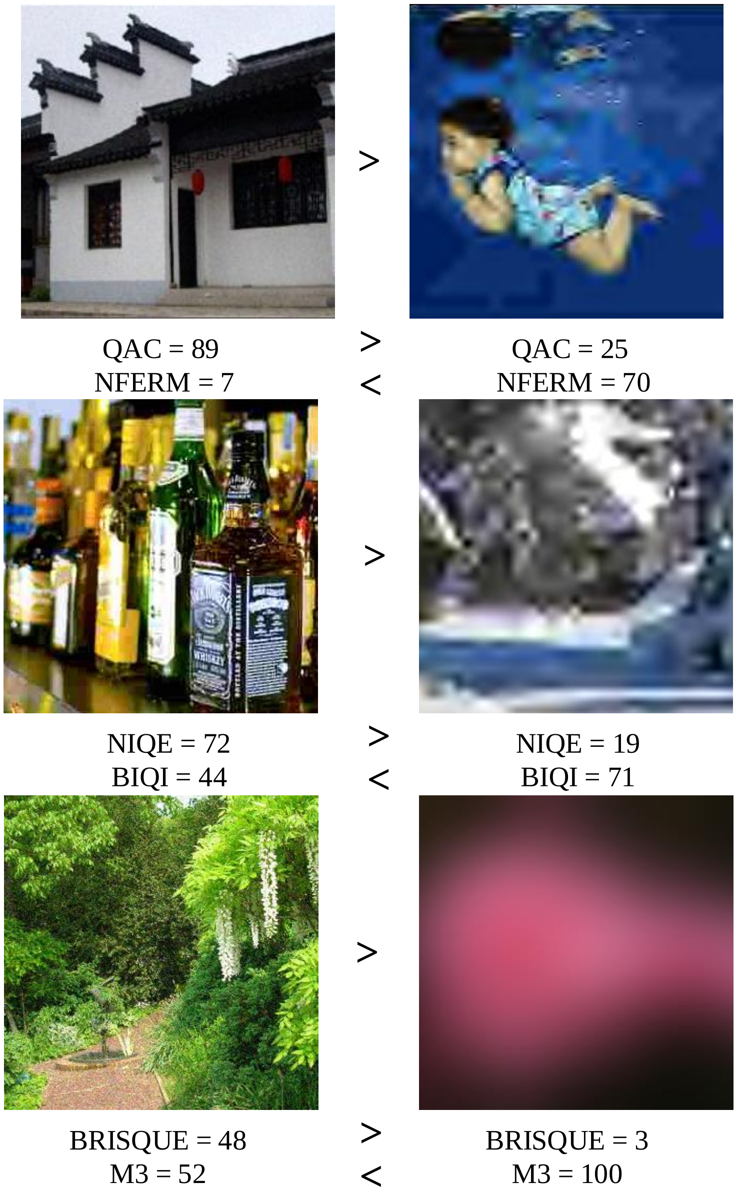


Figure 3.7: Sample DIPs from the Exploration database, where the left images have clearly better quality than that of the right images. A top performing BIQA model is able to give concordant opinions, whereas a less competitive model tends to perform randomly or provide discordant opinions.

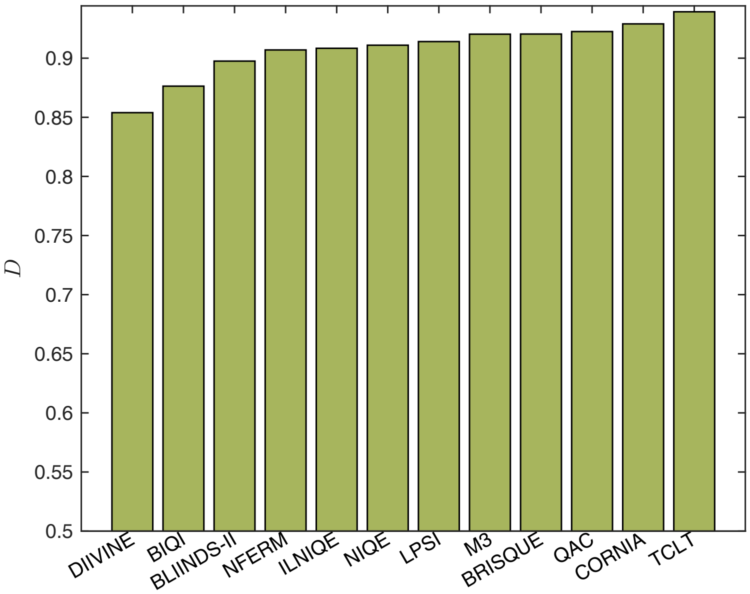
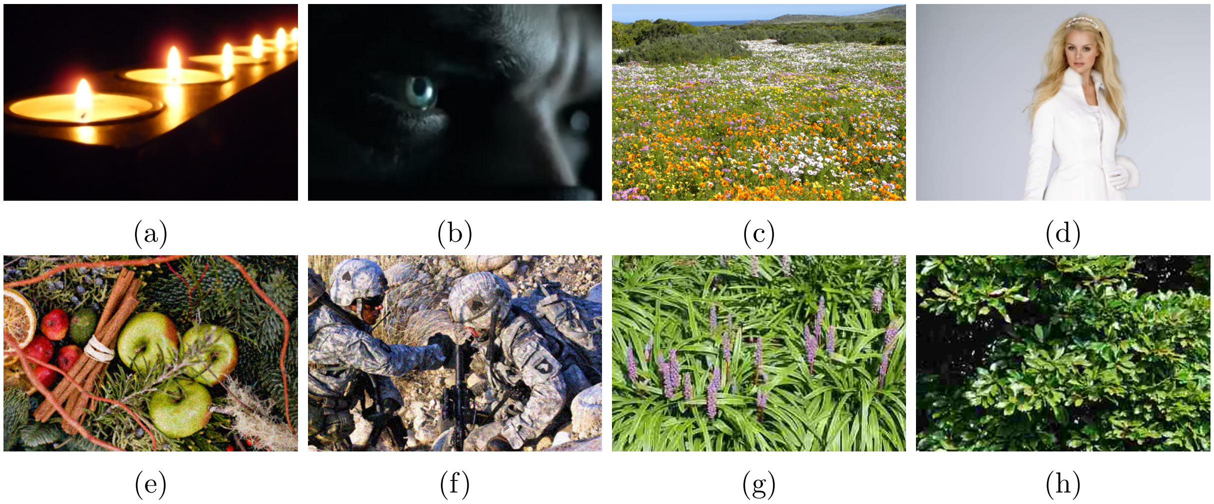


Figure 3.8: The D-test results of BIQA models on the Exploration database.

II [125], 3) BRISQUE [99], 4) CORNIA [184], 5) DIIVINE [103], 6) IL-NIQE [187], 7) LPSI [173], 8) M3 [177], 9) NFERM [38], 10) NIQE [101], 11) QAC [179], and 12) T-CLT [171]. The implementations of all algorithms are obtained from the original authors. For models that involve training, we use all images in the LIVE database [134] as the training set. We adopt a 4-parameter logistic nonlinear function [149] to map all model predictions to the MOS scale of LIVE [134]. As a result, the score range of all algorithms spans between $[0, 100]$, where a higher value indicates better perceptual quality.

3.3.1 D-Test Results

Fig. 3.8 shows the D-test results on the Exploration database of 12 BIQA measures. T-CLT [171], CORNIA [184], QAC [179], and BRISQUE [99] are among the top performing models. Despite their superior performance, by looking into their common failure cases, we are able to gain insights on their weaknesses. Some examples are shown in Fig. 3.9. In general, pristine images that are misclassified as distorted ones often exhibit low illumina-


(a) (b) (c) (d)

(e) (f) (g) (h)

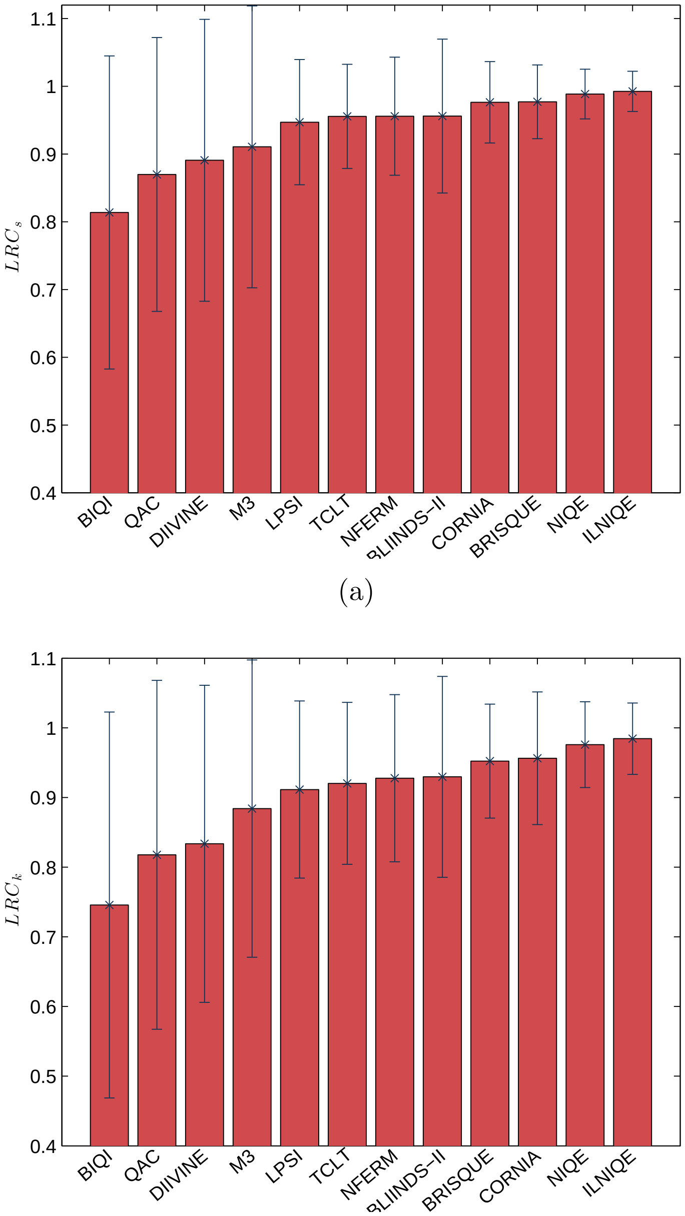
Figure 3.9: Failure cases of the top four BIQA models (TCLT [171], CORNIA [184], QAC [179], and BRISQUE [99]) in D-test on the Exploration database. (a)-(d) pristine images misclassified as distorted ones by the four models. (e)-(h) distorted images misclassified as pristine ones by the four models.

tion or low intensity variations. There are also exceptions. For example, complex textures as those in Fig. 3.9(c) resemble noise structures and may fool BIQA models. On the other hand, distorted images that are misclassified as pristine ones are often induced by white Gaussian noise and JPEG compression at mild distortion levels.

We also run D-test on LIVE [135] which has less than $1,000$ test images. The top performing BIQA models TCLT [171] and CORNIA [184] on the Exploration database perform perfectly on LIVE (achieving $D = 1$), which indicates no failure cases. This manifests the benefits of using the Exploration database which contains substantially more images to better distinguish BIQA models and to increase the chances of finding failure examples.

3.3.2 L-Test Results

We perform L-test on the Exploration database that includes $4,744 \times 4 = 18,976$ sets of images, each of which contains a list of images generated from the same source with the



(a)

(b)

Figure 3.10: The L-test results of BIQA models on the Exploration database. (a) LRC$_s$. (b) LRC$_k$.

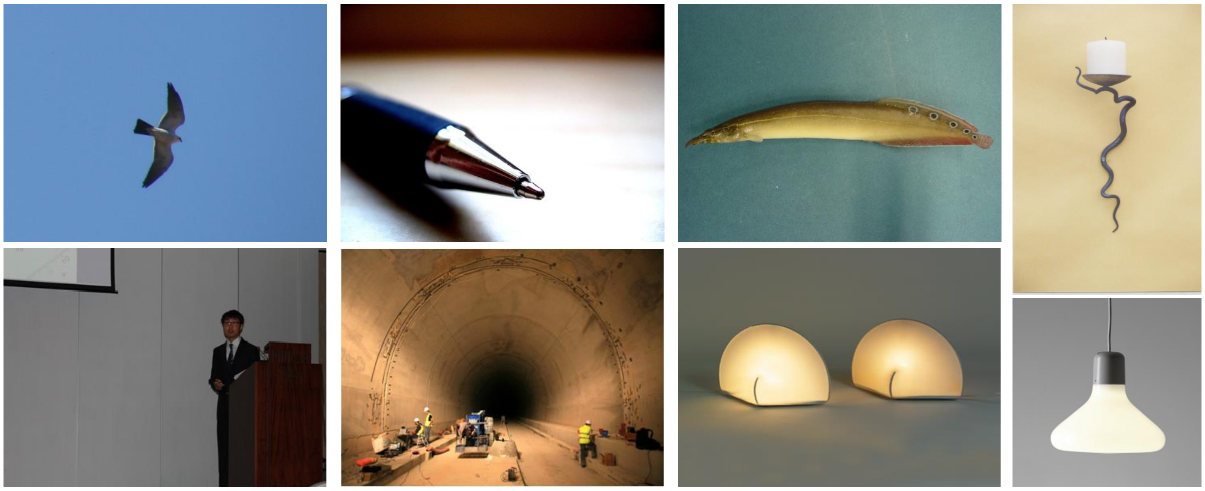


Figure 3.11: Failure cases of NIQE [101] in L-test induced by JPEG2000 compression on the Exploration database, where KRCC is less than 0.5.

same distortion type but at different distortion levels. Fig. 3.10 shows LRC$_s$ and LRC$_k$ results of 12 BIQA models, from which we have several observations. First, NIQE [101] and its feature enriched extension ILNIQE [187] outperform all other models. These methods are based on perception- and distortion-relevant NSS, without MOS for training. This reveals the power of NSS, which map images into a perceptually meaningful space for comparison. Second, although TCLT [171] performs the best in D-test, it is not outstanding in L-test. Third, training based models, such as BIQI [102] and DIIVINE [103] generally have lower overall consistency values and larger error bars, implying potential overfitting problems.

Furthermore, to demonstrate additional benefits of L-test, we focus on the best performing model NIQE [101], observe its main failure cases, and discuss how it can be improved. Fig. 3.11 shows sample failure cases which occur when JPEG2000 compression is applied. A common characteristic of these images is that they are combinations of strong edges and large smooth regions, which result in abundant ringing artifacts after JPEG2000 compression. The patch selection mechanism in NIQE [101] may mistakenly group such distorted structures to build the multi-variant Gaussian model, which may be close to that computed from a number of natural image patches. This explains the reverse orders of quality

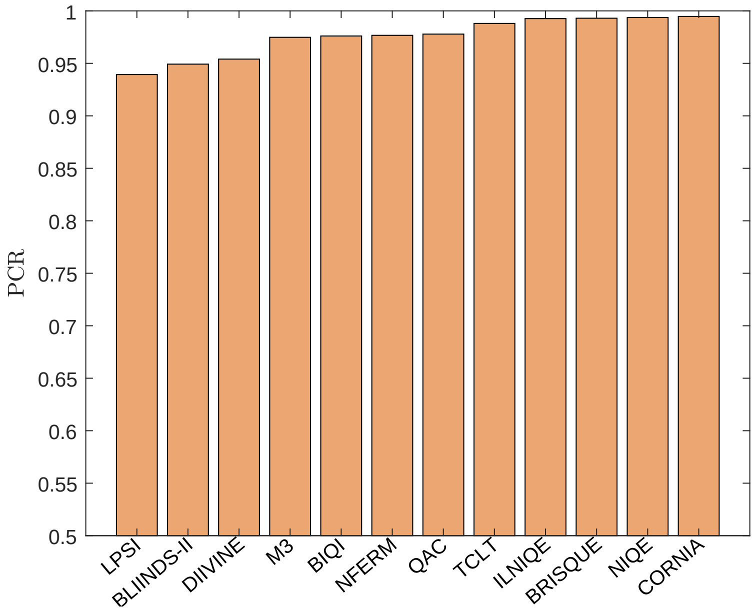


Figure 3.12: The P-test results of BIQA models on the Exploration database.

ranking. Potential ways of improving NIQE [101] include pre-screening ringing artifacts and training the model using natural image patches of more diverse content.

3.3.3 P-Test Results

We apply the proposed DIP generation engine on the Exploration database, resulting in more than 1 billion DIPs. Fig. 3.12 shows the pairwise preference consistency ratios of 12 BIQA measures, all of which achieve PCR $\geq$ 90%. This verifies the success of these algorithms in predicting image quality to a certain extent. Moreover, ILNIQE [187], NIQE [99], and CORNIA [184] are among top performing BIQA models, which conform to the results in L-test.

Although CORNIA [184] outperforms all other BIQA methods, it still gives $6,808,400$ wrong predictions. Representative failure cases are shown in Fig. 3.13. Specifically, CORNIA tends to favor artificial structures introduced in smooth regions, for example blocking structures in the sky in Fig. 3.13(a), and ringing around sharp edges in Fig. 3.13(e). This may be a consequence of its unsupervised feature learning mechanism that may not be capable of reliably differentiating real structures from artificially created distortions in smooth areas.

We also run P-test on LIVE [135] for comparison. Only $90,870$ DIPs have been generated, which is less than 0.01% of the DIPs generated from the Exploration database. All 12 BIQA algorithms perform perfectly on LIVE, achieving PCR = 1. This result manifests the value of the Exploration database and meanwhile shows the capability of P-test at exploiting large databases.

3.4 Summary

In this chapter, we build the Waterloo Exploration Database that contains $4,744$ pristine natural images and $94,880$ distorted images created from them. Furthermore, we present three test criteria (D-test, L-test, and P-test) and apply them to the Exploration database to assess 12 BIQA models, resulting in many useful findings.

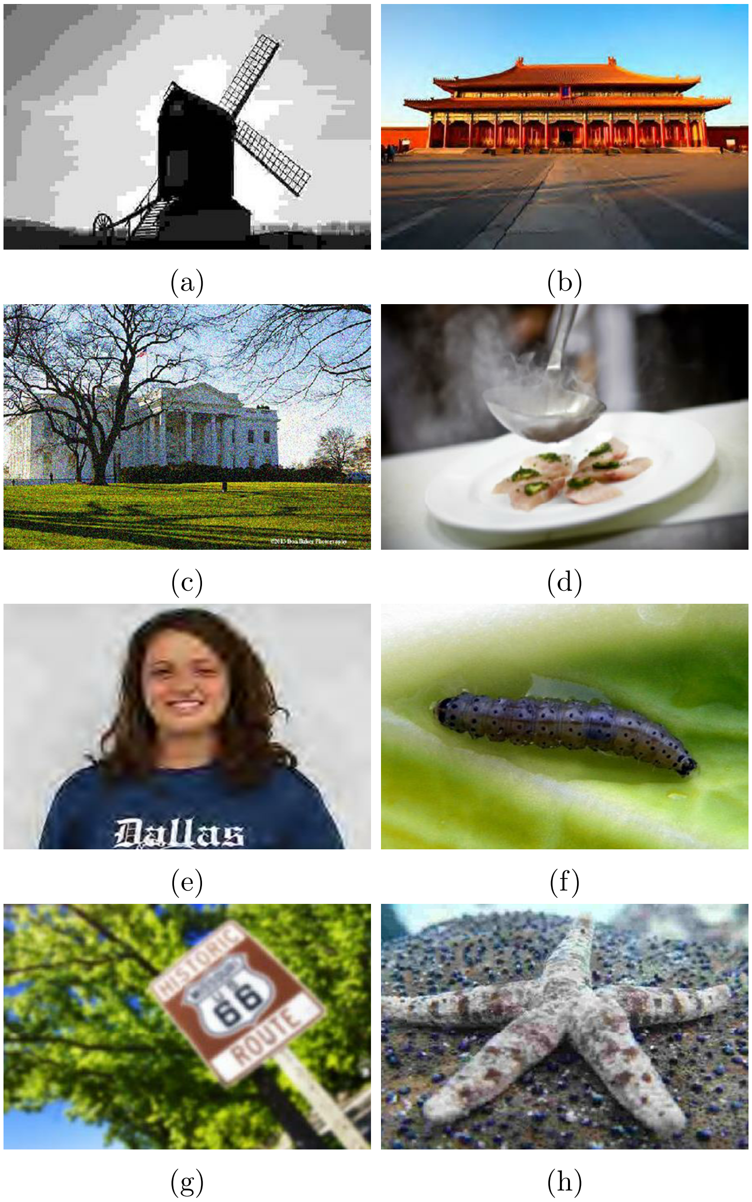

(a) (b)
(c) (d)
(e) (f)
(g) (h)

Figure 3.13: Failure cases of CORNIA [184] in P-test on the Exploration database. The left images have inferior quality compared with the right ones, but CORNIA [184] gives incorrect preference predictions. (a) CORNIA = 54. (b) CORNIA = 24. (c) CORNIA = 82. (d) CORNIA = 39. (e) CORNIA = 60. (f) CORNIA = 28. (g) CORNIA = 49. (h) CORNIA = 19.

Chapter 4

Group MAximum Differentiation Competition

In this chapter, we introduce the group MAximum Differentiation (gMAD) competition, a general methodology for comparing multiple computational models of a perceptually discriminable quantity. We adopt the fundamental idea behind the MAD competition [164] introduced by Wang and Simoncelli and extend it in several key ways. We apply gMAD to IQA and report the competition results on 16 objective IQA models.

4.1 gMAD Competition

Computational models of perceptual quantities are fundamental building blocks of man-made systems for processing sensory signals. With a group of such models of a perceptually discriminable quantity, how to efficiently compare their relative performance becomes a challenge. The standard approach in the literature is to select a number of samples from the stimulus space, collect subjective opinions for all stimuli, and compare model responses with collected subjective evaluations. A model that better accounts for the subjective data is superior. A major problem with this conventional methodology is the conflict between the possibly high dimensionality of the stimulus space and the limited scale of affordable subjective testing. For example, consider the space of all possible visual images. This stimulus space is of the same dimension as the number of pixels in the image, which is typically in the order of hundreds of thousands or millions. Since subjective testing is expensive and time-consuming, a typical "large-scale" subjective experiment allows for a maximum of a few thousand sample images to be examined, but no matter how these sample images are pre-selected, they are deemed to be extremely sparsely distributed in the image space.

Inspired by previous methods for efficient stimulus selection [167, 65, 113] and texture model assessment [27, 191, 117], Wang and Simoncelli proposed a novel psychophysical methodology, namely the MAD competition, to accelerate the model comparison process by minimizing the subjective experimental burden [164]. Given two computational models, MAD works by falsifying a model and one that is more difficult to be falsified is considered better. MAD gives us an opportunity to largely reduce the required number of test stimuli because ideally even one counterexample is sufficient to falsify a model. When generating stimuli that have great capability to discriminate between two models, MAD first synthesizes a pair of stimuli that maximize/minimize the responses of one model while holding the other fixed. The procedure is then repeated with the roles of two models exchanged. A visual demonstration in the context of IQA is shown in Fig. 4.1. Several limitations of MAD impede its wide usage in practical applications. First, although testing stimuli are automatically synthesized, MAD often relies on the gradient information of competing models to solve a constrained and often nonconvex optimization problem. This is not plausible for computational models whose gradients are difficult to compute, if not impossible. Second, MAD-generated stimuli may be highly unnatural [164], whose practical implications on how to improve existing models in real-world applications are limited. Third, it only allows two models into the competition and the extension to account for multiple models is nontrivial.

We adopt the fundamental idea behind MAD [164] and extend it in several key ways, toward an efficient and practical methodology for comparing multiple perceptual models, which we name the group MAximum Differentiation (gMAD) competition. When attempting to falsify a model (denoted as the defender), we work with a large set of pre-selected

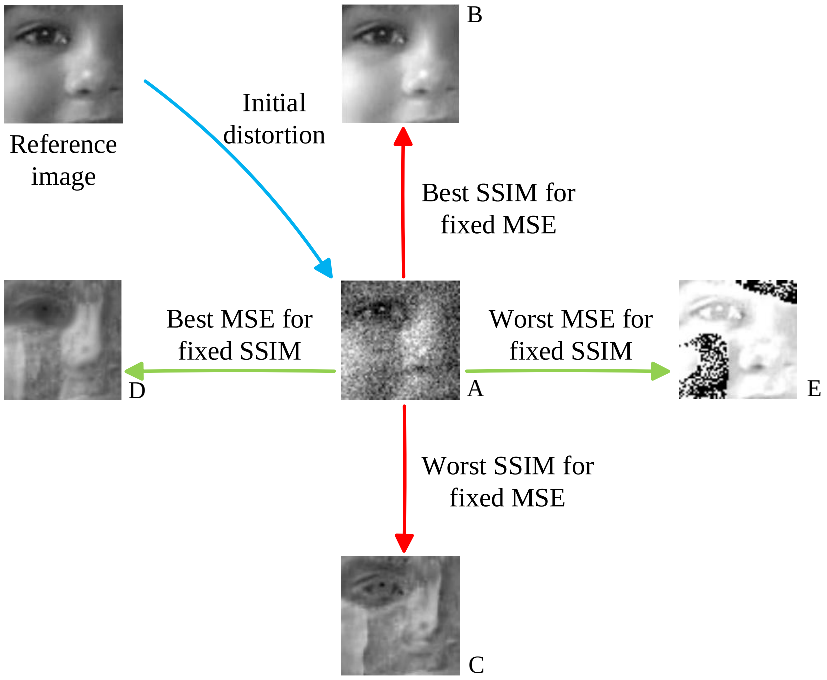


Figure 4.1: Illustration of the MAD competition method [164] with synthesized image pairs. Image by courtesy of Wang and Simoncelli [164].

stimuli rather than the whole stimulus space. This relaxation brings many advantages. First, it allows us to replace the iterative stimulus synthesis process with a search step in the stimulus set. As a result, gMAD is gradient-free, which permits almost all computational models into comparison. Second, we can easily specialize a competition by testing computational models on some specific types of stimuli. For example, if we compare multiple models of perceived image quality on a pre-selected image set, which contains only sharp and blurred images of different degrees, we in fact specialize a competition that compares the model abilities for blur perception [161]. Third, unlike MAD [164], we avoid generating highly unnatural stimuli that provide less practical insights in improving the competing models. Specifically, we first search for pairs of stimuli that maximize/minimize the responses of a group of other models (denoted by attackers) in the stimulus set, while fix responses of the defender. The attacks are optimal in the sense that the defender is most likely to be falsified. If instead, the defender survives from such an attack, it is a strong indicator that the defender is likely to be a robust and reliable model. gMAD runs this game among all models until every of them has played the defender role once. Subjective testing on pairs of such generated stimuli is then performed. To help summarize the relative performance of the competing models, we introduce notions of aggressiveness and resistance to indicate how strong an attacker in falsifying a defender and how resistant a defender to be defeated by an attacker, respectively. The pairwise aggressiveness and resistance measurements are aggregated into a global ranking using maximum likelihood of multiple options [147], which completes the gMAD competition. The framework is readily extensible, allowing future computational models to be added to the competition with minimally additional work.

We apply gMAD to the field of IQA [152, 153, 154] and report the competition results on 16 objective IQA models [157, 165, 189, 102, 125, 99, 184, 103, 187, 173, 177, 38, 101, 179, 171]. Careful investigations of selected extremal image pairs shed light on how to improve existing models and how to develop next-generation IQA models. To demonstrate the generality of the gMAD competition methodology, we investigate its usage in two more applications: image aesthetics evaluation [19] and video quality of experience (QoE) prediction [47], whose details are given in Appendix A.

4.1.1 Problem Formulation and General Methods

We first formulate the general model comparison problem. Suppose that we are interested in a stimulus space $\mathcal{S}$, from which we sample a set of stimuli $S \subset \mathcal{S}$ as our test set. We define a perceptually discriminable and continuous quantity $q(s) \in \mathcal{R}$ for all $s \in S$. A subjective assessment environment is assumed, where a human observer can compare the intensity of the perceptual quantity $q(s)$ for any stimulus $s \in S$ with the intensity of another stimulus s'. Given a group of computational models $\{Z_i\}_{i=1}^{M}$, each of which takes a stimulus s as input and predicts $q(s)$, we aim for comparing their relative performance in predicting the perceptual quantity $q(\cdot)$ based on a limited number of subjective tests.

The underlying philosophy in conventional approaches of model comparison is to validate a model. Subjective testing is first carried out on a set of stimuli. Average subjective opinions of those stimuli are then compared with model predictions, and a model that explains subjective data the best is the winner. Practicing such a philosophy requires the models to be validated using a sufficient number of test stimuli in the stimulus space, which is a major challenge because the stimulus space can be enormous, but the total number of test stimuli that can be obtained in a realistic subjective experiment is only in the order of thousands (if not fewer).

The MAD [164] and the current gMAD competition methodologies give up the conventional philosophy. Instead of trying to validate a model, MAD and gMAD attempt to falsify it, which have the freedom to explore the stimulus space $\mathcal{S}$ and a large set of stimuli S, respectively, before subjective testing. MAD [164] searches for optimal stimuli via synthesis, which involves a constrained optimization problem and typically calls for the gradient projection method. This requirement excludes models whose gradients are difficult to compute (*e.g.,* models are not continuous in the stimulus space). Different from MAD [164], gMAD does not rely on gradient computation and directly searches for stimulus pairs from a stimulus set S.

4.1.2 gMAD Competition Procedure

The details of the gMAD competition procedure are as follows. We work with a stimulus set $S = \{s_i\}_{i=1}^{N}$ that samples from a stimulus space $\mathcal{S}$ and compare M computational models $\{Z_i\}_{i=1}^{M}$of a perceptually discriminable quantity $q(\cdot)$. We divide $q(\cdot)$ into K intensity levels, within which the responses of the defender model are considered fixed.

- **Step 1**. Apply all M models to all stimuli in S, which results in a model prediction matrix $\hat{\mathbf{Q}}$ of M rows and N columns, where the entry $\hat{\mathbf{Q}}_{ij}$ is the prediction of $q(s_j)$ given by Z_i;

- **Step 2**. Choose Z_1 as the defender by setting the index $i = 1$. The rest of $M - 1$ models are the attackers;

- **Step 3**. Choose the first intensity level $k = 1$ from a total of K levels, where $k \in \{1, 2, \cdots, K\}$;

- **Step 4**. At the i-th row of $\hat{\mathbf{Q}}$, search for all stimuli whose model responses of Z_i (serve as the defender) lie in the k-th intensity level. This results in a subset of stimuli S_{ik}, which are considered to have similar perceptual quantities by the defender Z_i;

- **Step 5**. Choose Z_j as the current attacker from $M - 1$ attacker models, where $j \neq i$;

- **Step 6**. Within S_{ik}, find a pair of stimuli s^l_{ijk} and s^u_{ijk} that correspond to the minimal and maximal responses of Z_j. This stimulus pair is referred to as the gMAD counterexample suggested by the attacker Z_j, attempting to falsify the defender Z_i at the intensity level k;

- **Step 7**. Carry out a subjective quality discriminative test on s^l_{ijk} and s^u_{ijk}, whose details depend on the testing materials and will be given later;

- **Step 8**. Choose another model Z_j as the current attacker and repeat Steps 6-7 until all attacker models are exhausted;

- **Step 9**. Choose the next intensity level by setting $k = k + 1$ and repeat Steps 4-8 until $k = K$ (all intensity levels are exhausted);

- **Step 10**. Choose the next defender model Z_i by setting $i = i + 1$ and repeat Steps 3-9 until $i = M$ (all competing models are exhausted);

- **Step 11**. Carry out statistical analysis on the subjective test results, whose details will be given later.

Several useful features of the gMAD competition methodology are worth mentioning here. First, gMAD does not depend on the samplers for constructing S. In other words, it can be applied to any collection of stimuli. Second, the number of stimulus pairs selected by gMAD for subjective testing is $M(M-1)K$, independent of the size N of the stimulus set S. As a result, applying gMAD to a larger set has no impact on the cost of subjective testing. Third, each selected stimulus pair is associated with two computational models, which hold different opinions on their perceptual quantities. Specifically, the defender believes that the pair have the same perceptual quantity while the attacker suggests that they have very different perceptual quantities. If the stimulus pair are easily differentiated by human subjects, they constitute strong evidence against the defender model. On the other hand, if the pair indeed have similar perceptual quantities, they provide strong evidence to support the defender model. Fourth, it is straightforward and cost-effective to add new computational models into the competition. No change is necessary for all the selected pairs and their corresponding subjective testing. The only additional work is to select a total of $2MK$ new stimulus pairs for subjective testing, half of which are for the case that the new model acts as the defender and the other half as the attacker.

4.2 Application to IQA models

In this section, we present in detail how to apply the gMAD competition to computational models of perceived image quality. These include the selection of the image set and IQA models, the environment and flow of subjective testing, the preprocessing of subjective data, the definition of aggressiveness and resistance measures, and the aggregation from pairwise measurements to a global ranking.

4.2.1 Experimental Setup

For the test database, we choose the Waterloo Exploration Database. A total of 16 IQA models are selected to participate in the gMAD competition to cover a wide variety of IQA methodologies with emphasis on BIQA models. These include FR models 1) PSNR, 2) SSIM [157], 3) MS-SSIM [165] and 4) FSIM [189], and blind models 5) BIQI [102], 6) BLINDS-II [125], 7) BRISQUE [99], 8) CORNIA [184], 9) DIIVINE [103], 10) IL-NIQE [187], 11) LPSI [173], 12) M3 [177], 13) NFERM [38], 14) NIQE [101], 15) QAC [179]

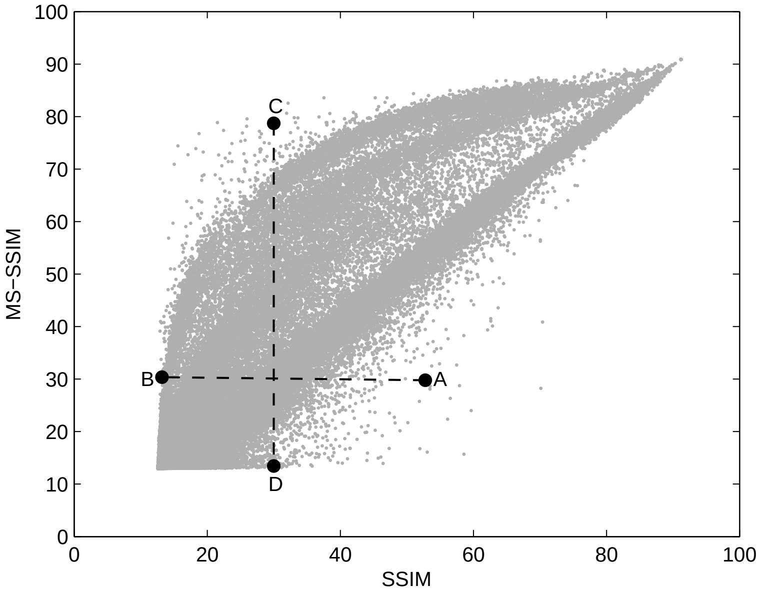
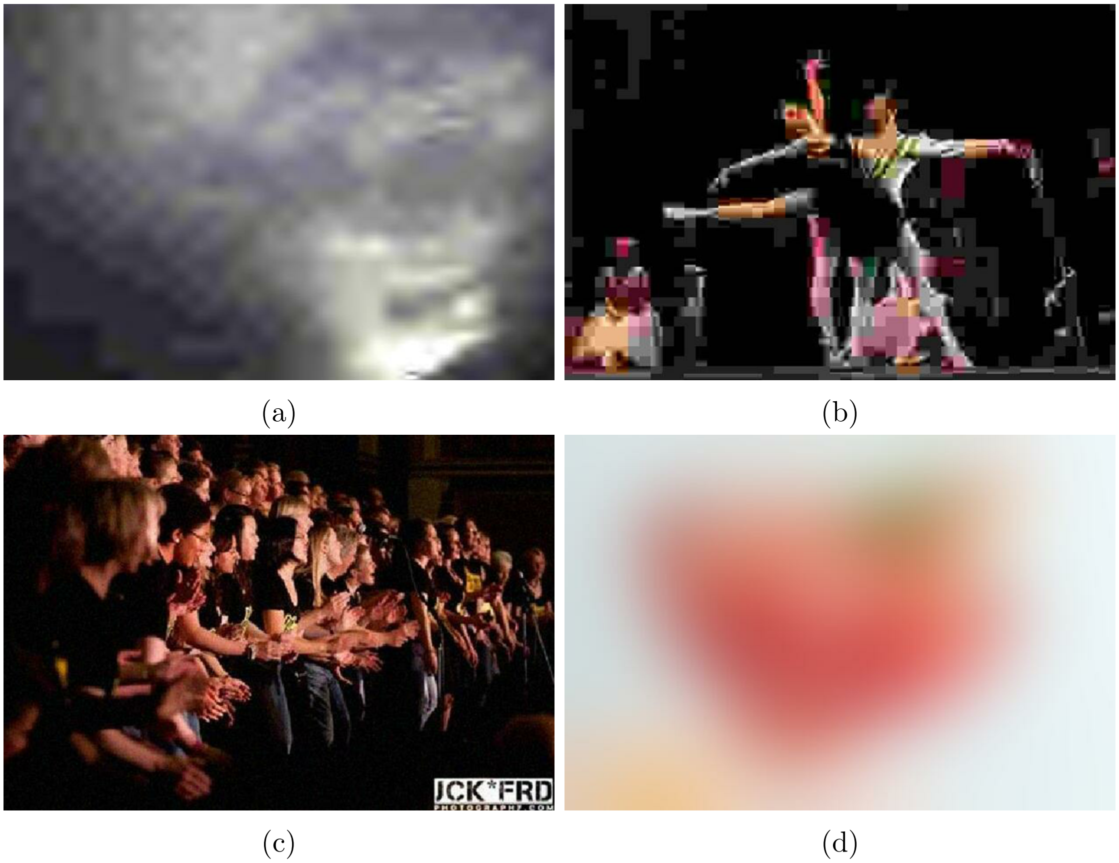


Figure 4.2: Extremal image pairs selected by gMAD on the Exploration database. A pair of images (A, B) is selected by maximizing/minimizing SSIM but holding MS-SSIM fixed. Similarly, a pair of images (C, D) is selected by maximizing/minimizing MS-SSIM but holding SSIM fixed.

and 16) TCLT [171]. The gradients of most models are extremely difficult to compute or approximate, therefore limiting the pairwise comparison using MAD [164]. The implementations of all algorithms are obtained from the original authors. For IQA models that involve training, we use all test images in the LIVE database [135] as the training set. To compensate for the nonlinearity of model predictions on the human perception of image quality and to make the comparison more consistent, we adopt a logistic nonlinear function as suggested in [149] to map all model predictions to the MOS scale of the LIVE database [135]. As a result, the score range of all algorithms spans between $[0, 100]$, where higher values indicate better perceptual quality.

For each defender model, we define 6 quality levels evenly spaced on the quality scale so that the selected subsets of images have a good coverage from low- to high-quality levels.



(a) (b)

(c) (d)

Figure 4.3: Image pairs found by gMAD between SSIM [157] and MS-SSIM [165] on the Exploration database. (a) MS-SSIM = 30, SSIM = 53. (b) MS-SSIM = 30, SSIM = 13. (c) SSIM = 30, MS-SSIM = 78. (d) SSIM = 30, MS-SSIM = 13.

The quality range within each subset of images is set to be within 1 std[1] of MOSs in the LIVE database [135]. Thus the images within the same subset have approximately the same quality by the defender model. The attacker models then choose extremal image pairs from the 6 subsets, as described previously. On the scatter plot, finding an extremal image pair corresponds to selecting points that have the longest distance in a given row or column, as exemplified in Fig. 4.2, where SSIM [157] competes with MS-SSIM [165]. The corresponding extremal image pairs are shown in Fig. 4.3, from which we can create a first

[1] Every image in the LIVE database has a MOS and an std associated with it, computed from all valid subjects. The std used here is in fact an average of stds for all images.

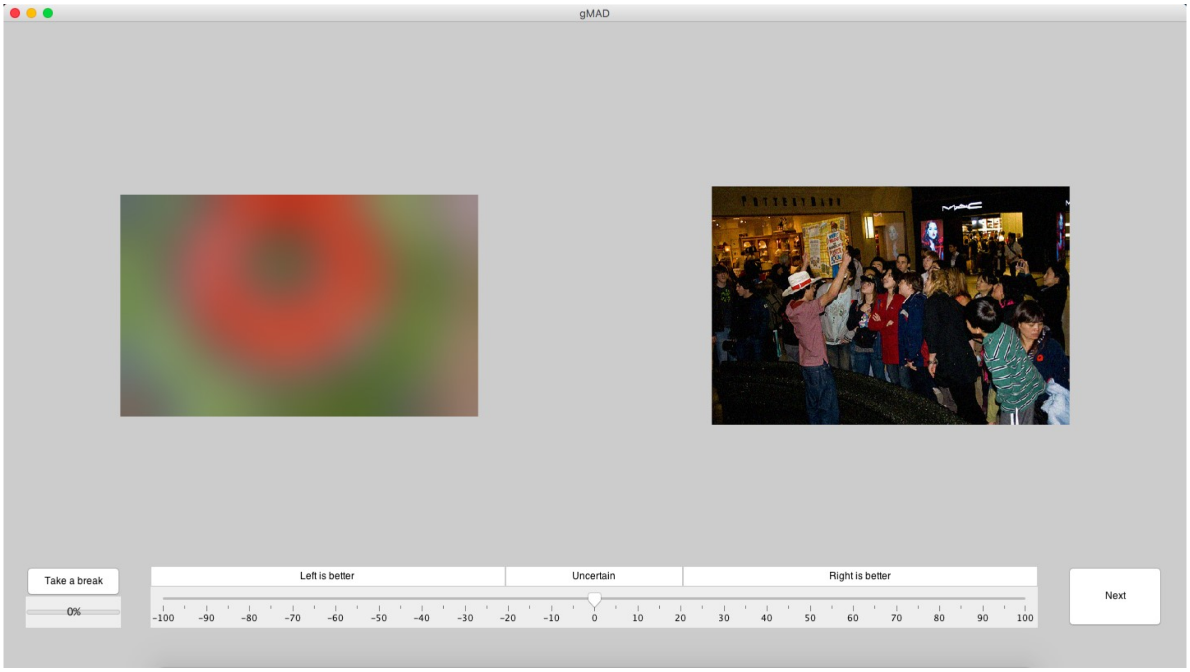


Figure 4.4: User interface for subjective testing.

impression on how the two models compete with each other. Specifically, images in the first row exhibit approximately the same perceptual quality (in agreement with MS-SSIM [165]) and those in the second row have drastically different perceptual quality (in disagreement with SSIM [157]). This may suggest that MS-SSIM is a solid improvement over SSIM. After the gMAD image pair selection process, a total of $16 \times (16 - 1) \times 6 = 1,440$ extremal image pairs are chosen for the subsequent subjective experiment.

4.2.2 Subjective Testing

A subjective user study is conducted in an office environment with a normal indoor illumination level and without the reflecting floor and ceiling walls. The display is a true-color LCD monitor at a resolution of $2,560 \times 1,600$ pixels and is calibrated in accordance with the recommendations of ITU-R BT.500 [149]. A customized MATLAB interface is created to render an image pair simultaneously at their exact pixel resolutions but in random spatial order. A scale-and-slider applet is used for assigning a quality score, as shown in Fig. 4.4. A total of 31 naïve subjects, including 16 males and 15 females, participate in the subjective experiment. All subjects have a normal or correct-to-normal visual acuity. Most of them are previously exposed to the general image processing field but do not have much experience in IQA. Each subject is introduced about the goal of the experiment, the experimental procedure, and the user interface. Sample extremal image pairs (independent of the test pairs) are shown to the subjects in a training session to familiarize them with image distortions and the whole test process. For each extremal image pair, the subject assigns a score between -100 and 100 to indicate his/her preference to either the left image $[-100, -20]$ (labeled as "left is better") or the right image $[20, 100]$ (labeled as "right is better"). In case the subject is uncertain about his/her decision, s/he can also assign a score between $[-20, 20]$ (labeled as "uncertain"), where a score of 0 indicates completely neutral. In contrast to the conventional paired comparison method, where the subject is only allowed to make a binary decision on his/her preference even when s/he is uncertain about the answer, our soft version of paired comparison better captures the subject's confidence when expressing his/her preference. During the experiment, subjects are allowed to move their positions to get closer or farther away from the screen for better observation. We divide the experiment into 4 sessions, each of which is limited to a maximum of 30 minutes, and subjects are asked to take a 5-minute break to minimize the influence of fatigue effect. All subjects participate in all sessions. Furthermore, in order to inspect if subjects are using consistent scoring strategies throughout the experiment, we repeat 10% of the image pairs (144 pairs to be specific) during the experiment.

4.2.3 Data Analysis

After collecting the raw subjective data, we adopt the outlier detection and subject rejection algorithm suggested in [149] to screen our pairwise data. Specifically, the raw score for an extremal image pair is considered to be an outlier if it is outside 2 stds of the mean score of that pair for the Gaussian case or outside $\sqrt{20}$ stds for the non-Gaussian case. A subject is removed if more than 5% of his/her evaluations are outliers. Moreover, a consistency check is conducted for each subject by making use of the image pairs that have been repeated. We define a consistency measure as the mean of stds of scores given by one subject to the repeated pairs. A subject is rejected if his/her consistency measure is more than 2 stds of consistency measures for all subjects. As a result, one subject is rejected due to inconsistent judgments. Among all scores given by the remaining valid subjects, about 1.4% of the total subjective evaluations are identified as outliers and are subsequently removed.

Since every extremal image pair in the gMAD competition are associated with two IQA models, we first compare these models in pairs and then aggregate the pairwise measurements into a global ranking using mature rank aggregation tools such as maximum likelihood for multiple options [147], hodgeRank [55], and ranking by eigenvectors [97]. We define the notions of aggressiveness and resistance. The aggressiveness measure a_{ij} indicates how strong the attacker model Z_i in falsifying the defender model Z_j and is computed by

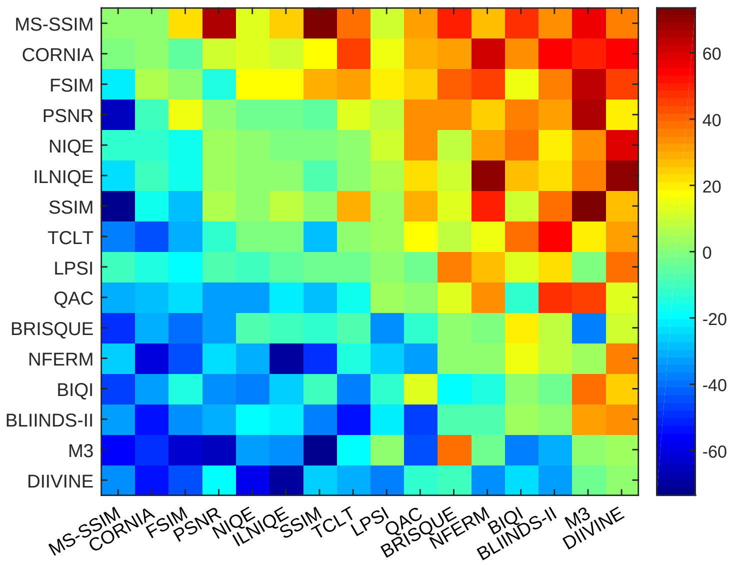
$$a_{ij} = \frac{\sum_{k=1}^{K} \theta_{jk} q_{ijk}}{\sum_{k=1}^{K} \theta_{jk}}\,, \tag{4.1}$$

where q_{ijk} is the MOS over all valid subjects on the extremal image pair selected from the k-th subset when Z_i and Z_j are the attacker and the defender, respectively. θ_{jk} is the number of sample images in the k-th subset, acting as a weighing factor. a_{ij} ranges between $[-100, 100]$ with a larger value indicating stronger aggressiveness of Z_i over Z_j. In general, a_{ij} is expected to be positive for a competitive model. It may also be negative, meaning that the order of the extremal image pair selected by Z_i contradicts to human judgments, which indicates a strong failure case of Z_i. The pairwise aggressiveness measurements of all models form an aggressiveness matrix $\mathbf{A}$.

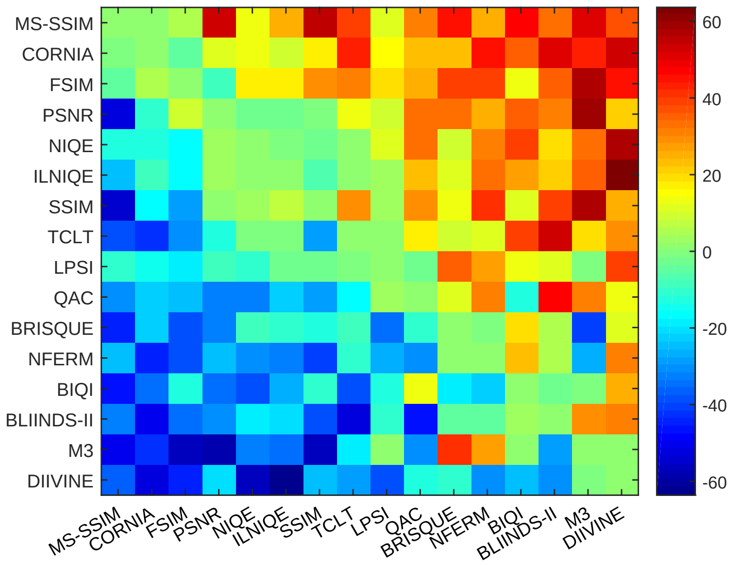
Different from aggressiveness, the resistance measure r_{ij} indicates how resistant the defender model Z_i to be defeated by the attacker model Z_j and is defined by

$$r_{ij} = \frac{\sum_{k=1}^{K} \theta_{ik}(100 - |q_{jik}|)}{\sum_{k=1}^{K} \theta_{ik}}\,. \tag{4.2}$$

r_{ij} ranges between $[0, 100]$ with a higher value indicating better resistance of Z_i as the defender against Z_j as the attacker. The pairwise resistance measurements of all models



(a)



(b)

Figure 4.5: Pairwise gMAD competition matrices: Each entry indicates the aggressiveness or the resistance of the row IQA model against the column IQA model. (a) Aggressiveness matrix. (b) Resistance matrix. $\mathbf{A} - \mathbf{A}^T$ and $\mathbf{R} - \mathbf{R}^T$ are drawn here for better visibility.

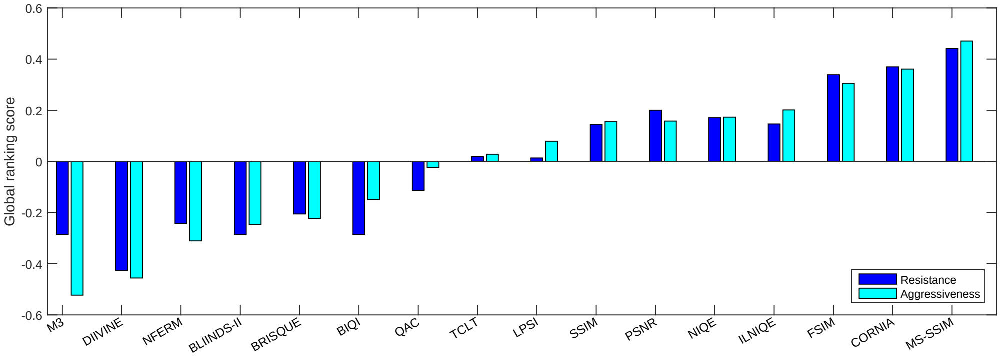


Figure 4.6: Global ranking of IQA models in terms of aggressiveness and resistance.

form a resistance matrix $\mathbf{R}$. We show the aggressiveness matrix $\mathbf{A}$ and the resistance matrix $\mathbf{R}$ of 16 IQA models from the gMAD competition in Fig. 4.5, where the higher value of an entry (warmer color), the stronger aggressiveness and resistance of the corresponding row model against the column model.

We aggregate the pairwise comparison results into a global ranking via a maximum likelihood method for multiple options [147]. Specifically, let $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_M] \in \mathcal{R}^M$ be the vector of ranking scores of M IQA models. We define the log-likelihood of $\boldsymbol{\mu}$ as

$$\mathcal{L}(\mathbf{A}|\boldsymbol{\mu}) = \sum_{ij} a_{ij} \log \left(\Phi(\mu_i - \mu_j) \right) \,, \tag{4.3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). To find the maximum likelihood solution, we need to solve

$$\begin{aligned} \underset{\boldsymbol{\mu}}{\text{argmax}} \quad & \sum_{ij} a_{ij} \log \left(\Phi(\mu_i - \mu_j) \right) \\ \text{subject to} \quad & \sum_{i} \mu_i = 0 \,. \end{aligned} \tag{4.4}$$

The constraint $\sum_i \mu_i = 0$ is added to resolve the translation ambiguity and therefore leads to a unique solution. Other constraints such as setting the first ranking score to zero $\mu_1 = 0$ are also applicable. It is not hard to show that (4.4) is a convex optimization problem [147] and can be efficiently solved using mature numerical optimization toolboxes such as CVX [17]. When $M = 2$, the maximum likelihood estimate reduces to the Thurstone's law [144, 83] and has a closed form solution (assuming $\mu_1 + \mu_2 = 0$)

$$\mu_1 = -\mu_2 = \Phi^{-1}\left(\frac{a_{12}}{a_{12} + a_{21}}\right) , \tag{4.5}$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of the standard normal. The pairwise resistance measurements can be aggregated in a similar fashion. We have also applied other ranking aggregation algorithms such as hodgeRank [55] and ranking by eigenvectors [97]. Very similar results are obtained.

We show the global ranking results in Fig. 4.6, from which we have several interesting observations. First, an IQA model that has stronger aggressiveness generally exhibits stronger resistance, which is confirmed by a high KRCC of 0.87 between them. Second, in general, FR-IQA models are more competitive than BIQA ones, which is not surprising because FR models make use of more information. Third, the best performance overall is obtained by MS-SSIM [165], which is a multi-scale version of SSIM [157] and significantly improves upon it. This suggests that incorporating multi-scale analysis is beneficial in improving the performance of IQA models. Fourth, CORNIA [184], NIQE [101] and its feature enriched version ILNIQE [187] perform the best among all BIQA models. They represent two popular NSS either hand-crafted or learned from data. Fifth, a model that is worth noting is LPSI [173], which essentially reduces the feature space to one dimension and without using MOSs for training, but it outperforms sophisticated machine learning-based approaches such as BRISQUE [99] and DIIVINE [103] which adopt several features for training. Finally, machine learning-based IQA models, though performed very well in existing publicly available databases, generally do not perform well in the current gMAD competition. This may be because the training samples are not sufficient to represent the population of real-world natural images and thus the risk of overfitting is high.

4.3 Summary

In this chapter, we propose a new methodology, namely the gMAD competition, for efficiently and practically comparing computational models of perceptually discriminable quantities. Working by falsifying models, gMAD automatically searches from a large-scale pre-selected stimulus set for a small and fixed number of model-dependent stimulus pairs. gMAD is particularly useful for comparing models that operate on a high dimensional stimulus space and that are mathematically not well-behaved. gMAD also provides two well-defined quantities (aggressiveness and resistance) to indicate the relative performance of competing models, through which useful insights can be gained to design better models.

Chapter 5

Blind Image Quality Assessment by Learning-to-Rank Discriminable Image Pairs

In this chapter, we first summarize the limitations of existing BIQA models, which motivate us to adopt learning-to-rank (L2R) algorithms to learn BIQA models. We show that a vast amount of reliable training data in the form of so-called quality-discriminable image pairs (DIP) can be generated at very low cost. We adopt a pairwise L2R algorithm (RankNet [8]) to learn BIQA models from DIPs. We also extend the pairwise learning paradigm to a listwise one and learn BIQA models using listwise L2R algorithms.

5.1 Pairwise L2R Approach for OU-BIQA

As discussed in Chapter 2, many BIQA models are developed by supervised learning [102, 125, 99, 184, 103, 173, 177, 38, 171] and share a common two-stage structure: 1) perception- and/or distortion-relevant features (denoted by $\mathbf{x}$) are extracted from a test image; and 2) a quality prediction function $f(\mathbf{x})$ is learned by statistical machine learning algorithms. The performance and robustness of these approaches rely heavily on the quality and quantity of the ground truth data for training. The most common type of ground truth data is in the form of the MOS, which is the average of quality ratings given by multiple subjects. Therefore, these models are often referred to as opinion-aware (OA) BIQA models and may incur the following drawbacks. First, collecting MOSs via subjective testing is expensive. As a result, even the largest publicly available IQA database, TID2013 [115], provides only $3,000$ images with MOSs. This limited number of training images is deemed extremely sparsely distributed in the entire image space [164]. As such, the generalizability of BIQA models learned from small training samples is questionable on real-world images. Second, among thousands of sample images, only a few dozen source reference images can be included, considering the combinations of reference images, distortion types and levels. For example, the TID2013 database [115] includes 25 source images only. It is extremely unlikely that this limited number of reference images sufficiently represent the variations that exist in real-world images. Third, since these BIQA models are trained with individual images to make independent quality predictions, the cost function is blind to the relative perceptual order between images. As a result, the learned models are weak at ordering images with respect to their perceptual quality.

In this chapter, we show that a vast amount of reliable training data in the form of so-called DIPs can be generated by exploiting large-scale databases with diverse image content. Each DIP is associated with a perceptual uncertainty measure to indicate the confidence level of its quality discriminability. We show that such DIPs can be generated at very low cost without resorting to subjective testing. We then employ RankNet [8], a neural network-based pairwise L2R algorithm [79, 39], to learn an opinion-unaware (OU) BIQA (meaning that no subjective opinions are used for training) model by incorporating the uncertainty measure into the loss function. Extensive experiments on four benchmark IQA databases demonstrate that the DIP inferred quality (dipIQ) indices significantly outperform previous OU-BIQA models. We also conduct another set of experiments in which we train the dipIQ indices using different feature representations as inputs and compare them with OA-BIQA models using the same representations. The generalizability and robustness of dipIQ are improved across all four IQA databases and verified by the gMAD competition methodology [90] on the Exploration database [86]. Furthermore, we extend the proposed pairwise L2R approach for OU-BIQA to a listwise L2R one by evoking

ListNet [9] (a listwise L2R extension of RankNet [8]) and transforming DIPs to quality-discriminable image lists (DIL) for training. The resulting DIL inferred quality (dilIQ) index leads to an additional performance gain.

5.1.1 DIP Generation

The DIP generation process presented here is similar to that in P-test, except that we do not enforce the constraint that a DIP is clearly quality-discriminable. Specifically, we first choose the three best-trusted FR-IQA models, namely MS-SSIM [165], VIF [132], and GSMD [180]. A nonlinear logistic function suggested in [134] is then adopted to map the predicted scores of the three models to the MOS scale of the LIVE database [135]. As a result, the score range of the three models spans $[0, 100]$, where higher values indicate better perceptual quality. We associate each candidate image pair with a nonnegative T, which is equal to the smallest score difference of the three FR models. Intuitively, the perceptual uncertainty level of quality discriminability should decrease monotonically with the increase of T. By varying T, we can generate a number of DIPs with different perceptual uncertainty levels of quality discriminability. To quantify the level of uncertainty, we employ a raised-cosine function given by

$$U(T) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\frac{\pi T}{T_c}\right)\right) & \text{if } T \leq T_c \\ 0 & \text{otherwise} \end{cases}, \tag{5.1}$$

where $U(T)$ lies in $[0, 1]$, with a higher value indicating a greater degree of uncertainty and T_c is a constant, above which the uncertainty is zero. In the current implementation, we set $T_c = 20$. Fig. 5.1 shows the uncertainty as a function of T and some representative DIPs, where the left images have better quality in terms of the three FR-IQA models with $T > 0$. All the shown DIPs are generated from the training image set that will be described later. It is clear that setting T close to zero produces the highest level of uncertainty of quality discriminability. Careful inspection of the two image pairs at the top of Fig. 5.1 reveals that the uncertainty manifests itself in two ways. First, the right image at the top left pair has better perceived quality to many human observers compared with the

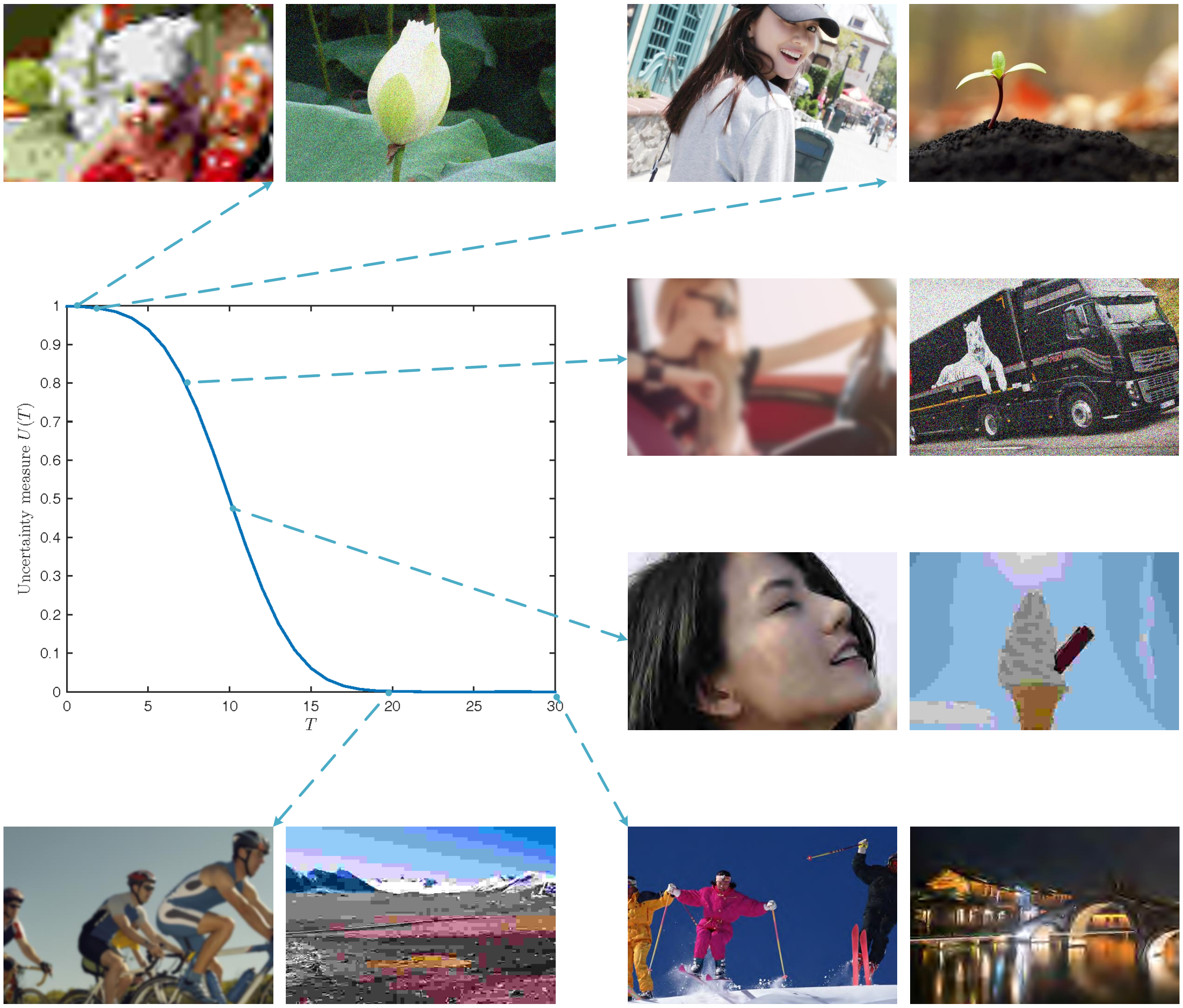


Figure 5.1: Illustration of the perceptual uncertainty of quality discriminability of DIPs as a function of T. The left images of all DIPs have better quality in terms of the three FR-IQA models with $T > 0$. However, the quality discriminability differs significantly. All images are originated from the 700 training images.

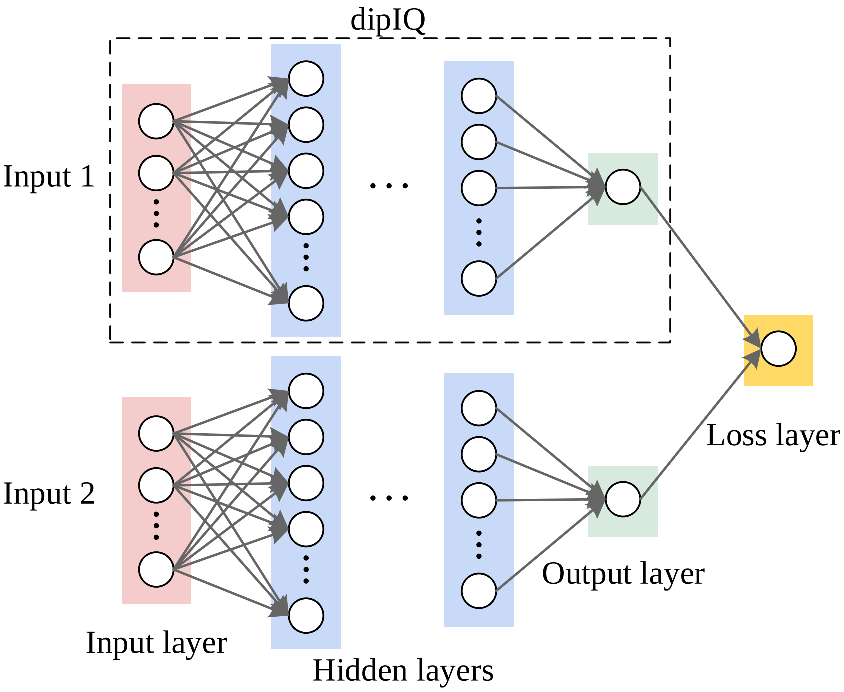


Figure 5.2: The architecture of RankNet [8].

left one, which disagrees with the three FR-IQA models. Second, both images in the top right pair have distortions that are barely perceived by the human eye. In other words, they have very similar perceptual quality. The perceptual uncertainty generally decreases if T increases and when $T > 20$, the DIP is clearly discriminable, further justifying the selection of $T_c = 20$.

5.1.2 RankNet

Given a number of DIPs, a pairwise L2R algorithm would make use of their perceptual order to learn quality models while taking the inherent perceptual uncertainty into account. Here, we revisit RankNet [8], a pairwise L2R algorithm that was the first of its kind used by commercial search engines [79]. We extend it to learn from DIPs associated with uncertainty. Fig. 5.2 shows RankNet's architecture, which is based on classical neural networks and has two parallel streams to accommodate a pair of inputs. The two-stream weights are shared, which is achieved by using the same initializations and the same gradients during backpropagation [8]. The quality prediction function $f(\mathbf{x})$, namely the dipIQ index, is implemented by one of the streams, and the loss function is defined on a pair of images with the help of f. Specifically, let $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ be the outputs of the first and second streams, whose difference is converted to a probability using

$$\hat{p}_{ij}(f) = \frac{\exp\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)}{1 + \exp\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)}\,, \tag{5.2}$$

based on which we define the cross entropy loss as

$$\begin{aligned} &\ell(f; \mathbf{x}_i, \mathbf{x}_j, p_{ij}) = -p_{ij} \log \hat{p}_{ij} - (1 - p_{ij}) \log(1 - \hat{p}_{ij}) \\ &= -p_{ij}\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right) + \log\left(1 + \exp\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)\right)\,, \end{aligned} \tag{5.3}$$

where p_{ij} is the ground truth label associated with the training pair, consisting of the i-th and j-th images. In the case of DIPs used in our work, p_{ij} is always 0 or 1, indicating that the quality of the i-th image is worse or better than the j-th one. Within the mini-batch stochastic gradient minimization framework, we define the empirical loss function using the perceptual uncertainty of each DIP as a weighting factor

$$\ell_b(f) = \sum_{\langle i,j \rangle \in \mathcal{B}} (1 - U_{ij}) \ell(f; \mathbf{x}_i, \mathbf{x}_j, p_{ij})\,, \tag{5.4}$$

where $\mathcal{B}$ is the batch containing the DIP indices currently being trained. As Eq. (5.4) makes clear, DIPs with higher uncertainty contribute less to the overall loss. With some derivations, we obtain the gradient of ℓ_b with respect to the model parameters collectively denoted by $\mathbf{w}$ as follows

$$\begin{aligned} \frac{\partial \ell_b(f)}{\partial \mathbf{w}} = \sum_{\langle i,j \rangle \in \mathcal{B}} &\left(-p_{ij} + \frac{\exp\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)}{1 + \exp\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right)}\right) \\ &\left(1 - U_{ij}\right)\left(\frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}} - \frac{\partial f(\mathbf{x}_j)}{\partial \mathbf{w}}\right)\,. \end{aligned} \tag{5.5}$$

In the case of a linear dipIQ containing no hidden layers and no nonlinear activations, Eq. (5.3) is reduced to

$$\begin{aligned}\ell(\mathbf{w};\mathbf{x}_i,\mathbf{x}_j,p_{ij}) = &- p_{ij}\left(\mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j)\right)\\ &+ \log(1+\exp(\mathbf{w}^T(\mathbf{x}_i-\mathbf{x}_j))\,,\end{aligned} \tag{5.6}$$

which is easily recognized as logistic regression. The convexity of Eq. (5.6) ensures the global optimality of the solution. We investigate both linear and nonlinear dipIQ cases with the cross entropy as loss. In fact, other probability distribution measures can also be adopted as alternatives. For example, Tsai *et al.* [146] proposed a fidelity loss measure from quantum physics. We find in our experiments that the fidelity loss impairs performance, so we use the cross entropy loss throughout the chapter.

We select RankNet [8] as our first choice of pairwise L2R algorithms for two reasons. First, it is capable of handling a large number of training samples using stochastic or mini-batch gradient descent algorithms. By contrast, the training of other pairwise L2R methods such as RankSVM [57], even with a linear kernel, is painfully slow. Second, since RankNet [8] embodies classical neural network architectures, we embrace the latest advances in training deep neural networks [44, 68] and can easily upscale the network by adding more hidden layers to learn powerful nonlinear quality prediction functions.

5.1.3 Implementation Details

Training Set Construction: We download 840 high-quality and high-resolution natural images from `http://www.ivsky.com/` to represent scenes we see in the real-world. They can be roughly clustered into seven groups: human, animal, plant, landscape, cityscape, still-life, and transportation. Sample source images are shown in Fig. 5.3. We preprocess each source image by down-sampling it using a bicubic kernel so that its maximum height or width is 768. After that, we add four distortion types (JPEG compression, JPEG2000 compression, white Gaussian noise contamination, and Gaussian blur) with five distortion levels to each source image, following the procedures described in Chapter 3. As a result, our training set consists of $17,640$ images. We randomly hold out 140 source images and

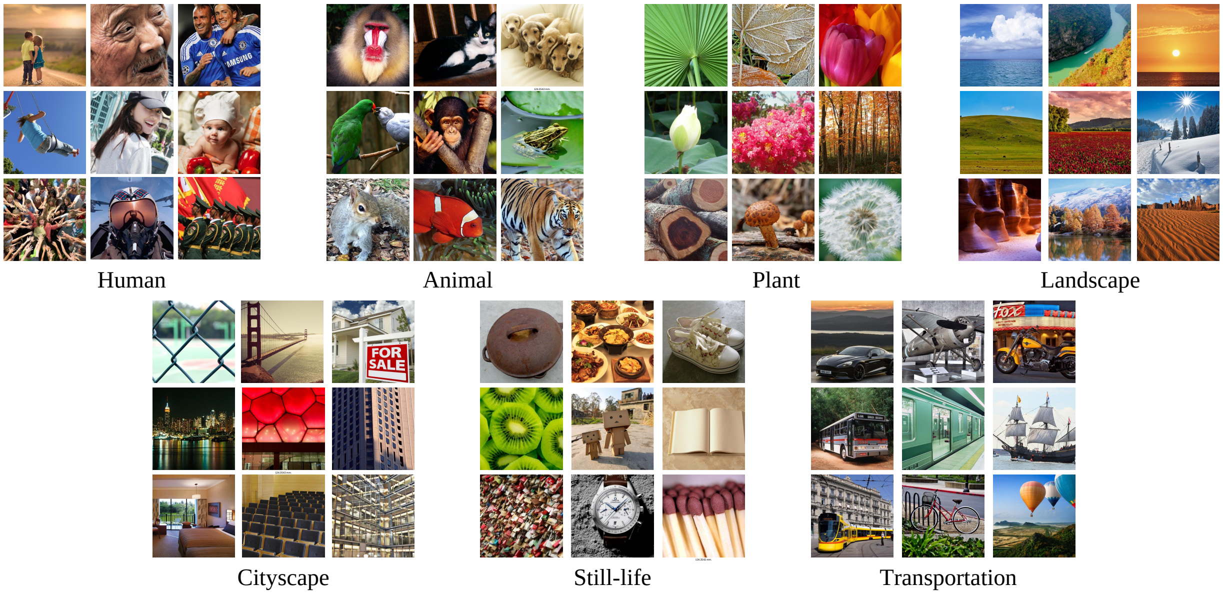


Figure 5.3: Sample source images in the training set. All images are cropped for better visibility.

their corresponding distorted images and use them as the validation set. For the rest $14,700$ images, we adopt the proposed DIP generation engine to produce more than 80 million DIPs, which constitute our training data.

Base Feature: We adopt the feature representation in CORNIA [181] as input because it performs the best among 12 BIQA models and even outperforms three FR-IQA models in the gMAD competition on the Exploration database, as shown in Chapter 4. In addition, one of the top performing OU-BIQA models, BLISS [183] also adopts CORNIA features and trains on synthetic scores. As such, we offer a fair testing bed to compare dipIQ learned by a pairwise L2R approach (RankNet [8]) with BLISS [183] learned by a regression method (SVR).

RankNet Instantiation: We investigate both linear and nonlinear dipIQ models, denoted by dipIQ* and dipIQ, respectively. The input dimension to RankNet is $20,000$, equaling the feature dimension in CORNIA [184]. The loss layer is implemented by the cross entropy function in Eq. (5.3). For dipIQ*, the input layer is directly connected to the output layer without adding hidden layers or going through nonlinear transforms. The use of the cross entropy loss ensures the convexity of the optimization problem. For dipIQ, we add 3 hidden layers, which have a 256 - 128 - 3 structure. All layers are fully connected, followed by ReLU [105] as the nonlinearity activation. We choose the node number of the third hidden layer to be 3 so that we can visualize the three-dimensional embedding of test images. Other choices are somewhat ad hoc, and a more exploration of alternative architectures could potentially lead to further performance improvements.

The RankNet training procedure generally follows Simonyan and Zisserman [139]. Specifically, the training is carried out by optimizing the cross entropy function using mini-batch gradient descent with momentum. The weights of the two streams in RankNet are shared. The batch size is set to 512, and momentum to 0.9. The training is regularized by weight decay (the ℓ_2 penalty multiplier set to 5×10^{-4}). The learning rate is fixed to 10^{-4}. Since we have a plenty of DIPs (more than 80 million) for training, each DIP is exposed to the learning algorithm once. The learning stops when the entire set of DIPs have been swept. The weights that achieve the lowest validation set loss are used for testing.

5.1.4 Experimental Protocols

Databases: Four IQA databases are used to compare dipIQ with state-of-the-art BIQA measures. They are LIVE [135], CSIQ [69], TID2013 [115], and the Exploration database. The first three are small databases that are widely adopted to benchmark objective IQA models. Each test image is associated with an MOS to represent its perceptual quality. In our experiments, we only consider the distortion types shared by them, namely JP2K, JPEG, WN, and BLUR. As a result, LIVE [135], CSIQ [69], and TID2013 [115] contain 634, 600, and 500 test images, respectively. The Exploration database has been described in Chapter 3. Although MOSs are not available, innovative evaluation criteria are employed to compare BIQA measures on the Exploration database.

Evaluation Criteria: We use five evaluation criteria to compare the performance of BIQA measures. The first two are included in previous tests carried out by the video quality experts group (VQEG) [149]. Others are proposed in Chapter 3 to take into account image databases without MOSs. Details are given as follows.

- Spearman's rank-order correlation coefficient (SRCC) [149] is defined as

  $$\mathrm{SRCC} = 1 - \frac{6\sum_i \Delta_i^2}{N(N^2-1)}\,, \tag{5.7}$$

  where N is the number of images in the database and Δ_i is the difference between the i-th image's ranks in MOSs and model predictions.

- Pearson linear correlation coefficient (PLCC) [149] is computed as

  $$\mathrm{PLCC} = \frac{\sum_i (q_i - q_m)(\hat{q}_i - \hat{q}_m)}{\sqrt{\sum_i (q_i - q_m)^2}\sqrt{\sum_i (\hat{q}_i - \hat{q}_m)^2}}\,, \tag{5.8}$$

  where q_i and $\hat{q}_i$ stand for the MOS and the model prediction of the i-th image, respectively.

- D-test (Chapter 3.2.1).

- L-test (Chapter 3.2.2).

- P-test (Chapter 3.2.3).

SRCC and PLCC are applied to LIVE [135], CSIQ [69] and TID2013 [115], while D-test, L-test and P-test are applied to the Exploration database. Note that the use of PLCC requires a nonlinear function $\tilde{q} = (\beta_1 - \beta_2)/(1 + \exp(-(\hat{q} - \beta_3)/|\beta_4|)) + \beta_2$ to map the raw model predictions to the MOS scale[1]. Following Mittal *et al.* [99] and Ye *et al.* [183], in our experiments we randomly choose 80% reference images along with their corresponding distorted images to estimate $\{\beta_i | i = 1, 2, 3, 4\}$ and use the rest 20% images for testing. This procedure is repeated $1,000$ times and the median SRCC and PLCC values are reported.

[1] The modified logistic regression function $\tilde{q} = \beta_1(\frac{1}{2} - \frac{1}{\exp{(\beta_2(\hat{q}-\beta_3))}}) + \beta_4\hat{q} + \beta_5$ is not used here because it does not necessarily preserve the monotonicity of the nonlinear mapping [149]. Nevertheless, two nonlinear mappings give very similar median SRCC and PLCC values.

Table 5.1: Median SRCC and PLCC results across $1,000$ sessions on LIVE [135]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.908	0.894	0.984	0.814	0.883
SSIM [157]	0.961	0.974	0.970	0.952	0.947
QAC [179]	0.876	0.951	0.925	0.911	0.869
NIQE [101]	0.924	0.945	0.972	**0.941**	0.920
ILNIQE [187]	0.901	0.944	**0.979**	0.927	0.918
BLISS [183]	0.925	**0.956**	0.967	0.936	0.945
dipIQ*	**0.946**	**0.956**	**0.976**	**0.962**	**0.952**
dipIQ	**0.956**	**0.969**	0.975	0.940	**0.958**

PLCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.912	0.896	0.987	0.812	0.874
SSIM [157]	0.968	0.980	0.972	0.951	0.937
QAC [179]	0.876	0.960	0.895	0.912	0.855
NIQE [101]	0.932	0.956	**0.979**	**0.951**	0.912
ILNIQE [187]	0.912	0.966	0.976	0.936	0.913
BLISS [183]	0.933	**0.972**	0.978	0.948	0.945
dipIQ*	**0.958**	0.953	0.951	**0.950**	**0.948**
dipIQ	**0.964**	**0.980**	**0.983**	0.948	**0.957**

5.1.5 Experimental Results

Comparison with FR and OU-BIQA Models: We compare dipIQ with two well-known FR-IQA models: PSNR (whose largest value is clipped at 60 dB in order to perform a reasonable parameter estimation) and SSIM [157] (whose implementation used in the chapter involves a down-sampling process [158]) and previous OU-BIQA models, including QAC [179], NIQE [101], ILNIQE [187], and BLISS [183]. The implementations of QAC [179], NIQE [101], and ILNIQE [187] are obtained from the original authors. To the best of our knowledge, the complete implementation of BLISS [183] is not publicly available. Therefore, to make a fair comparison we train BLISS [183] on the same 700 reference images and their distorted versions, which have been used to train dipIQ. The labels are synthesized using the method in [183]. The training toolbox and parameter settings are inherited from the original paper [183].

Table 5.2: Median SRCC and PLCC results across $1,000$ sessions on CSIQ [69]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.941	0.901	0.943	0.936	0.928
SSIM [157]	0.962	0.956	0.912	0.965	0.935
QAC [179]	0.884	0.913	0.850	0.839	0.840
NIQE [101]	0.926	0.882	0.836	0.908	0.883
ILNIQE [187]	0.924	0.905	0.867	0.867	0.887
BLISS [183]	0.932	**0.927**	0.879	0.922	0.920
dipIQ*	**0.938**	0.926	**0.887**	**0.925**	**0.924**
dipIQ	**0.944**	**0.936**	**0.904**	**0.932**	**0.930**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.954	0.908	0.961	0.937	0.918
SSIM [157]	0.973	0.983	0.908	0.956	0.930
QAC [179]	0.898	0.942	0.865	0.855	0.847
NIQE [101]	0.944	0.946	0.824	0.935	0.900
ILNIQE [187]	0.942	0.956	0.880	0.903	0.914
BLISS [183]	0.954	0.970	0.895	0.947	0.939
dipIQ*	**0.955**	**0.971**	**0.903**	**0.951**	**0.946**
dipIQ	**0.959**	**0.975**	**0.927**	**0.958**	**0.949**

Tables 5.1, 5.2, and 5.3 list comparison results between dipIQ and existing OU-BIQA models in terms of median SRCC and PLCC values on LIVE [135], CSIQ [69], and TID2013 [115], respectively. Both dipIQ* and dipIQ outperform all previous OU-BIQA models on LIVE [135] and CSIQ [69], and are comparable to ILNIQE [187] on TID2013 [115]. Although both dipIQ* and BLISS [183] learn a linear prediction function using CORNI-A features as inputs [184], we observe consistent performance gains of dipIQ* across all three databases over BLISS [183]. This may be because dipIQ* learns from more reliable data (DIPs) with uncertainty weighting, whereas the training labels (synthetic scores) for BLISS are noisier, as exemplified in Fig. 5.4. It is not hard to observe that Fig. 5.4(a) has clearly worse perceptual quality than Fig. 5.4(b), which in turn has approximately the same quality compared with Fig. 5.4(c). Both two cases are in disagreement with synthetic scores [183].

To ascertain that the improvement of dipIQ is statistically significant, we carry out a

Table 5.3: Median SRCC and PLCC results across $1,000$ sessions on TID2013 [115]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
PSNR	0.898	0.929	0.942	0.965	0.924
SSIM [157]	0.950	0.935	0.896	0.969	0.924
QAC [179]	0.883	0.885	0.668	0.879	0.837
NIQE [101]	0.901	0.873	0.854	0.821	0.812
ILNIQE [187]	**0.912**	0.873	**0.890**	0.815	**0.881**
BLISS [183]	0.906	0.893	0.856	0.872	0.836
dipIQ*	0.909	**0.903**	0.854	**0.884**	0.857
dipIQ	**0.926**	**0.932**	**0.905**	**0.922**	**0.877**
PLCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**ALL4**
PSNR	0.933	0.925	0.963	0.958	0.911
SSIM [157]	0.970	0.968	0.902	0.958	0.927
QAC [179]	0.892	0.929	0.719	0.877	0.829
NIQE [101]	0.912	0.928	0.859	0.848	0.819
ILNIQE [187]	0.929	0.944	**0.899**	0.816	0.890
BLISS [183]	0.930	**0.963**	0.863	0.872	0.862
dipIQ*	**0.937**	**0.963**	0.851	**0.892**	**0.894**
dipIQ	**0.948**	**0.973**	**0.906**	**0.928**	**0.894**

two sample T-test (with a 95% confidence) between PLCC values obtained by different models on LIVE [135]. After comparing every possible pair of OU-BIQA models, the results are summarized in Table 5.4, where a symbol "1" means the row model performs significantly better than the column model, a symbol "0" means the opposite, and a symbol "-" indicates that the row and column models are statistically indistinguishable. It can be observed that dipIQ is statistically better than dipIQ*, which is better than all previous OU-BIQA models.

Table 5.5 shows the results on the Exploration database. dipIQ* and dipIQ outperform all previous OU-BIQA models in D-test and P-test, and are competitive in L-test, whose performance is slightly inferior to NIQE [101] and ILNIQE [187]. By learning from examples with a variety of image content, dipIQ is able to reduce the number of incorrect preference predictions in P-test down to around $130,000$ out of more than 1 billion candidate DIPs.

To gain intuition to why the generalizability of dipIQ is excellent even without MOSs

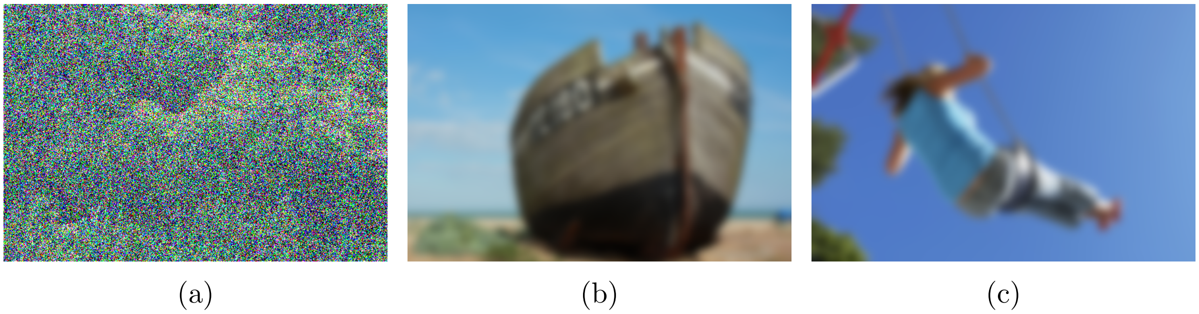


Figure 5.4: The noisiness of the synthetic score [183]. (a) Synthetic score = 10. (b) Synthetic score = 10. (c) Synthetic score = 40. (a) has clearly worse perceptual quality than (b), which in turn has approximately the same quality compared with (c). Both two cases are in disagreement with the synthetic score [183]. Images are selected from the training set.

for training, we visualize the three-dimensional embedding of the LIVE database [135] in Fig 5.5, using the learned three-dimensional features from the third hidden layer of dipIQ. We can see that the learned representation is able to cluster test images according to the distortion type, and meanwhile align them with respect to their perceptual quality in a meaningful way, where high-quality images are clamped together regardless of image content.

Comparison with OA-BIQA Models: In the second set of experiments, we train dipIQ using different feature representations as inputs and compare with OA-BIQA models using the same feature representations and MOSs for training. BRISQUE [99] and DIIVINE [103] are selected as representative features extracted from the spatial and wavelet domain, respectively. We also compare dipIQ with CORNIA [184], whose features are adopted as the default input to dipIQ. We re-train BRISQUE [99], DIIVINE [103], and CORNIA [184] on the LIVE database, whose learning tools and parameter settings follow their respective papers. We adjust the dimension of the input layer of dipIQ to accommodate features of different dimensions and train them on the 700 reference images and their distorted versions, as described previously. All models are tested on CSIQ [69], TID2013 [115], and the Exportation database [86]. From Tables 5.6, 5.7, and 5.8, we observe that dipIQ consistently performs better than the corresponding OA-BIQA model on

Table 5.4: Statistical significance matrix based on the hypothesis testing. A symbol "1" means that the performance of the row model is statistically better than that of the column model, a symbol "0" means that the row model is statistically worse, and a symbol "-" means that the row and column models are statistically indistinguishable

PLCC	PSNR	SSIM	QAC	NIQE	ILNIQE	BLISS	dipIQ*	dipIQ
PSNR	-	0	1	0	0	0	0	0
SSIM [157]	1	-	1	1	1	0	0	0
QAC [179]	0	0	-	0	0	0	0	0
NIQE [101]	1	0	1	-	-	0	0	0
ILNIQE [187]	1	0	1	-	-	0	0	0
BLISS [183]	1	1	1	1	1	-	0	0
dipIQ*	1	1	1	1	1	1	-	0
dipIQ	1	1	1	1	1	1	1	-

CSIQ [69] and the Exploration database, and is comparable on TID2013 [115]. The reason we do not obtain noticeable performance gains on TID2013 [115] may be that TID2013 [115] has 18 references images originated from LIVE [135], based on which the OA-BIQA models have been trained. This creates dependencies between training and testing sets. We may also draw conclusions about the effectiveness of the feature representations based on their performance under the same pairwise L2R framework: generally speaking, CORNIA [184] features > BRISQUE [99] features > DIIVINE [103] features.

We further compare dipIQ^B and BRISQUE [99] using the gMAD competition methodology on the Exploration database. Specifically, we first find a pair of images that have the maximum and minimum dipIQ^B values from a subset of images in the Exploration database, where BRISQUE [99] rates them to have the same quality. We then repeat this procedure, but with the roles of dipIQ^B and BRISQUE [99] exchanged. The two image pairs are shown in Fig. 5.6, from which we conclude that images in the first row exhibit approximately the same perceptual quality (in agreement with dipIQ^B) and those in the second row have drastically different perceptual quality (in disagreement with BRISQUE [99]). This verifies that the robustness of dipIQ^B is significantly improved over BRISQUE [99] using the same feature representation and MOSs for training. Similar gMAD competition results are obtained across all quality levels, and for dipIQ^D versus DIIVINE [103] and

Table 5.5: The D-test, L-test, and P-test results on the Exploration database. #IPP: number of incorrect predictions in P-test

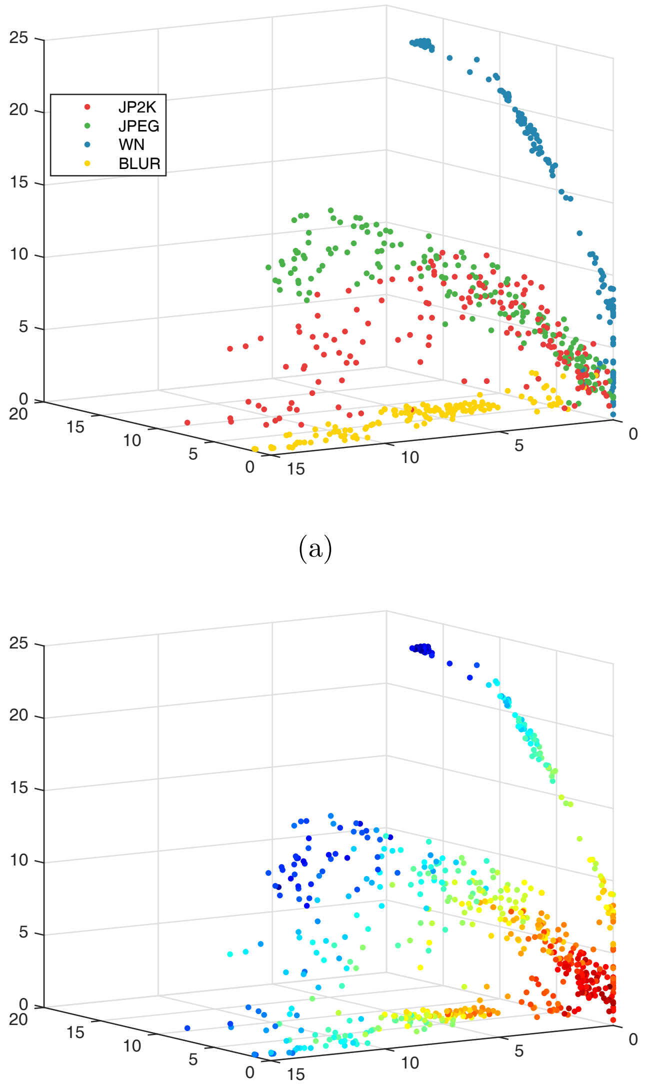
	D-test	L-test	P-test	#IPP
PSNR	1.0000	1.0000	0.9995	620,071
SSIM [157]	1.0000	0.9992	0.9991	1,131,457
QAC [179]	0.9226	0.8699	0.9779	28,447,590
NIQE [101]	0.9109	**0.9885**	0.9937	8,127,941
ILNIQE [187]	0.9084	**0.9926**	0.9927	9,435,319
BLISS [183]	0.9080	0.9801	0.9996	562,925
dipIQ*	**0.9209**	0.9863	**0.9996**	**465,069**
dipIQ	**0.9346**	0.9846	**0.9999**	**129,668**

dipIQ versus CORNIA [184].

In summary, the proposed pairwise L2R approach is proved to learn OU-BIQA models with improved generalizability and robustness compared with OA-BIQA models using the same feature representations and MOSs for training.

5.2 Listwise L2R Approach for OU-BIQA

We extend the proposed pairwise L2R approach for OU-BIQA to a listwise L2R one. Specifically, we first construct three-element DILs by concatenating DIPs. For example, given two DIPs $\langle i, j\rangle$ and $\langle j, k\rangle$ with the same level of uncertainty, we create a list $\langle i, j, k\rangle$ with the ground truth label $p_{ijk} = 1$, indicating that the quality of the i-th image is better than the j-th image, whose quality is better than the k-th image. The uncertainty level is transferred as well. We then employ ListNet [9], a listwise L2R extension of RankNet [8] to learn OU-BIQA models. The major differences between ListNet and RankNet are twofold. First, ListNet can have multiple streams with the same weights to accommodate a list of inputs, where each stream is implemented by a classical neural network architecture similar to RankNet, as shown in Fig. 5.2. We instantiate a three-stream ListNet to fit three-element DILs. Second, the loss function of ListNet is defined using the concept of permutation probability. More specifically, we define a permutation $\pi = \langle \pi(1), \pi(2), \ldots, \pi(N)\rangle$ on a list



(a)

(b)

Figure 5.5: Three dimensional embedding of the LIVE database [135]. (a) Color encodes distortion type. (b) Color encodes quality; the warmer, the better. The learned features from the third hidden layer of dipIQ are able to cluster images according to the distortion type, and meanwhile align them according to their perceptual quality in a meaningful way.

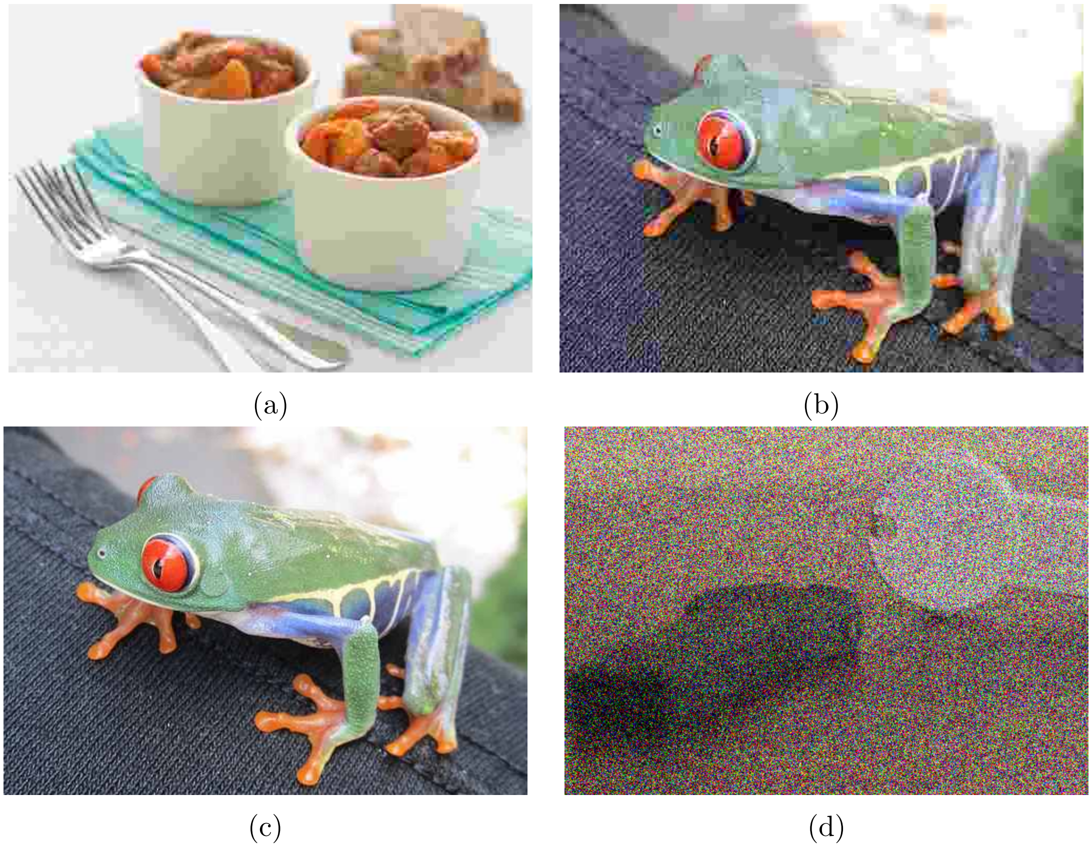

(a) (b) (c) (d)

Figure 5.6: The gMAD competition between dipIQ^B and BRISQUE [99]. (a) best BRISQUE for fixed dipIQ^B. (b) worst BRISQUE for fixed dipIQ^B. (c) best dipIQ^B for fixed BRISQUE. (d) worst dipIQ^B for fixed BRISQUE.

Table 5.6: Median SRCC and PLCC results across $1,000$ sessions, trained on LIVE [135] and tested on CSIQ [69]. The superscripts B and D indicate that the input features of dipIQ are from BRISQUE [99] and DIIVINE [103], respectively

SRCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [99]	0.894	0.916	**0.934**	0.915	0.909
dipIQ^B	**0.938**	**0.938**	**0.934**	**0.943**	**0.926**
DIIVINE [103]	0.844	0.819	0.881	0.884	0.835
dipIQ^D	**0.930**	**0.939**	**0.904**	**0.920**	**0.912**
CORNIA [184]	0.916	0.919	0.787	0.928	0.915
dipIQ	**0.944**	**0.936**	**0.904**	**0.932**	**0.930**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [99]	0.937	0.960	**0.947**	0.936	0.937
dipIQ^B	**0.956**	**0.974**	0.945	**0.959**	**0.943**
DIIVINE [103]	0.898	0.818	0.903	0.909	0.855
dipIQ^D	**0.949**	**0.973**	**0.924**	**0.944**	**0.942**
CORNIA [184]	0.947	0.960	0.777	0.953	0.934
dipIQ	**0.959**	**0.975**	**0.927**	**0.958**	**0.949**

of N instances as a bijection from $\{1, 2, .., N\}$ to itself, where $\pi(j)$ denotes the instance at position j in the permutation. The set of all possible permutations of N instances is termed as Π. We define the probability of permutation π given the list of predicted scores $\{f(\mathbf{x}_i)\}$ as

$$\hat{p}_\pi(f) = \prod_{j=1}^{N} \frac{\exp\big(f(\mathbf{x}_{\pi(j)})\big)}{\sum\limits_{k=j}^{N} \exp\big(f(\mathbf{x}_{\pi(k)})\big)}\,, \tag{5.9}$$

which satisfies $\hat{p}_\pi(f) > 0$ and $\sum_{\pi\in\Pi} \hat{p}_\pi(f) = 1$ as proved in [9]. The loss function can then be defined as the cross entropy function between the ground truth and permutation probabilities

$$\ell(f; \{\mathbf{x}_i\}, \{p_\pi\}) = -\sum_{\pi\in\Pi} p_\pi \log(\hat{p}_\pi)\,. \tag{5.10}$$

When $N = 2$, the loss function of ListNet [9] in Eq. (5.10) becomes equivalent to that of RankNet [8] in Eq. (5.3). In the case of three-element DILs, we have $p_\pi = 1$, if $\pi = \langle i, j, k\rangle$

Table 5.7: Median SRCC and PLCC results across $1,000$ sessions, trained on LIVE [135] and tested on TID2013 [115]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [99]	0.906	0.894	0.889	0.886	**0.883**
dipIQ^B	**0.927**	**0.921**	**0.921**	**0.917**	**0.883**
DIIVINE [103]	0.857	0.680	0.879	0.859	0.795
dipIQ^D	**0.912**	**0.889**	**0.887**	**0.905**	**0.872**
CORNIA [184]	0.907	0.912	0.798	**0.934**	**0.893**
dipIQ	**0.926**	**0.932**	**0.905**	0.922	0.877

PLCC	JP2K	JPEG	WN	BLUR	ALL4
BRISQUE [99]	0.919	0.950	0.886	0.884	**0.901**
dipIQ^B	**0.942**	**0.957**	**0.923**	**0.906**	0.883
DIIVINE [103]	0.901	0.696	**0.882**	0.860	0.794
dipIQ^D	**0.945**	**0.947**	0.881	**0.896**	**0.892**
CORNIA [184]	0.923	0.960	0.778	**0.934**	**0.904**
dipIQ	**0.948**	**0.973**	**0.906**	0.928	0.894

and $p_\pi = 0$ otherwise. Therefore, the loss function in Eq. (5.10) can be simplified as

$$\begin{aligned}\ell(f;\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k,p_{ijk}) &= -f(\mathbf{x}_i) - f(\mathbf{x}_j)\\ &+ \log\left(\sum_{l\in\{i,j,k\}} \exp\left(f(\mathbf{x}_l)\right)\right) + \log\left(\sum_{l\in\{j,k\}} \exp\left(f(\mathbf{x}_l)\right)\right),\end{aligned} \tag{5.11}$$

base on which we define the batch-level loss as

$$\ell_b(f) = \sum_{\langle i,j,k\rangle\in\mathcal{B}} (1 - U_{ijk})\ell(f;\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k,p_{ijk}), \tag{5.12}$$

where U_{ijk} is the uncertainty level of the list, transferred from the corresponding DIPs. The gradients of Eq. (5.12) w.r.t. the parameters $\mathbf{w}$ can be easily derived. Note that ListNet [9] does not add new parameters.

We generate 50 million DILs from the available DIPs as the training data for ListNet [9]. The training procedure is exactly the same as training RankNet [8]. The training stops

Table 5.8: The D-test, L-test, and P-test results on the Exploration database, trained on LIVE [135]

	D-test	L-test	P-test	#IPP
BRISQUE [99]	0.9204	**0.9772**	0.9930	9,004,685
dipIQ^B	**0.9265**	0.9753	**0.9996**	**503,911**
DIIVINE [103]	0.8538	0.8908	0.9540	59,053,011
dipIQ^D	**0.9191**	**0.9588**	**0.9983**	**2,124,199**
CORNIA [184]	0.9290	0.9764	0.9947	6,808,400
dipIQ	**0.9346**	**0.9846**	**0.9999**	**129,668**

Table 5.9: Median SRCC and PLCC results across $1,000$ sessions on LIVE [135] using ListNet [9]

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.956	0.969	0.975	0.940	0.958
dilIQ	0.956	0.966	**0.976**	**0.953**	0.958
PLCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.964	0.980	0.983	0.948	0.957
dilIQ	0.964	0.978	**0.985**	**0.956**	0.954

when the entire set of image lists have been swept once. The weights that achieve the lowest validation set loss are used for testing.

We list the comparison results between the DIL inferred quality (dilIQ) index trained by ListNet [9] and the baseline dipIQ on LIVE [135], CSIQ [69], TID2013 [115], and the Exploration database in Tables 5.9, 5.10, 5.11, and 5.12, respectively. Remarkable performance improvements have been achieved on CSIQ and TID2013. This may be because the ranking position information is made explicit to the learning process. dilIQ is comparable to dipIQ on LIVE and the Exploration database.

Table 5.10: Median SRCC and PLCC results across $1,000$ sessions on CSIQ [69] using ListNet [9]

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.944	0.936	0.904	0.932	0.930
dilIQ	0.930	0.925	0.893	**0.939**	**0.936**
PLCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**ALL**
dipIQ	0.959	0.975	0.927	0.958	0.949
dilIQ	0.954	0.968	0.920	**0.960**	**0.954**

Table 5.11: Median SRCC and PLCC results across $1,000$ sessions on TID2013 [115] using ListNet [9]

SRCC	JP2K	JPEG	WN	BLUR	ALL
dipIQ	0.926	0.932	0.905	0.922	0.877
dilIQ	0.918	0.849	0.905	**0.925**	**0.891**
PLCC	**JP2K**	**JPEG**	**WN**	**BLUR**	**ALL**
dipIQ	0.948	0.973	0.906	0.928	0.894
dilIQ	0.948	0.923	0.903	**0.929**	**0.915**

5.3 Summary

In this chapter, we propose an OU-BIQA model, namely dipIQ, using RankNet [8]. The inputs to the dipIQ training model are an enormous number of DIPs, not obtained by expensive subjective testing but automatically generated with the help of most-trusted FR-IQA models at low cost. Extensive experimental results demonstrate the effectiveness of the proposed dipIQ indices with higher accuracy and improved robustness in content variations. We also learn an OU-BIQA model, namely dilIQ, using a listwise L2R approach, which achieves an additional performance gain.

Table 5.12: The D-test, L-test, and P-test results on the Exploration database using ListNet [9]

	D-test	L-test	P-test	#IPP
dipIQ	0.9346	0.9846	0.9999	129,668
dilIQ	0.9346	**0.9893**	0.9998	198,650

Chapter 6

End-to-end Blind Image Quality Assessment Using Deep Neural Networks

In this chapter, we leverage the recent advances in deep neural networks (DNN) and propose a multi-task DNN for BIQA. We first summarize the drawbacks of previous studies in end-to-end optimization of BIQA. We then present in detail the decomposition of BIQA into two different but related subtasks and the construction of our multi-task DNN.

6.1 Motivations

As is clear from Chapter 2, early BIQA models are mainly based on hand-crafted features [103, 99, 125, 26], which rely heavily on knowledge of the probabilistic structures of our visual world, the mechanisms of image degradations, and the functionalities of the HVS [150, 138]. Built upon feature representations, a quality prediction function is learned using MOSs. Typically, the knowledge-driven feature extraction and data-driven quality prediction stages are designed separately. With the recent exciting development of DNN methodologies [68], a fully data-driven end-to-end BIQA solution becomes possible.

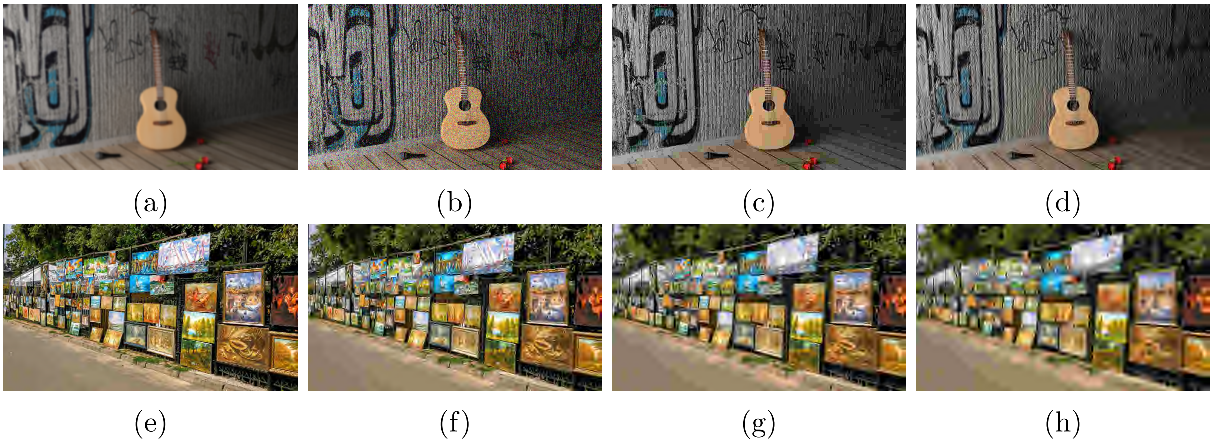


Figure 6.1: Images (a)-(d) with different distortion types have similar quality while images (e)-(h) of the same distortion type have different quality, according to our subjective testing. (a) Gaussian blurring. (b) Gaussian noise contamination. (c) JPEG compression. (d) JPEG2000 compression. (e)-(h) JPEG2000 compression with increasing compression ratios from left to right.

Although DNN has shown great promises in many vision tasks [68, 139, 42], end-to-end optimization of BIQA is challenging due to the lack of sufficient ground truth samples for training. Previous DNN-based BIQA methods tackle this challenge in three ways. Methods of the first kind [4] directly inherit the architectures and weights from pre-trained networks for general image classification tasks [123] followed by fine-tuning. The performance and efficiency of such networks depend highly on the generalizability and relevance of the tasks used for pre-training. The second kind of methods [59, 60, 5] work with image patches by assigning the MOS of an image to all patches within it. This approach suffers from three limitations. First, the concept of quality without context (*e.g.*, the quality of a single 32×32 patch) is not well defined [150, 156]. Second, local image quality within context (*e.g.*, the quality of a 32×32 patch within a large image) varies across spatial locations even when the distortion is homogeneously applied [157]. Third, patches with similar statistical behaviors (*e.g.*, smooth and blurred regions) may have substantially different quality [87]. Methods of the third kind [62] make use of FR-IQA models for quality annotation. Their performance is directly affected by that of FR-IQA models, which may be inaccurate across distortion levels [90] and distortion types [115].

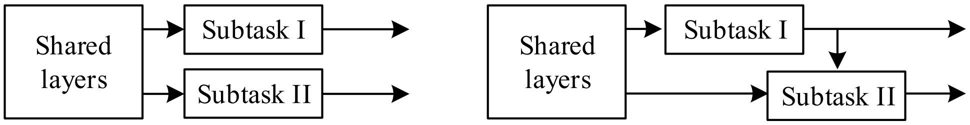


Figure 6.2: Left: Traditional multi-task learning [60]. Right: Proposed multi-task learning structure.

We describe a framework for end-to-end BIQA based on multi-task learning. Motivated by previous works [102, 60], we decompose the BIQA problem into two subtasks. Subtask I classifies an image into a specific distortion type from a set of pre-defined categories. Subtask II predicts the perceptual quality of the same image, taking advantage of distortion information obtained from Subtask I. On the one hand, two subtasks are related because quality degradation arises from distortion and the quality level is also affected by the distortion amount. On the other hand, they are different because images with different distortion types may exhibit similar quality while images with the same distortion may have drastically different quality, as shown in Fig. 6.1. The subtasks are accomplished by two sub-networks of linear convolutions and nonlinearities, and with shared features at early stages. Feature sharing not only greatly reduces the computation, but also enables the network to pre-train the shared layers via Subtask I, for which large-scale training data (distortion type) can be automatically generated at low cost. By doing so, we largely reduce the label noise problem. Unlike traditional multi-task learning, Subtask II of our method depends on the outputs of Subtask I, forming a casual structure as shown in Fig. 6.2. The structure makes the distortion information transparent to Subtask II for better quality prediction. We define a layer that is differential with respect to both convolutional activations and outputs of Subtask I to guarantee the feasibility of backpropagation. After pre-training, the entire network is end-to-end optimized using a variant of the stochastic gradient descent method. In addition, instead of using ReLU [105], we adopt a generalized divisive normalization (GDN) joint nonlinearity as the activation function that is inspired biologically, and has proven effective in assessing image quality [73], Gaussianizing image densities [2], and compressing digital images [3]. We empirically show that GDN has the capability to reduce model parameters/layers and meanwhile maintain similar quality prediction performance. We evaluate the resulting Multi-task End-to-end Optimized deep neural Network (MEON) based image quality index on four publicly available IQA databases and demonstrate that it achieves state-of-the-art performance compared with existing BIQA models. Finally, we investigate the generalizability and robustness of MEON using the gMAD competition methodology [90] on the Exploration database [86]. We observe that MEON significantly outperforms the most recent DNN-based BIQA model [5] and is highly competitive with MS-SSIM [165], a well-known FR-IQA model.

6.2 MEON for BIQA

Our work is motivated by two previous methods. In BIQI [102], Moorthy and Bovik proposed a two-step framework for BIQA, where an image is first classified into a particular distortion category, and then the distortion-specific quality prediction is performed [102]. The two steps of BIQI are optimized separately. Unlike BIQI, we are aiming at an end-to-end solution, meaning that feature representation, distortion type identification, and quality prediction are eventually optimized jointly. In [60], Kang *et al.* simultaneously estimated image quality and distortion type via a traditional multi-task DNN. However, simultaneous multi-task training requires ground truths of distortion type and subjective quality to be both available, which largely limits the total number of valid training samples. In addition, the quality prediction subtask is ignorant of the output from the distortion identification subtask. As a result, the performance is less competitive.

In the proposed MEON index, we take a raw image of $256 \times 256 \times 3$ as input and predict its perceptual quality score. How larger images are handled will be explained later. MEON consists of two subtasks accomplished by two sub-networks. Sub-network I aims to identify the distortion type in the form of a probability vector, which indicates the likelihood of each distortion and is fed as partial input to Sub-network II whose goal is to predict the image quality. Each subtask involves a loss function. Since Sub-network II relies on the output of Sub-network I, the two loss terms are not independent. We pre-train the shared layers in MEON via Subtask I and then jointly optimize the entire network with a unified loss function.

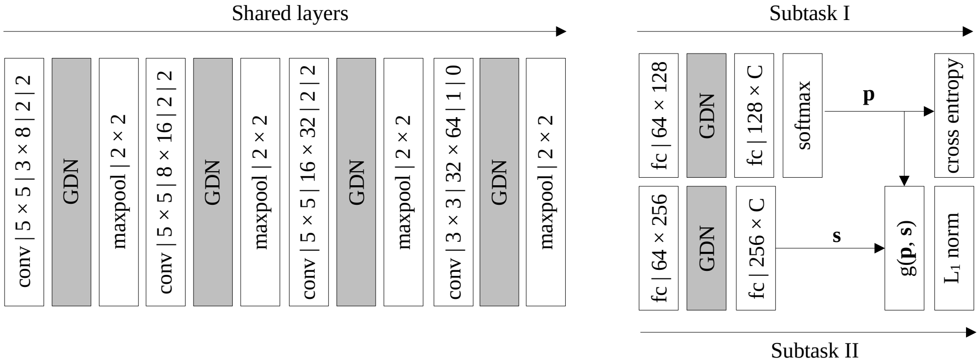


Figure 6.3: Illustration of MEON configurations for BIQA, highlighting the GDN nonlinearity. We follow the style and convention in [3] and denote the parameterization of the convolutional layer as "height × width | input channel × output channel | stride | padding".

6.2.1 GDN as Activation Function

Since Nair and Hinton revealed the importance of the ReLU nonlinearity in accelerating the training of DNNs [105], ReLU and its variants [41, 14] have become the dominant activation functions in DNNs. However, the joint statistics of linear filter responses after ReLU exhibit strong higher-order dependencies [2, 3]. As a result, ReLU generally requires a substantially large number of model parameters to achieve good performance for a particular task. These higher-order statistics may be significantly decorrelated through the use of a joint nonlinear gain control mechanism [129, 85] inspired by models of visual neurons [43, 10]. Previous studies also showed that incorporating the local gain control operation in DNNs improves the generalizability in image classification [68] and object recognition [53], where the parameters are predetermined empirically and fixed during training. Here, we adopt a GDN transform that has been previously demonstrated to work well in density estimation [2] and image compression [3]. Specifically, given a V-dimensional linear convolutional activation $\mathbf{x}(m,n) = (x_1(m,n), \cdots, x_V(m,n))$ at spatial location (m,n), the GDN transform is defined as

$$y_i(m,n) = \frac{x_i(m,n)}{\left(\beta_i + \sum_{j=1}^{V} \gamma_{ij} x_j(m,n)^2\right)^{\frac{1}{2}}}, \tag{6.1}$$

where $\mathbf{y}(m,n) = (y_1(m,n), \cdots, y_V(m,n))$ is the normalized activation vector at spatial location (m,n). The weight matrix $\boldsymbol{\gamma}$ and the bias vector $\boldsymbol{\beta}$ are parameters in GDN to be optimized. Both of them must be confined to $[0, +\infty)$ so as to ensure the legitimacy of the square root operation in the denominator and are shared across spatial locations. GDN is a differentiable transform and can be trained with any preceding or subsequent layers. Moreover, GDN is proven to be iteratively invertible under mild assumptions [2], which preserves better information than ReLU.

During training, we need to backpropagate the gradient of the loss ℓ through the GDN transform and compute the gradients with respect to its inputs and parameters. According to the chain rule

$$\frac{\partial \ell}{\partial x_j(m,n)} = \sum_{i=1}^{V} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial x_j(m,n)}, \tag{6.2}$$

$$\frac{\partial \ell}{\partial \beta_i} = \sum_{m=1}^{H} \sum_{n=1}^{W} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial \beta_i}, \tag{6.3}$$

$$\frac{\partial \ell}{\partial \gamma_{ij}} = \sum_{m=1}^{H} \sum_{n=1}^{W} \frac{\partial \ell}{\partial y_i(m,n)} \frac{\partial y_i(m,n)}{\partial \gamma_{ij}}, \tag{6.4}$$

where H and W denote the spatial sizes of the GDN transformed coefficients and

$$\frac{\partial y_i(m,n)}{\partial x_j(m,n)} = \begin{cases} \frac{\beta_i + \sum_{k \neq i} \gamma_{ik} x_k(m,n)^2}{\left(\beta_i + \sum_{k=1}^{V} \gamma_{ik} x_k(m,n)^2\right)^{\frac{3}{2}}} & i = j \\ \frac{-\gamma_{ij} x_i(m,n) x_j(m,n)}{\left(\beta_i + \sum_{k=1}^{V} \gamma_{ik} x_k(m,n)^2\right)^{\frac{3}{2}}} & i \neq j \end{cases}, \tag{6.5}$$

$$\frac{\partial y_i(m,n)}{\partial \beta_i} = \frac{-x_i(m,n)}{2\left(\beta_i + \sum_{j=1}^{V} \gamma_{ij} x_j(m,n)^2\right)^{\frac{3}{2}}}, \tag{6.6}$$

$$\frac{\partial y_i(m,n)}{\partial \gamma_{ij}} = \frac{-x_i(m,n) x_j(m,n)^2}{2\left(\beta_i + \sum_{j=1}^{V} \gamma_{ij} x_j(m,n)^2\right)^{\frac{3}{2}}}. \tag{6.7}$$

Some DNNs incorporate the batch normalization (BN) transform [51] that whitens the responses of linear filters to reduce the internal covariate shift and to rescale them in a reasonable operating range. GDN is different from BN in many ways. First, during testing, the mean and variance parameters are fixed and BN is simply an affine transform applied to the input. By contrast, GDN offers high nonlinearities especially when it is cascaded in multiple stages. Second, BN jointly normalizes all the activations across the mini-batch and over all spatial locations, which makes it an element-wise operation. Although the parameters in GDN are shared across the space similar to BN, the normalization of one activation at one location involves all activations across the channel, making it spatially adaptive.

6.2.2 Network Architecture

We denote our input mini-batch training data set by $\left\{\left(\mathbf{X}^{(k)}, \mathbf{p}^{(k)}, q^{(k)}\right)\right\}_{k=1}^{N}$, where $\mathbf{X}^{(k)}$ is the k-th raw input image, $\mathbf{p}^{(k)}$ is a multi-class indicator vector with only one entry activated to encode the ground truth distortion type, and $q^{(k)}$ is the MOS of the k-th input image. As depicted in Fig. 6.3, we first feed $\mathbf{X}^{(k)}$ to the shared layers, which are responsible for transforming raw image pixels into perceptually meaningful and distortion relevant feature representations. It consists of four stages of convolution, GDN, and maxpooling, whose model parameters are collectively denoted by $\mathbf{W}$. The parameterizations of convolution, maxpooling, and connectivity from layer to layer are detailed in Fig. 6.3. We reduce the spatial size by a factor of 4 after each stage via convolution with a stride of 2 or without padding, and 2×2 maxpooling. As a result, we represent a $256 \times 256 \times 3$ raw image by a 64-dimensional feature vector. On top of the shared layers, Sub-network I appends two fully connected layers with an intermediate GDN transform to increase nonlinearity, whose parameters are denoted by $\mathbf{w}_1$. We adopt the softmax function to encode the range to $[0, 1]$

$$\hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1) = \frac{\exp\left(y_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)\right)}{\sum_{j=1}^{C} \exp\left(y_j^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)\right)}, \tag{6.8}$$

where $\hat{\mathbf{p}}^{(k)} = (\hat{p}_1^{(k)}, \cdots, \hat{p}_C^{(k)})$ is a C-dimensional probability vector of the k-th input in a mini-batch, which indicates the probability of each distortion type. We take pristine images into account and use one entry to represent the "pristine" category. $\hat{\mathbf{p}}^{(k)}$ is the quantity fed to sub-network II and creates the causal structure. For Subtask I, we consider the batch-level cross entropy loss

$$\ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1) = -\sum_{k=1}^{N}\sum_{i=1}^{C} p_i^{(k)} \log \hat{p}_i^{(k)}(\mathbf{X}^{(k)}; \mathbf{W}, \mathbf{w}_1)\,. \tag{6.9}$$

Since we feed pristine images into Sub-network I by adding the "pristine" category, our training set is mildly unbalanced. Specifically, the number of images suffering from a particular distortion is K times as many as pristine images, where K is the number of distortion levels. It is straightforward to offset such class imbalance by adding weights in Eq. (6.9) according to the proportion of each distortion type. In our experiments, instead of over-weighting pristine images in the loss function, we over-sample them K times during training. By doing so, we expose our network to pristine images more often, which is beneficial for learning strong discriminative features to handle mild distortion cases.

Sub-network II takes the shared convolutional features and the estimated probability vector $\hat{\mathbf{p}}^{(k)}$ from Sub-network I as inputs. It predicts the perceptual quality of $\mathbf{X}^{(k)}$ in the form of a scalar value ranging between $[0, 100]$, where a lower score indicates worse perceptual quality. As in Sub-network I, to increase nonlinearity, we append two fully connected layers with an intermediate GDN layer, whose parameters are collectively denoted by $\mathbf{w}_2$. We double the node number of the first fully connected layer compared with that of Sub-network I, because predicting image quality is expected to be more difficult than identifying the distortion type. After the second fully connected layer, the network produces a score vector $\mathbf{s}^{(k)}$, whose i-th entry represents the perceptual quality score corresponding to the i-th distortion type. We define a layer that combines $\hat{\mathbf{p}}^{(k)}$ and $\mathbf{s}^{(k)}$ to yield an overall quality score

$$\hat{q}^{(k)} = g(\hat{\mathbf{p}}^{(k)}, \mathbf{s}^{(k)})\,. \tag{6.10}$$

We continue by completing the definition of $g(\cdot)$. First, in order to achieve theoretically valid backpropagation, g should be differentiable with respect to both $\hat{\mathbf{p}}^{(k)}$ and $\mathbf{s}^{(k)}$. Second, pairs $(\hat{p}_i^{(k)}, s_i^{(k)})$ and $(\hat{p}_j^{(k)}, s_j^{(k)})$ should be interchangeable in g to reflect the equal treatment of each distortion type under no privileged information. Third, g needs to be intuitively reasonable. For example, more emphasis should be given to $s_i^{(k)}$ if $\hat{p}_i^{(k)}$ is larger; $\hat{q}^{(k)}$ should be monotonically non-decreasing with respect to each entry of $\mathbf{s}^{(k)}$. Here, we adopt a probability-weighted summation [102] as a simple implementation of g

$$\hat{q}^{(k)} = g(\hat{\mathbf{p}}^{(k)}, \mathbf{s}^{(k)}) = \hat{\mathbf{p}}^{(k)T}\mathbf{s}^{(k)} = \sum_{i=1}^{C} \hat{p}_i^{(k)} \cdot s_i^{(k)}\,, \tag{6.11}$$

which is easily seen to obey all the properties listed above. For subtask II, we use the ℓ_1-norm as the batch-level loss function

$$\ell_2(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_2) = \|\mathbf{q} - \hat{\mathbf{q}}\|_1 = \sum_{k=1}^{K} |q^{(k)} - \hat{q}^{(k)}|\,. \tag{6.12}$$

We have also tried the ℓ_2-norm as the loss and observed similar performance. This is different from patch-based DNN methods [5] which show a clear preference to the ℓ_1-norm due to a high degree of label noise in the training data.

We now define the overall loss function of MEON as

$$\ell(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1, \mathbf{w}_2) = \ell_1 + \lambda\ell_2\,, \tag{6.13}$$

where λ is the balance weight to account for the scale difference between the two terms or to impose relative emphasis on one over the other.

We finish this subsection by highlighting the causal structure of MEON. In addition to the special treatment through Eq. (6.10) and Eq. (6.11), the gradient of ℓ with respect to $\hat{p}_i^{(k)}$ in Sub-network I

$$\frac{\partial \ell}{\partial \hat{p}_i^{(k)}} = \frac{\partial \ell_1}{\partial \hat{p}_i^{(k)}} + \lambda \frac{\partial \ell_2}{\partial \hat{p}_i^{(k)}} \tag{6.14}$$

$$= -\frac{p_i^{(k)}}{\hat{p}_i^{(k)}} - \lambda \mathrm{sign}\left(q^{(k)} - \hat{q}^{(k)}\right) s_i^{(k)}\,, \tag{6.15}$$

depends on the gradient backpropagated from Sub-network II.

6.2.3 Training and Testing

The success of DNN is largely owing to the availability of large-scale labeled training data. However, in BIQA, it is difficult to source accurate MOSs at a large scale because subject-rated images available in existing IQA databases are limited in size for training. Fortunately, our special design of MEON allows us to divide the training into two steps: pre-training and joint optimization. At the pre-training step, we minimize the loss function in Subtask I

$$(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1) = \operatorname{argmin} \ell_1(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1)\,. \tag{6.16}$$

The training set used for pre-training can be efficiently generated without subjective testing. Details will be discussed in Chapter 6.3.1. At the joint optimization step, we initialize $(\mathbf{W}, \mathbf{w}_1)$ with $(\hat{\mathbf{W}}, \hat{\mathbf{w}}_1)$ and minimize the overall loss function

$$(\mathbf{W}^\star, \mathbf{w}_1^\star, \mathbf{w}_2^\star) = \operatorname{argmin} \ell(\{\mathbf{X}^{(k)}\}; \mathbf{W}, \mathbf{w}_1, \mathbf{w}_2)\,. \tag{6.17}$$

During testing, given an image, we extract $256 \times 256 \times 3$ sub-images with a stride of J. The final distortion type is computed by majority vote among all predicted distortion types of the extracted sub-images. Similarly, the final quality score is obtained by simply averaging all predicted scores.

6.3 Experiments

In this section, we first describe the experimental setups including implementation details of MEON, IQA databases, and evaluation criteria. We then compare MEON with classic and state-of-the-art BIQA models. Finally, we conduct a series of ablation experiments to identify the contributions of the core factors in MEON.

6.3.1 Experimental Setup

Implementation Details: Both pre-training and joint optimization steps adopt the Adam optimization algorithm [63] with a mini-batch of 40. For pre-training, we start with the learning rate $\alpha = 10^{-2}$ and subsequently lower it by a factor of 10 when the loss plateaus, until $\alpha = 10^{-4}$. For joint optimization, α is fixed to 10^{-4}. Other parameters in Adam are set by default [63]. The learning rates for biases are doubled. The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in GDN are projected to nonnegative values after each update. Additionally, we enforce $\boldsymbol{\gamma}$ to be symmetric by averaging it with its transpose as recommended in [3]. The balance weight in Eq. (6.13) is set to account for the scale difference between the two terms (0.2 for LIVE [135] and 1 for TID2013 [115]). During testing, the stride J is set to 128. We augment the training data by randomly horizontal flipping and changing their contrast and saturation within the range that is indiscernible to human eyes. Since quality changes with scales which correspond to different viewing distances, we do not augment training data across scales.

Similar in Chapter 5, we select 840 high-resolution natural images with nearly pristine quality as the basis to construct the dataset for pre-training. Some representative images are shown in Fig. 6.4. We down-sample each image to further reduce possible compression artifacts, keeping a maximum height or width of 768. All $C-1$ distortion types (excluding the "pristine" category) are applied to those images, each with 5 distortion levels. As previously described, we over-sample pristine images to balance the class labels during pre-training. As a result, our dataset contains a total of $C \times 840 \times 5$ images with ground truth labels automatically generated.

IQA Databases: We compare MEON with classic and state-of-the-art BIQA models on four standard IQA databases. They are LIVE [135], CSIQ [69], TID2013 [115], and the Exploration database [86]. In the first set of experiments, we consider four distortion types that are common in the four databases: JP2K, JPEG, WN, and BLUR. This leaves us 634, 600, 500, and 94880 test images in LIVE [135], CSIQ [69], TID2013 [115], and the Exploration database, respectively. In the second set of experiments, we investigate the effectiveness of MEON on handling more distortion types (24 to be specific) by considering all $3,000$ test images in TID2013 [115].

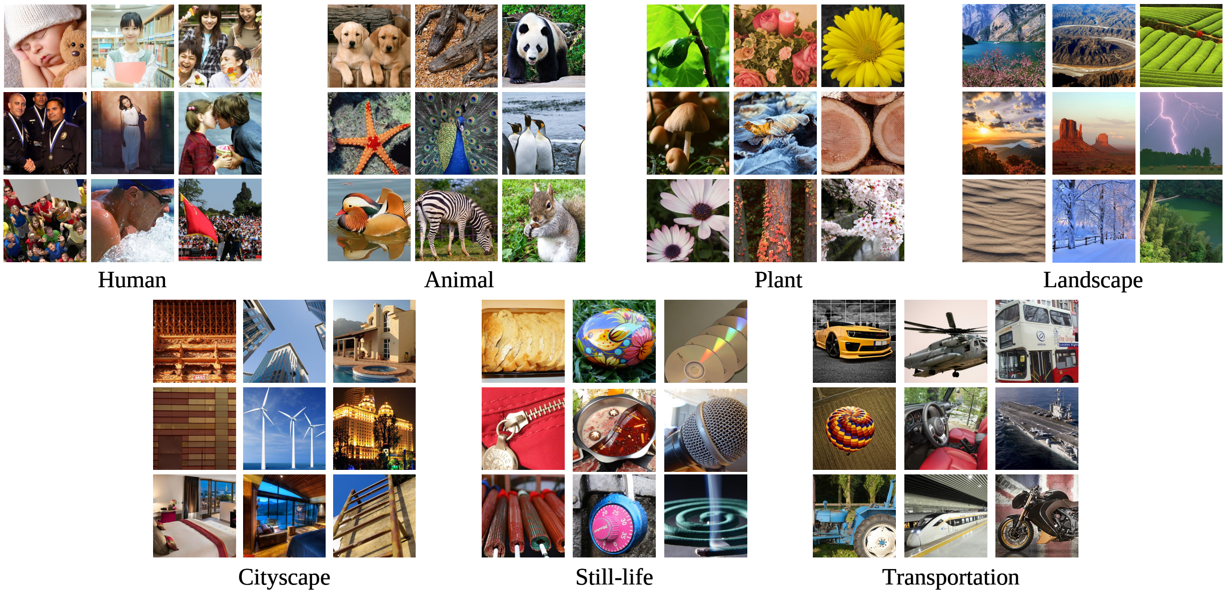


Figure 6.4: Sample source images used for pre-training. All images are cropped for better visibility.

Evaluation Criteria: Five evaluation criteria are adopted as follows and their detailed descriptions are given in Chapters 3 and 5.

- SRCC (Eq. (5.7)).
- PLCC (Eq. (5.8)).
- D-test (Chapter 3.2.1).
- L-test (Chapter 3.2.2).
- P-test (Chapter 3.2.3).

We apply SRCC and PLCC to LIVE [135], CSIQ [69], and TID2013 [115]. The other three tests are used in the Exploration database [86].

6.3.2 Experimental Results

Results on Four Distortions: We compare MEON with classic and state-of-the-art BIQA models on four common distortion types in LIVE [135], CSIQ [69], TID2013 [115], and the Exploration database [86]. The competing algorithms are chosen to cover a diversity of design philosophies, including three classic ones: DIIVINE [103], BRISQUE [99] and CORNIA [184], and five state-of-the-art ones: ILNIQE [187], BLISS [183], HOSA [174], dipIQ [88] and deepIQA [5]. In order to make a fair comparison, all models are re-trained/validated on the full LIVE database and tested on CSIQ, TID2013, and the Exploration database. As for MEON, we randomly select 23 reference and their corresponding distorted images in LIVE for training and leave the rest 6 reference and their distorted images for validation. The model parameters with the lowest validation loss are chosen. When testing, we follow the common practice of Mittal *et al.* [99] and Ye *et al.* [183] and randomly choose 80% reference images along with their corresponding distorted images to estimate the parameters $\{\beta_i | i = 1, 2, 3, 4\}$ of a nonlinear function

$$\tilde{q} = \frac{\beta_1 - \beta_2}{1 + \exp\left(\frac{-\hat{q} + \beta_3)}{|\beta_4|}\right)} + \beta_2 \,, \tag{6.18}$$

which is used to map model predictions to the MOS scale. The rest 20% images are left out for testing. This procedure is repeated $1,000$ times and the median SRCC and PLCC values are reported.

Tables 6.1, 6.2, and 6.3 show the results on CSIQ [69], TID2013 [115], and the Exploration database [86], respectively, from which the key observations are as follows. First, MEON achieves state-of-the-art performance on all three databases. Although there is slight performance bias towards JPEG and WN, MEON aligns all distortions pretty well across the perceptual space. Second, MEON significantly outperforms DIIVINE [103], an improved version of BIQI [102] with more advanced NSS. The performance improvement is largely due to the joint end-to-end optimization for feature and multi-task learning. Third, MEON performs the best in D-test on the Exploration database, which is no surprise because we are optimizing a more fine-grained version of D-test through Subtask I. More

Table 6.1: Median SRCC and PLCC results across $1,000$ sessions on CSIQ [69]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [103]	0.844	0.819	0.881	0.884	0.835
BRISQUE [99]	0.894	0.916	**0.934**	0.915	0.909
CORNIA [184]	0.916	0.919	0.787	**0.928**	0.914
ILNIQE [187]	0.924	0.905	0.867	0.867	0.887
BLISS [183]	**0.932**	0.927	0.879	0.922	0.920
HOSA [174]	0.920	0.918	0.895	0.915	0.918
dipIQ [88]	**0.944**	**0.936**	0.904	**0.932**	**0.930**
deepIQA [5]	0.907	0.929	0.933	0.890	0.871
MEON	0.898	**0.948**	**0.951**	0.918	**0.932**
PLCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [103]	0.898	0.818	0.903	0.909	0.855
BRISQUE [99]	0.937	0.960	**0.947**	0.936	0.937
CORNIA [184]	0.947	0.960	0.777	**0.953**	0.934
ILNIQE [187]	0.942	0.956	0.880	0.903	0.914
BLISS [183]	**0.954**	0.970	0.895	0.947	0.939
HOSA [174]	0.946	0.958	0.912	0.940	0.942
dipIQ [88]	**0.959**	**0.975**	0.927	**0.958**	**0.949**
deepIQA [5]	0.931	0.951	0.933	0.906	0.891
MEON	0.925	**0.979**	**0.958**	0.946	**0.944**

specifically, the network learns not only to classify the image into pristine and distorted classes but also to identify the specific distortion type whenever distorted. Fourth, we observe stronger generalizability of MEON on the Exploration database compared with another DNN-based method, deepIQA [5]. We believe the performance improvement arises because 1) the proposed novel learning framework has the quality prediction subtask regularized by the distortion identification subtask; 2) images instead of patches are used as inputs to reduce the label noise; 3) the pre-training step enables the network to start from a more task-relevant initialization, resulting in a better local optimum.

As a by-product, MEON outputs the distortion information of a test image, whose accuracy on CSIQ [69], TID2013 [115], and the Exploration database [86] is shown in Table 6.4. Empirical justifications for the correlation of the two subtasks can be easily seen,

Table 6.2: Median SRCC and PLCC results across $1,000$ sessions on TID2013 [115]

SRCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [103]	0.857	0.680	0.879	0.859	0.795
BRISQUE [99]	0.906	0.894	0.889	0.886	0.883
CORNIA [184]	0.907	0.912	0.798	**0.934**	0.893
ILNIQE [187]	0.912	0.873	0.890	0.815	0.881
BLISS [183]	0.906	0.893	0.856	0.872	0.836
HOSA [174]	**0.933**	0.917	0.843	0.921	**0.904**
dipIQ [88]	0.926	**0.932**	0.905	**0.922**	0.877
deepIQA [5]	**0.948**	**0.921**	**0.938**	0.910	0.885
MEON	0.911	0.919	**0.908**	0.891	**0.912**

PLCC	JP2K	JPEG	WN	BLUR	ALL4
DIIVINE [103]	0.901	0.696	0.882	0.860	0.794
BRISQUE [99]	0.919	0.950	0.886	0.884	0.900
CORNIA [184]	0.928	0.960	0.778	**0.934**	0.904
ILNIQE [187]	0.929	0.944	0.899	0.816	0.890
BLISS [183]	0.930	0.963	0.863	0.872	0.862
HOSA [174]	**0.952**	0.949	0.842	0.921	**0.918**
dipIQ [88]	0.948	**0.973**	0.906	**0.928**	0.894
deepIQA [5]	**0.963**	0.960	**0.943**	0.897	**0.913**
MEON	0.924	**0.969**	**0.911**	0.899	0.912

where a lower classification error of a particular distortion generally leads to better quality prediction performance on that distortion and vice versa (*e.g.*, WN and BLUR). Since the statistical behaviors of WN have obvious distinctions with the other three distortions, MEON predicts WN nearly perfectly. On the other hand, it confounds JP2K with BLUR sometimes because JP2K often introduces significant blur at low bitrates. When the distortion level is mild, MEON occasionally labels distorted images as pristine, which is not surprising because the HVS is also easily fooled by such cases. Finally, there is still much room for improvement to correctly classify pristine images. We conjecture that adding more training data in the pre-training step may help improve the results.

Moreover, we let MEON play the gMAD competition game [90] with deepIQA [5]. We choose the Exploration database [86] as the playground. An image pair is automatically

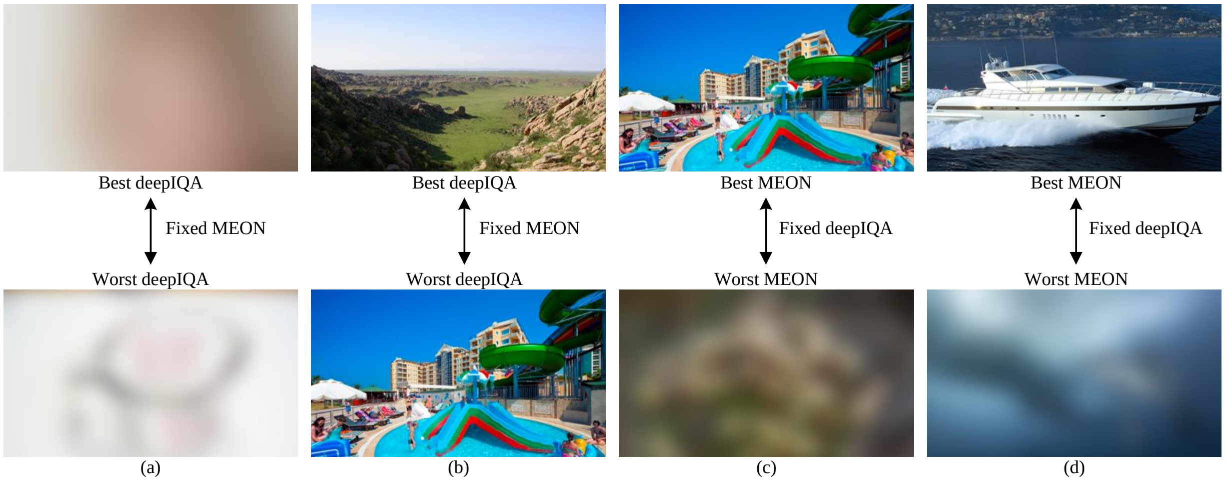


Figure 6.5: The gMAD competition results between MEON and deepIQA [5]. (a) Fixed MEON at the low-quality level. (b) Fixed MEON at the high-quality level. (c) Fixed deepIQA at the low-quality level. (d) Fixed deepIQA at the high-quality level.

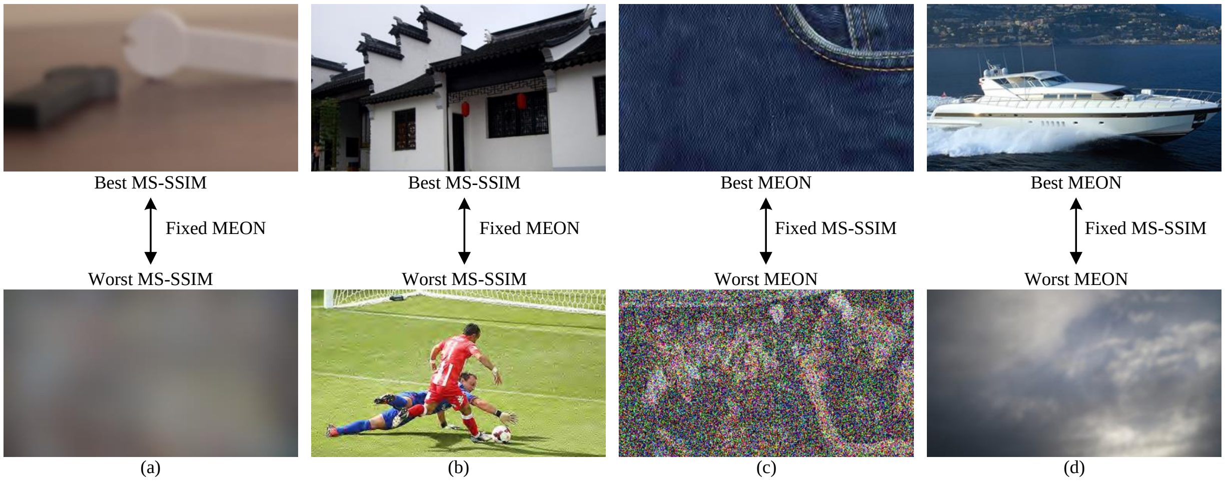


Figure 6.6: The gMAD competition results between MEON and MS-SSIM [165]. (a) Fixed MEON at the low-quality level. (b) Fixed MEON at the high-quality level. (c) Fixed MS-SSIM at the low-quality level. (d) Fixed MS-SSIM at the high-quality level.

Table 6.3: The D-test, L-test, and P-test results on the Exploration database

	D-test	L-test	P-test
DIIVINE [103]	0.8538	0.8908	0.9540
BRISQUE [99]	0.9204	0.9772	0.9930
CORNIA [184]	0.9290	0.9764	0.9947
ILNIQE [187]	0.9084	**0.9926**	0.9927
BLISS [183]	0.9080	0.9801	**0.9996**
HOSA [174]	0.9175	0.9647	0.9983
dipIQ [88]	**0.9346**	**0.9846**	**0.9999**
deepIQA [5]	0.9074	0.9467	0.9628
MEON	**0.9384**	0.9669	0.9984

searched for the maximum quality difference in terms of MEON, while keeping deepIQA [5] predictions at the same quality level. The procedure is then repeated with the roles of the two models exchanged. Four such image pairs are shown in Fig. 6.5 (a)-(d), where MEON considers pairs (a) and (b) of the same quality at low- and high-quality levels respectively, which is in close agreement with our visual observations. By contrast, deepIQA incorrectly predicts the top images of (a) and (b) to have much better quality than that of the bottom images. Similar conclusions can be drawn by examining pairs (c) and (d), where the roles of the two models are reversed. The results of gMAD provide strong evidence that the generalizability of MEON is significantly improved over deepIQA [5]. We further compare MEON through gMAD with MS-SSIM [165]. Fig. 6.6 (a)-(d) show the results, from which we observe that MEON is highly competitive with MS-SSIM [165] in the sense that both methods are able to fail each other by successfully finding strong counterexamples. Specifically, MS-SSIM [165] tends to over-penalize WN but under-penalize BLUR. MEON is able to reveal such weaknesses of MS-SSIM, which can be easily discerned in the bottom images of Fig. 6.6 (c) and (d). On the other hand, MS-SSIM takes advantage of the fact that MEON does not handle BLUR and JP2K well enough and finds counterexamples from those distortions.

Results on More Distortion Types: We investigate the scalability of our multi-task learning framework to handle more distortion types by training and testing on the full TID2013 database [115]. For pre-training, we make our best effort to reproduce 15 out of

Table 6.4: The confusion matrices produced by MEON on CSIQ [69], TID2013 [115], the Exploration database. The column and the raw contain ground truth and predicted distortion types, respectively

Accuracy		JP2K	JPEG	WN	BLUR	Pristine
CSIQ	JP2K	**0.847**	0.007	0.000	0.093	0.053
	JPEG	0.040	**0.820**	0.000	0.027	0.113
	WN	0.000	0.000	**0.947**	0.013	0.040
	BLUR	0.067	0.006	0.000	**0.827**	0.100
	Pristine	0.067	0.000	0.100	0.166	**0.667**
TID2013	JP2K	**0.944**	0.016	0.000	0.040	0.000
	JPEG	0.032	**0.968**	0.000	0.000	0.000
	WN	0.000	0.000	**1.000**	0.000	0.000
	BLUR	0.088	0.008	0.000	**0.848**	0.056
	Pristine	0.160	0.000	0.040	0.000	**0.800**
Exploration	JP2K	**0.985**	0.000	0.000	0.015	0.000
	JPEG	0.006	**0.994**	0.000	0.000	0.000
	WN	0.000	0.000	**1.000**	0.000	0.000
	BLUR	0.003	0.000	0.000	**0.997**	0.000
	Pristine	0.213	0.050	0.067	0.234	**0.436**

Table 6.5: Median SRCC results across 10 sessions on the full TID2013 database with 24 distortion types. The distortion index is the same as in the original paper [115]

SRCC	#01	#02	#03	#04	#05	#06	#07	#08	#09
DIIVINE [103]	0.756	0.464	0.869	0.374	0.794	0.704	0.650	**0.900**	0.814
BRISQUE [99]	0.674	0.550	0.804	0.222	0.824	0.749	0.677	0.855	0.492
CORNIA [184]	0.496	0.130	0.655	0.373	0.715	0.647	0.632	0.844	0.688
ILNIQE [187]	**0.924**	**0.847**	**0.947**	**0.786**	**0.908**	**0.847**	**0.933**	0.869	**0.846**
HOSA [174]	**0.833**	0.575	0.808	0.432	0.906	0.817	0.783	**0.903**	**0.873**
deepIQA [5]	—	—	—	—	—	—	—	—	—
MEON	0.813	**0.722**	**0.926**	**0.728**	**0.911**	**0.901**	**0.888**	0.887	0.797

SRCC	#10	#11	#12	#13	#14	#15	#16	#17	#18
DIIVINE [103]	0.795	0.804	0.514	**0.892**	**0.215**	**0.389**	0.124	0.189	0.280
BRISQUE [99]	0.751	0.696	0.285	0.719	0.158	0.362	0.253	0.102	0.200
CORNIA [184]	0.758	0.866	0.587	0.603	**0.282**	-0.025	0.194	0.145	-0.006
ILNIQE [187]	**0.901**	**0.930**	0.400	0.708	-0.173	0.000	**0.328**	0.080	0.103
HOSA [174]	**0.903**	**0.920**	**0.712**	**0.743**	0.143	0.330	**0.279**	**0.307**	**0.414**
deepIQA [5]	—	—	—	—	—	—	—	—	—
MEON	0.850	0.891	**0.746**	0.716	0.116	**0.500**	0.177	**0.252**	**0.684**

SRCC	#19	#20	#21	#22	#23	#24	ALL
DIIVINE [103]	0.691	0.340	0.690	0.769	0.700	0.795	0.632
BRISQUE [99]	0.587	0.211	0.546	**0.842**	0.770	0.764	0.572
CORNIA [184]	0.461	**0.560**	0.648	0.646	0.672	0.867	0.611
ILNIQE [187]	**0.773**	0.507	**0.911**	0.822	**0.801**	**0.878**	0.534
HOSA [174]	0.711	**0.537**	0.756	0.840	**0.821**	**0.903**	0.707
deepIQA [5]	—	—	—	—	—	—	**0.761**
MEON	**0.849**	0.406	**0.772**	**0.857**	0.779	0.855	**0.808**

the 24 distortions in TID2013 and apply them to the 840 high-quality images. As a result, only parameters of the shared layers **W** are provided with meaningful initializations. Since BLISS [183] and dipIQ [88] cannot be trained without all distorted images originated from the 840 high-quality ones, we exclude them from the comparison. For joint optimization, we follow Bosse *et al.* [5] and use 15, 5, and 5 reference and their corresponding distorted images for training, validation, and testing, respectively. Median SRCC results are reported based on 10 random splits in Table 6.5. All other competing BIQA models except deepIQA [5] are re-trained, validated, and tested in exactly the same way. Since the training codes of deepIQA are not available, we copy the results from the original paper for reference (note that the random seeds for the 10 data splits may be different).

From Table 6.5, we observe that MEON outperforms previous BIQA models by a clear margin, aligning 24 distortions in the perceptual space remarkably well. By contrast, although ILNIQE [187] does an excellent job in predicting image quality under the same distortion type, which is also reflected in its superior performance in L-test on the Exploration database, it fails to align distortion types correctly. Moreover, all competing BIQA models including MEON do not perform well on mean shift (#16) and contrast change (#17) cases. This is not surprising for methods that adopt spatial normalization as preprocessing, such as BRISQUE [99], CORNIA [184], ILNIQE [187], and HOSA [174] because the mean and contrast information has been removed at the very beginning. Moreover, mean shift and contrast change may not be considered as distortions at all because modest mean shift may not affect perceptual quality and contrast change (*e.g.*, contrast enhancement) often improves image quality.

Ablation Experiments: We conduct a series of ablation experiments to single out the core contributors of MEON. We first train Sub-network II with random initializations as a simple single-task baseline. We also experiment with the traditional multi-task learning framework by directly producing an overall quality score. From Table 6.6, we observe that without pre-training, MEON achieves the best performance. Moreover, pre-training brings the prediction accuracy to the next stage. We conclude that the proposed multi-task learning framework and the pre-training mechanism are keys to the success of MEON.

Next, we analyze the impact of the GDN transform on model complexity and quality

Table 6.6: Median SRCC results of ablation experiments across $1,000$ sessions on CSIQ [69] and TID2013 [115]

	CSIQ	TID2013
Single-task	0.844	0.850
Traditional multi-task	0.885	0.871
MEON w/o pre-training	0.894	0.880
MEON with pre-training	**0.932**	**0.912**

Table 6.7: SRCC performance comparison of configurations with different activation functions and model complexities

	CSIQ	TID2013
ReLU + single layer	0.922	0.891
ReLU + double layers	0.924	0.900
ReLU + double layers + BN	0.930	**0.918**
MEON (GDN + single layer)	**0.932**	0.912

prediction performance. We start from a baseline by replacing all GDN layers with ReLU. We then double all convolutional and fully connected layers in both Sub-networks I and II with ReLU nonlinearity to see whether a deeper network improves the performance. Last, we introduce the BN transform right before each ReLU layer. The results are listed in Table 6.7, from which we see that simply replacing GDN with ReLU leads to inferior performance. The network with a deeper architecture slightly improves the performance. When combined with BN, it achieves competitive performance compared with the proposed method. This suggests that GDN may be an effective way to reduce model complexity without sacrificing the performance. Specifically in our experiments, GDN is able to half the layers and parameters of the network while achieving similar performance using ReLU.

6.4 Summary

We propose a novel multi-task learning framework for BIQA, namely MEON, by decomposing the BIQA task into two subtasks with dependent loss functions. We end-to-end optimize MEON for both distortion identification and quality prediction. The resulting MEON index demonstrates state-of-the-art performance, which we believe arises from pre-training for better initializations, multi-task learning for mutual regularization, and GDN for biologically inspired feature representations. In addition, we show the scalability of MEON to handle more distortion types and its strong competitiveness against state-of-the-art BIQA approaches in gMAD competitions.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we attempt to investigate new evaluation and design methodologies for BIQA. We first aim at addressing the IQA model comparison problem to overcome the conflict between the enormous size of image space and the limited resource for subjective testing. To this end, we build the Waterloo Exploration Database and introduce three test criteria (D-test, L-test, and P-test) that are independent of subjective testing. Moreover, we propose a general methodology, namely the gMAD competition, to compare multiple computational models for perceptually discriminable quantities and apply it to IQA model comparison.

The second part of this thesis focuses on BIQA model learning. We first learn robust BIQA models using mature L2R algorithms from millions of DIPs, which can be automatically generated at very low cost. We then exploit the fact that the distortion type information of images is readily available and propose a multi-task DNN for BIQA by decomposing it into two subtasks. Highly competitive performance against the state-of-the-art is achieved by the novel approaches proposed in this thesis, including the dipIQ, dilIQ, and MEON algorithms.

7.2 Future Work

The current work can be extended in many ways, some of which are listed as follows.

Waterloo Exploration Database Release II: Although the current Waterloo Exploration Database is the largest in the IQA field, it is still small relative to the image space for IQA predictions. Therefore, it is necessary to be extended to a larger one, based on which D-test, L-test, and P-test are more powerful to distinguish between BIQA models. The construction of the Waterloo Exploration Database does not involve subjective testing; the only human intervention is to screen high-quality images from the Internet. Therefore, it is readily extended by adding more pristine images, more distortion types and/or more distortion levels.

Extending the gMAD Competition: Although we have applied gMAD to three different perceptual quantities—image quality, image aesthetics, and video QoE—there are a much wider variety of scenarios that gMAD can come into play. To give a few examples, these include comparisons of image/video emotion predictors in the field of cognitive vision [58], the relative attributes (sportiness and furriness) estimators in the field of semantic image search [66], machine translation quality estimators in the field of computational linguistics [36], and thermal comfort models in the field of thermal environment of buildings [107].

The current gMAD requires computational models to produce continuous-valued responses. How to adapt gMAD to account for discrete-valued models has great potentials to impact other computer vision and machine learning applications. For example, instead of building a larger database than ImageNet [123], it is of great interest to see how the current image classification algorithms behave in a discrete version of gMAD setting with a low and manageable subjective testing cost. On the other hand, the current gMAD requires computational models to be scalar-valued, manifesting themselves in predicting a perceptual quantity. It is interesting to extend gMAD to include vector-valued models. A direct application is to compare the robustness of different feature representations in a computational vision task.

Extending L2R Approaches for BIQA: The current L2R approaches for BIQA open the door to a new class of OU-BIQA models and many exciting directions are worth exploring. First, novel image pair and list generation engines may be developed to account for situations that reference images are not available (or do not ever exist). Second, in practice, a pair of images may be regarded as having indiscriminable quality. Such knowledge could be obtained either from subjective testing (*e.g.*, paired comparison between images) or from the image source (*e.g.*, two pristine images acquired from the same source), and is informative in constraining the behavior of an objective quality model. The current learning framework needs to be improved in order to learn from such quality-indiscriminable image pairs. Third, given the powerful DIP generation engine developed in the current work and the remarkable success of recent DNNs, it may become feasible to develop end-to-end BIQA models using the proposed L2R schemes, aiming for even stronger robustness and generalizability.

Distortion-Unaware BIQA: Another design philosophy of BIQA that is worth exploring is the statistical modeling of natural undistorted images. Quality predictions of a distorted image with an unknown distortion type can be performed by quantifying its departure from the statistical regularities. By doing so, we make BIQA models distortion-unaware and can apply them to evaluation distorted images of any possible distortion type, as opposed to existing BIQA models that need to be trained on specific distortions and cannot be generalized to unseen distortion types.

Blind Video Quality Assessment (BVQA): Although considerable progress has been made in BIQA, the progress on developing BVQA models has been relatively slow due to the complexities of video spatial/temporal features and perceptual spatiotemporal characteristics. It would be interesting to extend the current BIQA frameworks to BVQA for real time video monitoring, resource allocation, and rate distortion optimization.

References

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

[2] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. Density modeling of images using a generalized normalization transformation. In *International Conference on Learning Representations*, 2016.

[3] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

[4] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *CoRR*, abs/1602.05531, 2016.

[5] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *CoRR*, abs/1612.01697, 2016.

[6] Alan C. Bovik. Meditations on video quality. *IEEE Multimedia Communications E-Letter*, 4(4):4–10, May 2009.

[7] Alan C. Bovik. *Handbook of Image and Video Processing*. Academic Press, 2010.

[8] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, pages 89–96, 2005.

[9] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. In *International Conference on Machine Learning*, pages 129–136, 2007.

[10] Matteo Carandini and David J. Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, Jan. 2012.

[11] Damon M. Chandler and Sheila S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, 16(9):2284–2298, Sep. 2007.

[12] Cisco IBSG Youth Focus Group. Cisco IBSG youth survey, Nov. 2010. [Online]. Available: http://www.cisco.com/c/dam/en_us/about/ac79/docs/ppt/Video_Disruption_SP_Strategies_IBSG.pdf.

[13] Willard H. Clatworthy. Partially balanced incomplete block designs with two associate classes and two treatments per block. *Journal of Research of the National Bureau of Standards*, 54(4):177–190, Apr. 1955.

[14] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *CoRR*, abs/1511.07289, 2015.

[15] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, Sep. 1995.

[16] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[17] CVX Research Inc. CVX: Matlab software for disciplined convex programming, version 2.0, Aug. 2012. [Online]. Available: http://cvxr.com/cvx.

[18] Scott J. Daly. Visible differences predictor: An algorithm for the assessment of image fidelity. In *SPIE/IS&T Symposium on Electronic Imaging: Science and Technology*, pages 2–15, 1992.

[19] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, pages 288–301, 2006.

[20] Herbert A. David. *The Method of Paired Comparisons*. DTIC Document, 1963.

[21] Zhengfang Duanmu, Kede Ma, and Zhou Wang. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *ACM Multimedia*, pages 1–9, 2017.

[22] Zhengfang Duanmu, Abdul Rehman, Kai Zeng, and Zhou Wang. Quality-of-experience prediction for streaming video. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2016.

[23] Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. A quality-of-experience index for streaming video. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):154–166, Feb. 2017.

[24] Alexander Eichhorn, Pengpeng Ni, and Ragnhild Eg. Randomised pair comparison: An economic and robust method for audiovisual quality assessment. In *ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 63–68, 2010.

[25] Ulrich Engelke, Maulana Kusuma, Hans-Jürgen Zepernick, and Manora Caldera. Reduced-reference metric design for objective perceptual quality assessment in wireless imaging. *Signal Processing: Image Communication*, 24(7):525–547, Aug. 2009.

[26] Yuming Fang, Kede Ma, Zhou Wang, Weisi Lin, Zhijun Fang, and Guangtao Zhai. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, 22(7):838–842, Jul. 2015.

[27] Olivier D. Faugeras and William K. Pratt. Decorrelation methods of texture feature extraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(4):323–332, Jul. 1980.

[28] David J. Field. What is the goal of sensory coding? *Neural Computation*, 6(4):559–601, Jul. 1994.

[29] Peter Fröhlich, Sebastian Egger, Raimund Schatz, Michael Mühlegger, Kathrin Masuch, and Bruno Gardlo. QoE in 10 seconds: Are short video clip lengths sufficient for quality of experience assessment? In *IEEE International Workshop on Quality of Multimedia Experience*, pages 242–247, 2012.

[30] Fei Gao, Dacheng Tao, Xinbo Gao, and Xuelong Li. Learning to rank for blind image quality assessment. *IEEE Transactions on Neural Networks and Learning Systems*, 26(10):2275–2290, Oct. 2015.

[31] Xinbo Gao, Fei Gao, Dacheng Tao, and Xuelong Li. Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning. *IEEE Transactions on Neural Networks and Learning Systems*, 24(12):2013–2026, Dec. 2013.

[32] Xinbo Gao, Wen Lu, Dacheng Tao, and Xuelong Li. Image quality assessment based on multiscale geometric analysis. *IEEE Transactions on Image Processing*, 18(7):1409–1423, Jul. 2009.

[33] Wilson S. Geisler and Randy L. Diehl. Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1420):419–448, Apr. 2002.

[34] Deepti Ghadiyaram and Alan C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, Jan. 2016.

[35] Bernd Girod. What's wrong with mean-squared error? In *Digital images and human vision*, pages 207–220, 1993.

[36] Yvette Graham. Improving evaluation of machine translation quality estimation. In *53rd Annual Meeting of the Association for Computational Linguistics*, pages 1804–1813, 2015.

[37] Michael Grubinger, Stefanie Nowak, and Paul Clough. *Data Sets Created in ImageCLEF*, pages 19–43. Springer Berlin Heidelberg, 2010.

[38] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Using free energy principle for blind image quality assessment. *IEEE Transactions on Multimedia*, 17(1):50–63, Jan. 2015.

[39] Li Hang. A short introduction to learning to rank. *IEICE Transactions on Information and Systems*, 94(10):1854–1862, Oct. 2011.

[40] Rania Hassen, Zhou Wang, and Magdy M. Salama. Image sharpness assessment based on local phase coherence. *IEEE Transactions on Image Processing*, 22(7):2798–2810, Jul. 2013.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[43] David J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(02):181–197, Aug. 1992.

[44] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, Jul. 2006.

[45] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2):177–196, Jan. 2001.

[46] Yuukou Horita, Keiji Shibata, Yoshikazu Kawayoke, and ZM Parvez Sazzad. Toyama-MICT image quality evaluation database, 2010. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.

[47] Tobias Hoßfeld, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. Quantification of YouTube QoE via crowdsourcing. In *IEEE International Symposium on Multimedia*, pages 494–499, 2011.

[48] Weilong Hou, Xinbo Gao, Dacheng Tao, and Xuelong Li. Blind image quality assessment via deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 26(6):1275–1286, Jun. 2015.

[49] Thomas S. Huang, James W. Burnett, and Andrew G. Deczky. The importance of phase in image processing filters. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23(6):529–542, Dec. 1975.

[50] David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, Jan. 1962.

[51] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[52] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM computing Surveys*, 31(3):264–323, Sep. 1999.

[53] Kevin Jarrett, Koray Kavukcuoglu, MarcAurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *IEEE International Conference on Computer Vision*, pages 2146–2153, 2009.

[54] Dinesh Jayaraman, Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Objective quality assessment of multiply distorted images. In *the Forty-Sixth IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1693–1697, 2012.

[55] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, Mar. 2011.

[56] Bin Jin, Maria V. O. Segovia, and Sabine Süsstrunk. Image aesthetic predictors based on weighted CNNs. In *IEEE International Conference on Image Processing*, pages 2291–2295, 2016.

[57] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[58] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5):94–115, Sep. 2011.

[59] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[60] Le Kang, Peng Ye, Yi Li, and David Doermann. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In *IEEE International Conference on Image Processing*, pages 2791–2795, 2015.

[61] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 419–426, 2006.

[62] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, Feb. 2017.

[63] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[64] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*, pages 662–679, 2016.

[65] Leonid L. Kontsevich and Christopher W. Tyler. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729–2737, Aug. 1999.

[66] Adriana Kovashka, Devi Parikh, and Kristen Grauman. WhittleSearch: Image search with relative attribute feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980, 2012.

[67] Peter Kovesi. Image features from phase congruency. *Journal of Computer Vision Research*, 1(3):1–26, Sum. 1999.

[68] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[69] Eric C. Larson and Damon M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *SPIE Journal of Electronic Imaging*, 19(1):1–21, Jan. 2010.

[70] Patrick Le Callet and Florent Autrusseau. Subjective quality assessment IRCCyN/IVC database, 2005. [Online]. Available: http://www.irccyn.ec-nantes.fr/ivcdb/.

[71] Chaofeng Li, Alan C. Bovik, and Xiaojun Wu. Blind image quality assessment using a general regression neural network. *IEEE Transactions on Neural Networks*, 22(5):793–799, May 2011.

[72] Jing Li, Marcus Barkowsky, and Patrick Le Callet. Boosting paired comparison methodology in measuring visual discomfort of 3DTV: Performances of three different designs. In *IS&T/SPIE Electronic Imaging*, pages 1–12, 2013.

[73] Qiang Li and Zhou Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):202–211, Apr. 2009.

[74] Xin Li. Blind image quality assessment. In *IEEE International Conference on Image Processing*, pages 449–452, 2002.

[75] Weisi Lin and C.-C. Jay Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, May 2011.

[76] Anmin Liu, Weisi Lin, and Manish Narwaria. Image quality assessment based on gradient similarity. *IEEE Transactions on Image Processing*, 21(4):1500–1512, Apr. 2012.

[77] Hantao Liu, Nick Klomp, and Ingrid Heynderickx. A no-reference metric for perceived ringing artifacts in images. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(4):529–539, Apr. 2010.

[78] Shizhong Liu and Alan C. Bovik. Efficient DCT-domain blind measurement and reduction of blocking artifacts. *IEEE Transactions on Circuits and Systems for Video Technology*, 12(12):1139–1149, Dec. 2002.

[79] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, Mar. 2009.

[80] Wentao Liu and Zhou Wang. A database for perceptual evaluation of image aesthetics. In *IEEE International Conference on Image Processing*, pages xx–xx, 2017.

[81] Xi Liu, Florin Dobrian, Henry Milner, Junchen Jiang, Vyas Sekar, Ion Stoica, and Hui Zhang. A case for a coordinated Internet video control plane. In *ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 359–370, 2012.

[82] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *IEEE International Conference on Computer Vision*, pages 990–998, 2015.

[83] R. Duncan Luce. Thurstone and sensory scaling: Then and now. *Psychological Review*, 101(2):271–277, Apr. 1994.

[84] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *European Conference on Computer Vision*, pages 386–399, 2008.

[85] Siwei Lyu. Divisive normalization: Justification and effectiveness as efficient coding transform. In *Advances in Neural Information Processing Systems*, pages 1522–1530, 2010.

[86] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database : New challenges for image quality

assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, Feb. 2017.

[87] Kede Ma, Huan Fu, Tongliang Liu, Zhou Wang, and Dacheng Tao. Local blur mapping: Exploiting high-level semantics by deep neural networks. *CoRR*, abs/1612.01227, 2016.

[88] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, Aug. 2017.

[89] Kede Ma, Wentao Liu, and Zhou Wang. Perceptual evaluation of single image dehazing algorithms. In *IEEE International Conference on Image Processing*, pages 3600–3604, 2015.

[90] Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group MAD competition − a new methodology to compare objective image quality models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2016.

[91] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 497–506, 2016.

[92] Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, Jul. 1989.

[93] Stephane G. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way.* Academic Press, 2008.

[94] James L. Mannos and David J. Sakrison. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, 20(4):525–536, Jul. 1974.

[95] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, Feb. 2004.

[96] Eftichia Mavridaki and Vasileios Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *IEEE International Conference on Image Processing*, pages 887–891, 2015.

[97] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2000.

[98] Xiongkuo Min, Kede Ma, Ke Gu, Guangtao Zhai, Zhou Wang, and Weisi Lin. Unified blind quality assessment of compressed natural, graphic and screen content images. *IEEE Transactions on Image Processing to appear*, 2017.

[99] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, Dec. 2012.

[100] Anish Mittal, Gautam S. Muralidhar, Joydeep Ghosh, and Alan C. Bovik. Blind image quality assessment without human training using latent quality factors. *IEEE Signal Processing Letters*, 19(2):75–78, Feb. 2012.

[101] Anish Mittal, Ravi Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, Mar. 2013.

[102] Anush K. Moorthy and Alan C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Processing Letters*, 17(5):513–516, May 2010.

[103] Anush K. Moorthy and Alan C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, Dec. 2011.

[104] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012.

[105] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *IEEE International Conference on Machine Learning*, pages 807–814, 2010.

[106] Netflix Inc. Per-title encode optimization, 2015. [Online]. Available: http://techblog.netflix.com/2015/12/per-title-encode-optimization.html.

[107] Fergus Nicol and Michael A. Humphreys. Adaptive thermal comfort and sustainable thermal standards for buildings. *Energy and Buildings*, 34(6):563–572, Jul. 2002.

[108] Seyfullah H. Oğuz, Yu Hen Hu, and Truong Q. Nguyen. Image coding ringing artifact reduction using morphological post-filtering. In *IEEE Workshop on Multimedia Signal Processing*, pages 628–633, 1998.

[109] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[110] Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, May 1981.

[111] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice-Hall Press, 2009.

[112] Ozgur Oyman and Sarabjot Singh. Quality of experience for HTTP adaptive streaming services. *IEEE Communications Magazine*, 50(4):20–297, Apr. 2012.

[113] Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507, Jul. 2005.

[114] Nikolay Ponomarenko and Karen Egiazarian. Tampere image database TID2008, 2008. [Online]. Available: http://www.ponomarenko.info/tid2008.

[115] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and

C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30(xx):57–77, Jan. 2015.

[116] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Jaakko Astola, Marco Carli, and Federica Battisti. TID2008 − a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, May 2009.

[117] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, 40(1):49–70, Oct. 2000.

[118] Umesh Rajashekar, Zhou Wang, and Eero P. Simoncelli. Quantifying color image distortions based on adaptive spatio-chromatic signal decompositions. In *IEEE International Conference on Image Processing*, pages 2213–2216, 2009.

[119] Abdul Rehman, Kai Zeng, and Zhou Wang. Display device-adapted video quality-of-experience assessment. In *SPIE Human Vision and Electronic Imaging*, pages 1–11, 2015.

[120] Azriel Rosenfeld. Picture processing by computer. *ACM Computing Surveys*, 1(3):147–176, Sep. 1969.

[121] Daniel L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548, Jan. 1994.

[122] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(9):533–536, Oct. 1986.

[123] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Fei-Fei Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec. 2015.

[124] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. A DCT statistics-based blind image quality index. *IEEE Signal Processing Letters*, 17(6):583–586, Jun. 2010.

[125] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, Aug. 2012.

[126] Robert J. Safranek and James D. Johnston. A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1945–1948, 1989.

[127] Ashirbani Saha and Qing Ming J. Wu. Utilizing image scales towards totally training free blind image quality assessment. *IEEE Transactions on Image Processing*, 24(6):1879–1892, Jun. 2015.

[128] Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson, and Peter L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, May 2000.

[129] Odelia Schwartz and Eero P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, Aug. 2001.

[130] Kalpana Seshadrinathan and Alan C. Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *IEEE Transactions on Image Processing*, 19(2):335–350, Feb. 2010.

[131] Claude E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, Jan. 2001.

[132] Hamid R. Sheikh and Alan C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb. 2006.

[133] Hamid R. Sheikh, Alan C. Bovik, and Lawrence Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *IEEE Transactions on Image Processing*, 14(11):1918–1927, Nov. 2005.

[134] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov. 2006.

[135] Hamid R. Sheikh, Zhou Wang, Alan C. Bovik, and Lawrence Cormack. Image and video quality assessment research at LIVE, 2006. [Online]. Available: http://live.ece.utexas.edu/research/quality/.

[136] Aleksandr Shnayderman, Alexander Gusev, and Ahmet M. Eskicioglu. An SVD-based grayscale image quality measure for local and global assessment. *IEEE Transactions on Image Processing*, 15(2):422–429, Feb. 2006.

[137] Eero P. Simoncelli, William T. Freeman, Edward H. Adelson, and David J. Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, Mar. 1992.

[138] Eero P. Simoncelli and Bruno A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, Mar. 2001.

[139] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[140] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, Aug. 2004.

[141] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Learning a blind measure of perceptual image quality. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 305–312, 2011.

[142] Huixuan Tang, Neel Joshi, and Ashish Kapoor. Blind image quality assessment using semi-supervised rectifier networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2877–2884, 2014.

[143] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, Dec. 2013.

[144] Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, Jul. 1927.

[145] Hanghang Tong, Mingjing Li, Hongjiang Zhang, and Changshui Zhang. Blur detection for digital images using wavelet transform. In *IEEE International Conference on Multimedia and Expo*, pages 17–20, 2004.

[146] Ming-Feng Tsai, Tie-Yan Liu, Tao Qin, Hsin-Hsi Chen, and Wei-Ying Ma. FRank: A ranking method with fidelity loss. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 383–390, 2007.

[147] Kristi Tsukida and Maya R. Gupta. How to analyze paired comparison data. Technical Report UWEETR-2011-0004, University of Washington, 2011.

[148] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, Dec. 2005.

[149] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, 2000. [Online]. Available: http://www.vqeg.org.

[150] Brian A. Wandell. *Foundations of Vision.* Sinauer Associates, 1995.

[151] Zhou Wang and Alan C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84, Mar. 2002.

[152] Zhou Wang and Alan C. Bovik. *Modern Image Quality Assessment.* Morgan & Claypool Publishers, 2006.

[153] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, Jan. 2009.

[154] Zhou Wang and Alan C. Bovik. Reduced- and no-reference image quality assessment: The natural scene statistic model approach. *IEEE Signal Processing Magazine*, 28(6):29–40, Nov. 2011.

[155] Zhou Wang, Alan C. Bovik, and Brian L. Evan. Blind measurement of blocking artifacts in images. In *IEEE International Conference on Image Processing*, pages 981–984, 2000.

[156] Zhou Wang, Alan C. Bovik, and Ligang Lu. Why is image quality assessment so difficult? In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 3313–3316, 2002.

[157] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr. 2004.

[158] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. The SSIM index for image quality assessment, 2004. [Online]. Available: https://ece.uwaterloo.ca/ z70wang/research/ssim/.

[159] Zhou Wang and Qiang Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, May 2011.

[160] Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *IEEE International Conference on Image Processing*, pages 477–480, 2002.

[161] Zhou Wang and Eero P. Simoncelli. Local phase coherence and the perception of blur. In *Advances in Neural Information Processing Systems*, pages 1435–1442, 2003.

[162] Zhou Wang and Eero P. Simoncelli. An adaptive linear system framework for image distortion analysis. In *IEEE International Conference on Image Processing*, pages 1160–1163, 2005.

[163] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *SPIE Human Vision and Electronic Imaging*, pages 1–11, 2005.

[164] Zhou Wang and Eero P. Simoncelli. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):1–13, Sep. 2008.

[165] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1398–1402, 2003.

[166] Zhou Wang, Guixing Wu, Hamid R. Sheikh, Eero P. Simoncelli, En-hui Yang, and Alan C. Bovik. Quality-aware images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, Jun. 2006.

[167] Andrew B. Watson and Denis G. Pelli. QUEST: A Bayesian adaptive psychometric method. *Attention, Perception, & Psychophysics*, 33(2):113–120, Mar. 1983.

[168] Stefan Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, Oct. 2012.

[169] Stefan Winkler. Image and video quality resources, 2016. [Online]. Available: http://stefan.winkler.net/resources.html/.

[170] Hong Ren Wu and Michael Yuen. A generalized block-edge impairment metric for video coding. *IEEE Signal Processing Letters*, 4(11):317–320, Nov. 1997.

[171] Qingbo Wu, Hongliang Li, Fanman Meng, King N. Ngan, Bing Luo, Chao Huang, and Bing Zeng. Blind image quality assessment based on multi-channel features fusion and label transfer. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(3):425–440, Mar. 2016.

[172] Qingbo Wu, Hongliang Li, King N. Ngan, and Kede Ma. Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator. *IEEE Transactions on Circuits and Systems for Video Technology to appear*, 2017.

[173] Qingbo Wu, Zhou Wang, and Hongliang Li. A highly efficient method for blind image quality assessment. In *IEEE International Conference on Image Processing*, pages 339–343, 2015.

[174] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, Sep. 2016.

[175] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *European Conference on Computer Vision*, pages 157–170, 2010.

[176] Long Xu, Weisi Lin, Jia Li, Xu Wang, Yihua Yan, and Yuming Fang. Rank learning on training set selection and image quality assessment. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2014.

[177] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C. Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, Nov. 2014.

[178] Wufeng Xue, Xuanqin Mou, Lei Zhang, and Xiangchu Feng. Perceptual fidelity aware mean squared error. In *IEEE International Conference on Computer Vision*, pages 705–712, 2013.

[179] Wufeng Xue, Lei Zhang, and Xuanqin Mou. Learning without human scores for blind image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, 2013.

[180] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695, Feb. 2014.

[181] Peng Ye and David Doermann. No-reference image quality assessment using visual codebooks. *IEEE Transactions on Image Processing*, 21(7):3129–3138, Jul. 2012.

[182] Peng Ye and David Doermann. Active sampling for subjective image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4249–4256, 2014.

[183] Peng Ye, Jayant Kumar, and David Doermann. Beyond human opinion scores: Blind image quality assessment based on synthetic scores. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4241–4248, 2014.

[184] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1098–1105, 2012.

[185] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Real-time no-reference image quality assessment based on filter learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 987–994, 2013.

[186] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over HTTP. *ACM SIGCOMM Computer Communication Review*, 45(4):325–338, Sep. 2015.

[187] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, Aug. 2015.

[188] Lin Zhang, Lei Zhang, and Xuanqin Mou. RFSIM: A feature based image quality assessment metric using Riesz transforms. In *IEEE International Conference on Image Processing*, pages 321–324, 2010.

[189] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug. 2011.

[190] Min Zhang, Xuanqin Mou, and Lei Zhang. Non-shift edge based ratio (NSER): An image quality assessment metric based on early vision features. *IEEE Signal Processing Letters*, 18(5):315–318, May 2011.

[191] Song-Chun Zhu, Yingnian Wu, and David Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27(2):107–126, Mar. 1998.

[192] Xiang Zhu and Peyman Milanfar. A no-reference sharpness metric sensitive to blur and noise. In *International Workshop on Quality of Multimedia Experience*, pages 64–69, 2009.

[193] Xiang Zhu and Peyman Milanfar. Automatic parameter selection for denoising algorithms using a no-reference measure of image content. *IEEE Transactions on Image Processing*, 19(12):3116–3132, Dec. 2010.

APPENDICES

Appendix A

Extended Applications of gMAD

The application scope of gMAD is far beyond IQA model comparison. As a general methodology, it can be used to compare any group of computational models that predict certain continuous quantities discriminable through human perception or other means. In this appendix, we demonstrate the gMAD competition methodology with two more examples: image aesthetics and video quality of experience (QoE).

A.1 Comparison of Image Aesthetics Models

As a highly subjective and abstract attribute, image aesthetics refers to the experience of beauty for subjects when viewing a photo [58]. It is generally accepted that image aesthetics is determined by a combination of low-level features such as composition, lighting, color arrangement and camera settings, and high-level semantics such as simplicity, realism, content type and topic emphasis [61, 143]. A successful objective image aesthetics model can be applied to many other fields such as image editing, image retrieval, and personal photo management.

Automatic assessment of image aesthetics is no easy task. Most existing image aesthetics models only make a binary decision on whether an image is a high-quality professional photo or a low-quality snapshot [19, 84, 82, 91]. Consequently, those models can only be

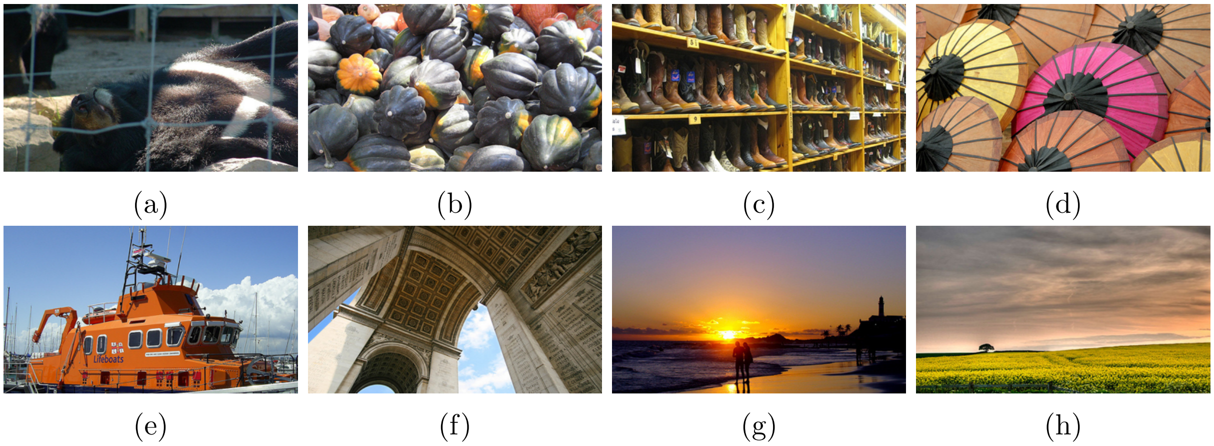


Figure A.1: Sample images from ImageNet [123] used for the gMAD competition of image aesthetics models. (a)-(h) Images with increasing degrees of perceived aesthetics according to our subjective testing. Images are cropped for better visibility.

tested on subject-rated image aesthetics databases with binary annotations [143, 37]. For databases that provide continuous-valued ground truths [104, 80], binarization is a must to take into account those binary classification-based models. In fact, the perceived aesthetics of real-world images is much richer than just two levels, making continuous-valued models more desirable in practice.

Here we aim to apply gMAD to compare continuous-valued aesthetics models. We first randomly select more than $170,000$ images from ImageNet [123] as the test database, whose content and aesthetics levels are very diverse. Images with small sizes have been manually removed. Sample images are shown in Fig. A.1. We select four image aesthetics models, which are GIST+SVR, aesthetics-aware features [96] with SVR (AAF+SVR), Jin16 [56], and Kong16 [64]. We implement our own version of GIST+SVR and AAF+SVR algorithms, and the codes of the other two models are obtained from the original authors. Specifically, for GIST [109], we work with 5 scales, 4 orientations and 16 blocks, and process RGB channels separately, resulting in a total of $5 \times 4 \times 16 \times 3 = 960$ features per image. Linear SVR [140] is adopted with hyperparameters optimized for the best prediction. For AAF, we choose the $1,323$-dimensional features proposed by Mavridaki and Mezaris [96], who implement a set of generally accepted photographic rules such as

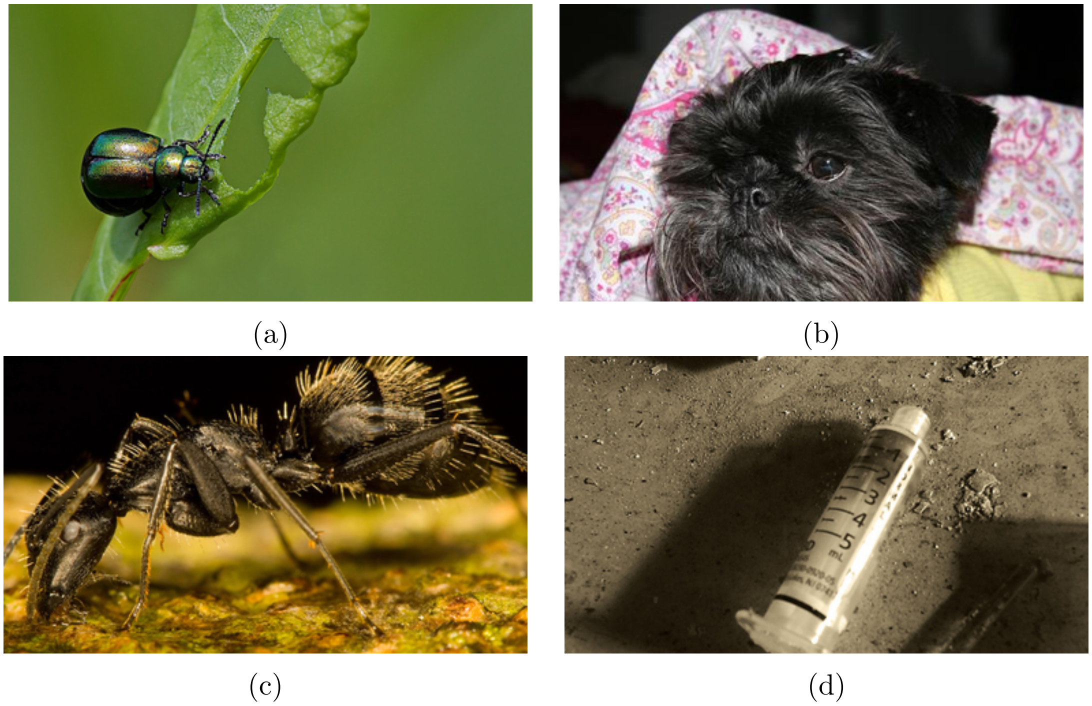
Table A.1: Global ranking results of image aesthetics models in the gMAD competition

Aesthetics model	Aggressiveness	Resistance
GIST+SVR [109]	−0.577	−0.097
AAF+SVR [96]	−0.189	−0.064
Kong16 [64]	0.145	−0.098
Jin16 [56]	**0.621**	**0.260**

simplicity, colorfulness, sharpness, image pattern, and composition. Linear SVR with the same hyperparameter optimization strategy is adopted. Jin16 [56] is a DNN-based algorithm that inherits the VGG16 [139] architecture and fine-tunes the weights for image aesthetics assessment using a weighted MSE loss. Kong16 [64] is another DNN-based model that fine-tunes the weights from AlexNet [68] using a weighted sum of a regression loss, a pairwise ranking loss, and an attribute loss. We train and validate GIST+SVR and AFF+SVR on AVA [104]. The weights of Jin16 [56] and Kong16 [64] fine-tuned from AVA [104] and AADB [64], respectively, are used for testing. Finally, we use the Waterloo IAA Database [80] to map all model predictions into the same perceptual space for comparison.

We choose 3 aesthetics levels and generate $4 \times 3 \times 3 = 36$ extremal image pairs. The subjective testing procedure is very similar as described in Chapter 4 and we only highlight the differences here. 30 subjects (18 males and 12 females) participate in the experiment. Each subject takes about 10 minutes to finish rating all the pairs. After running the outlier detection and subject rejection algorithm, all subjects are valid and 2.1% of the total ratings are identified as outliers and removed.

We list the global ranking results of the four image aesthetics models in terms of aggressiveness and resistance in Table A.1. It can be observed that Jin16 [56], a DNN-based model, exhibits the strongest aggressiveness and resistance. To take a closer look, we show two extremal image pairs, where Jin16 competes with Kong16 [64], another DNN-based algorithm, at the high (the third) aesthetics level in Fig. A.2. It is clear that Jin16 successfully falsifies Kong16 by finding the image pair at the first row, where the image (a) looks more beautiful than the image (b). At the same time, Jin16 survives from the



(a) (b)

(c) (d)

Figure A.2: gMAD competition between Jin16 [56] and Kong16 [64] at the high (the third) aesthetics level. (a) Best Jin16 for fixed Kong16. (b) Worst Jin16 for fixed Kong16. (c) Best Kong16 for fixed Jin16. (d) Worst Kong16 for fixed Jin16. Images are cropped for better visibility.

attack by Kong16 as evidenced by approximately the same aesthetics of the image pair at the second row according to our subjective testing. We conjecture that the superiority of Jin16 over Kong16 arises because 1) the backbone of Jin16—VGG16 [139]—is easier to generalize to novel tasks than AlexNet [68] used in Kong16; 2) the weighted loss that offsets the aesthetics level imbalance in Jin16 has more potentials to improve the performance than adding the pairwise ranking and attribute losses in Kong16. Moreover, it is not surprising that the general purpose feature representation GIST [109] for holistic scene modeling is defeated by AAF [96] under the same training configuration. After all, AAFs are motivated by years of practices for professional photography and are more relevant to image aesthetics. Finally, the hand-crafted AAFs are slightly better in terms of resistance than the end-to-end optimized Kong16 [64], which calls for larger training data, novel DNN architectures, and advanced optimization techniques to learn more robust image aesthetics models.

A.2 Comparison of Video QoE Models

Video streaming services have gained increasing popularity due to the fast deployment of network infrastructures and the booms of smart mobile devices. Being able to predict the QoE of end users is of great importance because it plays a dominant role in the user choices of different video streaming services according to a recent survey [12]. Three major factors affect the QoE for HTTP adaptive streaming (HAS) [23, 21]. The first is the presentation quality of video segments encoded in different bitrates, spatial resolutions, and frame-rates. The second is the stalling events due to bad network conditions, characterized by their frequencies and time durations. The third is the switchings of video segments of different bitrates, spatial resolutions, and frame-rates from one time segment to another, adapting to varying network conditions. Developing objective QoE models that jointly consider these three factors and their interactions is a sophisticated and challenging task. In recent years, many QoE models have been developed [130, 112, 22], trying to account for some of these factors or for some specific applications. However, most models have not been tested on or calibrated against subjective data with sufficient variations of video content and distortions. Note that the largest subject-rated streaming video database only contains hundreds of videos.

We build a large video streaming database as the playground for the gMAD competition of QoE models. Specifically, we first download 50 high-quality videos of size 4K and 24-30 frames per second (fps) from the Internet, which carry a Creative Commons license, and down-sample all videos to $1,920 \times 1,020$ to further damp possible compression artifacts. They are selected to cover sufficient content variations and motion patterns. Frames of representative videos are shown in Fig. A.3. From each video we extract a 10-second video clip [29], which is further divided into 5 non-overlapping 2-second segments. Each segment is encoded using H.264 into 5 representations selected from the Netflix's encoding ladder [106], indicating “bad”, “poor”, “fair”, “good”, and “excellent” presentation quality.

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure A.3: Representative frames from the test streaming video database for the gMAD competition of QoE models. (a) YellowStone: natural, high motion. (b) StreetDance: outdoor, high motion. (c) SplitTrailer: human, high motion. (d) CSGO: animation, high motion. (e) UCLY: indoor, slow motion. (f) WildAnimal: animal, slow motion. (g) Rose: plant, slow motion. (h) Food: still-life, slow motion. Frames are cropped for better visibility.

The details of the encoding ladder are given in Table A.2. After that, we prepend a stalling event to each encoded segment with a time duration of 0, 2, or 4 seconds, representing "no", "short", and "long" stalling events. We concatenate all possible combinations of 2-second segments from the same source content along with the stalling events, resulting in a total of $3^5 \times 5^5 \times 50 = 37,968,750$ test video clips.

We let 3 objective QoE models play the gMAD game. These are Liu12 [81], Yin15 [186], and SQI [23]. Liu12 [81] adopts bitrate and stalling percentage as two features. On top of Liu12, Yin15 [186] adds two more features—switching magnitude and initial buffering duration (the stalling event before video play). Linear regression is used for the two models. Instead of using bitrate as the indication of presentation quality, SQI [23] resorts to advanced video quality models such as SSIMplus [119] to predict the presentation quality and considers the interactions between video presentation quality and playback stalling experiences. We make use of the Waterloo QoE Database [22] and map all model responses to the same perceptual scale.

We choose 3 QoE levels and generate $3 \times 2 \times 3 = 18$ extremal video pairs. The same 30

Table A.2: Encoding ladder of video clips. kbps: kB per second

Representation	Bitrate (kbps)	Resolution
Bad	235	320×240
Poor	560	512×384
Fair	$1,050$	640×480
Good	$2,350$	$1,280 \times 720$
Excellent	$5,800$	$1,920 \times 1,080$

Table A.3: Global ranking results of QoE models in the gMAD competition

QoE model	Aggressiveness	Resistance
Liu12 [81]	-0.106	0.010
Yin15 [186]	-0.161	-0.112
SQI [23]	**0.267**	**0.102**

subjects in the subjective testing for image aesthetics participate in the current subjective experiment. Two video clips in the same pair are played consecutively but in random order. Subjects are allowed to replay them until they are confident about their relative QoE on the two video clips. Each subject takes about 20 minutes to finish the experiment. After subjective data screening, no subject is rejected and 3.0% of the total ratings are identified as outliers.

The global ranking results of Liu12 [81], Yin15 [186], and SQI [23] are listed in Table A.3. It is no surprise that SQI outperforms the other two QoE models in terms of both aggressiveness and resistance measures. We also show the extremal video pairs between SQI and Yin15 in Fig. A.4, where it is not hard to observe that SQI defeats Yin15 at all QoE levels. Although widely used, bitrate is a poor measure for video presentation quality because using the same bitrate to encode different video content results in drastically different quality. Instead of using bitrate as in Liu12 and Yin15, SQI uses SSIMplus [119] as the presentation quality estimator, which is in closer agreement with human perception of video quality. Taking into account the interactions between presentation quality and stalling events is another important ingredient for SQI to win the competition. However, SQI does not consider the switching effect to the overall QoE. We believe a joint modeling

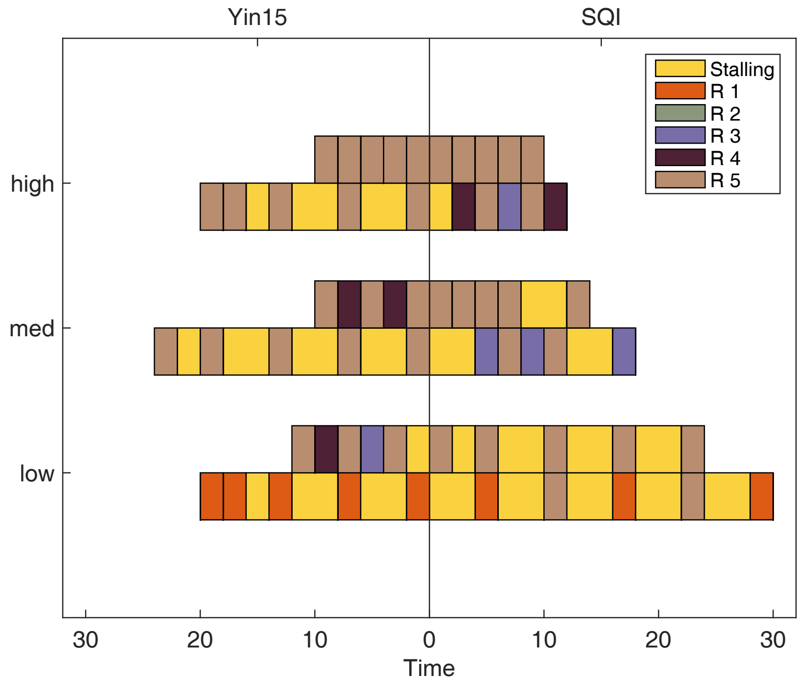


Figure A.4: gMAD competition between Yin15 [186] and SQI [23]. “R” stands for the representation level, characterized by bitrate and spatial resolution. In general, for the same content, the higher the representation level, the better the presentation quality. Left: Yin15 as the defender. Right: SQI as the defender.

of video presentation quality, stalling events, and switchings is a potential direction to further improve SQI. Compared to Liu12, Yin15 adds two more features, attempting to model the switching and initial buffering effects. Unfortunately, we observe a performance degradation in gMAD. This may be because Yin15 captures the switching effect using an oversimplified measure—the absolute difference between bitrates of two consecutive video segments, which may in turn hamper the overall performance. Specifically, bitrate and its difference exhibit a strong nonlinearity and (possibly non-monotonicity) to the overall QoE. Linearly incorporating it into the model may not be an appropriate choice. In addition, our latest results in [21] show that users have clearly different behaviors when experiencing positive and negative adaptations. In other words, the switching direction matters, but the absolute operation in Yin15 throws away such information. In summary, modeling the QoE of users when viewing streaming videos is challenging and current models only work to some degrees. A complete treatment of the aforementioned three factors is desirable to better predict streaming video QoE.

Appendix B

Publications During Ph.D. Studies

1. Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, accepted, 2017.

2. Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951-3964, Aug. 2017.

3. Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004-1016, Feb. 2017.

4. Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. Group MAD competition – a new methodology to compare objective image quality models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1664-1673, 2016.

5. Yuming Fang, Kede Ma, Zhou Wang, Weisi Lin, Zhijun Fang, and Guangtao Zhai. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters*, 22(7):838-842, Jul. 2015.