Blind Image Quality Assessment: From Heuristic-Based to Learning-Based

by

© Hao Cai

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

Memorial University of Newfoundland

July 2022

St. John's, Canada

Contents

Abstract

Image quality assessment (IQA) plays an important role in numerous digital image processing applications, including image compression, image transmission, and image restoration, etc. The goal of objective IQA is to develop computational models that can predict image quality in a way being consistent with human perception. Compared with subjective quality evaluations such as psycho-visual tests, objective IQA metrics have the advantages of predicting image quality automatically and effectively in a timely manner.

This thesis focuses on a particular type of objective IQA – blind IQA (BIQA), where the developed methods not only achieve objective IQA, but also are able to assess the perceptual quality of digital images without access to their pristine reference counterparts. Firstly, a novel blind image sharpness evaluator is introduced in Chapter 3, which leverages the discrepancy measures of structural degradation. Secondly, a “completely blind” quality assessment metric for gamut-mapped images is designed in Chapter 4, which does not need subjective quality scores during the model training. Thirdly, a general-purpose BIQA method is presented in Chapter 5, which can evaluate the quality of digital images without prior knowledge on the types of distortions. Finally, in Chapter 6, a deep neural network-based general-purpose BIQA method is proposed, which is fully data driven and trained in an end-to-end manner.

In summary, four BIQA methods are introduced in this thesis, where the first three are heuristic-based and the last one is learning-based. Unlike heuristics-based ones, the learning-based method does not involves manually engineered feature designs.

Acknowledgements

First and foremost, I want to express my sincere gratitude to my supervisor Dr. Minglun Gong for his continuous support and guidance throughout the years. I really appreciate his dedication of time, patience and encouragement to help me go through the whole journey.

Before graduation, I had the opportunity to visit University of Alberta for about a year. I would like to thank Dr. Li Cheng for the invitation and his constructive advice on my research.

I would like to thank Dr. Leida Li, Dr. Andrew Vardy, Sipan Ye, and many others for the thoughtful discussions on my work.

I would like to thank my supervisory committee members Dr. Wolfgang Banzhaf and Dr. Lourdes Pena-Castillo, and my thesis/proposal examiners Dr. Jie Liang, Dr. Matthew Hamilton, Dr. Xianta Jiang, Dr. Yuanzhu Chen, and Dr. Antonina Kolokolova for their valuable comments on my research. I would also like to thank the staff in our department and the School of Graduate Studies for the assistance in the program.

I would like to thank my labmates Mingjie Wang, Zili Yi, Wendong Mao, Xin Huang, Jun Zhou, Songyuan Ji, Shiyao Wang, Xue Cui, and Yiming Qian for the enjoyable time we had together. I want to thank all my friends in St. John's, Edmonton, Guelph, as well as other places around the world. You guys brought me so many wonderful memories.

Finally, I am very grateful to my family, especially my parents and my brother, for their unconditional love and care all the time.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Background

With the growing popularity of mobile devices, the last decade has seen a blossom of multimedia applications. People are thrilled by various apps, such as Facebook, Instagram, Tiktok, and Youtube, which can be used to share photos and videos with others. However, the quality of digital images is subject to degradation during the acquisition, transmission, and restoration processes [1, 2, 3, 4], as shown in Fig. 1.1[1]. For example, motion blur may be introduced by the motion of both image acquisition devices and moving targets, and the image compression process may lead to "blockiness" and "ringing" artifacts. These degradations usually lead to the loss of visual information, resulting in a poor experience for human viewers and difficulties for image processing and analysis at subsequent stages [5].

To obtain fine quality images from the massive amounts of available images, it

[1] *: This thesis focuses on natural (photographic) images which do not include images constructed randomly by computers.

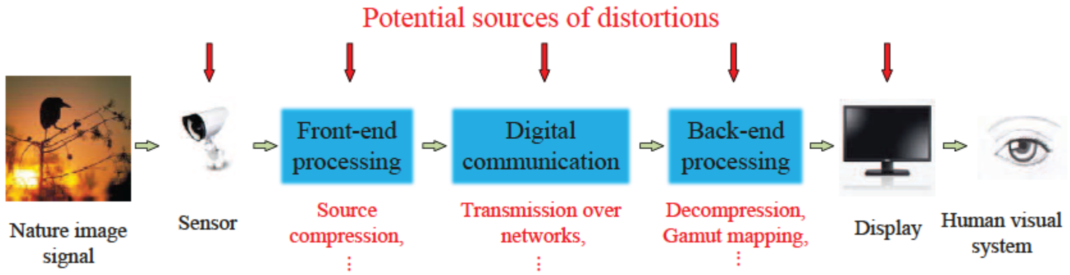


Figure 1.1: Potential sources of distortions in a simplified digital image processing pipeline [6].

is necessary to incorporate image quality assessment (IQA) in modern multimedia systems. Inevitably, the reliable assessment of image quality is of great importance in various sectors of image processing technologies, e.g., the quality monitoring of multimedia visual data services. With the help of IQA, an image processing or transmission system may have their parameters be fine-tuned, e.g., compression ratio for encoding solutions. In an image retrieval system, IQA metrics can be used to rank images. Moreover, IQA metrics can be utilized to benchmark image processing algorithms.

As Möller pointed out in Porikli *et al.* [7], visual quality is the result of a human judgment process, during which the perceiving human compares the features of the perceptual event to the features of some internal reference. Hence, the most straightforward approach for IQA is to perform psycho-visual tests following the standard international telecommunication union (ITU) recommendation [8], such as single-stimulus tests (e.g., absolute categorical rating) and double-stimulus tests (e.g., paired comparison, as shown in Fig. 1.2). For example, in a comparison test for gamut-mapped images, a pair of images obtained from different gamut mapping algorithms are shown either simultaneously or consecutively, the observers are asked which image

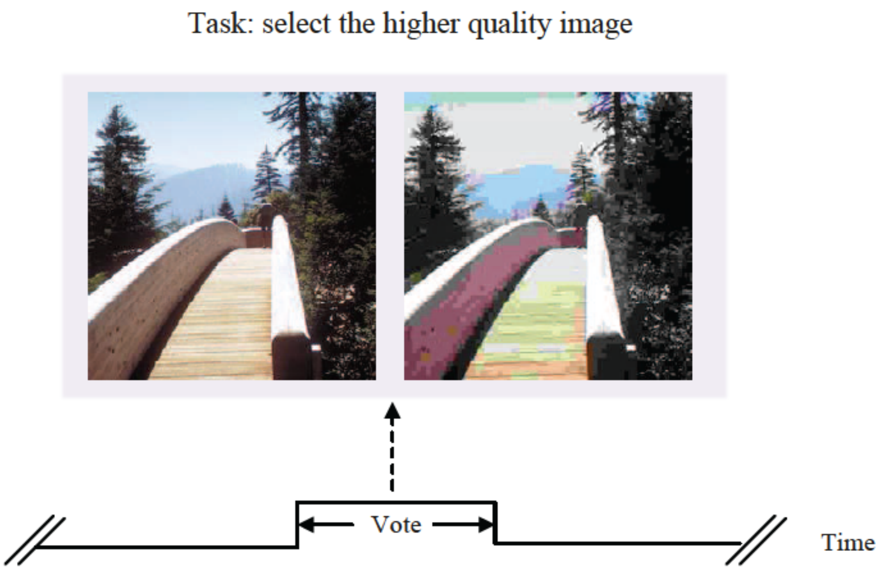


Figure 1.2: Overview of a paired comparison test [9].

is perceived to have better quality [10]. In a single-stimulus test, a reference image is usually presented to the observers as well, so as to reduce assessment bias due to the difference of image content [11]. The collected subjective ratings are typically averaged across all observers to obtain the mean opinion scores (MOS) or differential mean opinion scores (DMOS). For instance, in an absolute categorical rating test, the differential subjective score for the j-th test (distorted) image by the i-th observer can be computed as in Sheikh *et al.* [12]

$$d_{i,\text{dis}(j)} = r_{i,\text{ref}(j)} - r_{i,\text{dis}(j)}, \tag{1.1}$$

where $r_{i,\text{dis}(j)}$ denotes the raw subjective score for the j-th distorted image and $r_{i,\text{ref}(j)}$ is the raw subjective score for the corresponding reference image. To unify the inherent variability in visual quality judgment across different observers, $d_{i,\text{dis}(j)}$ is often

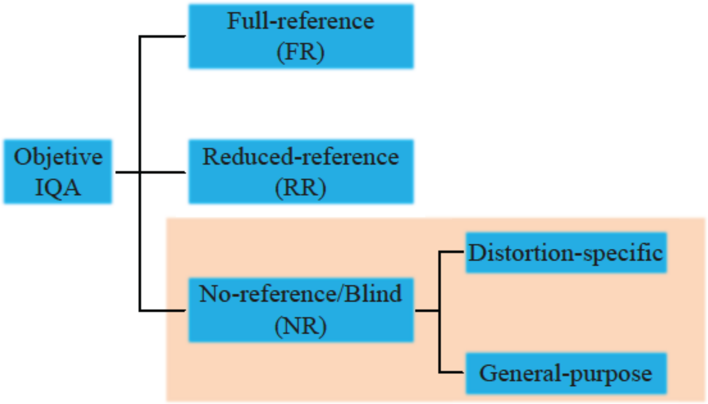


Figure 1.3: Categorization of objective IQA.

normalized into Z-score [13] as

$$z_{i,\text{dis}(j)} = \frac{d_{i,\text{dis}(j)} - \bar{d}_i}{\sigma_i}, \tag{1.2}$$

where $\bar{d}_i$ denotes the mean of raw subjective scores for all distorted images graded by the i-th observer and σ_i is the corresponding standard deviation. The Z-scores are then averaged across observers to compute the DMOS for the j-th distorted image.

The MOS or DMOS values produced from those psycho-visual experiments are generally accepted to be the "gold standard" for quality evaluation. However, these subjective tests are usually quite expensive, cumbersome, and time-consuming, and are thus not suitable for systems where real-time evaluation of image quality is required. On the other hand, objective IQA approaches can provide evaluation results more efficiently. The aim of objective IQA is to design computational models for measuring image distortions automatically and to predict image quality in a manner being consistent with human subjective perception at the same time [14].

The current objective IQA methods can be categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR) depending on the required amount of

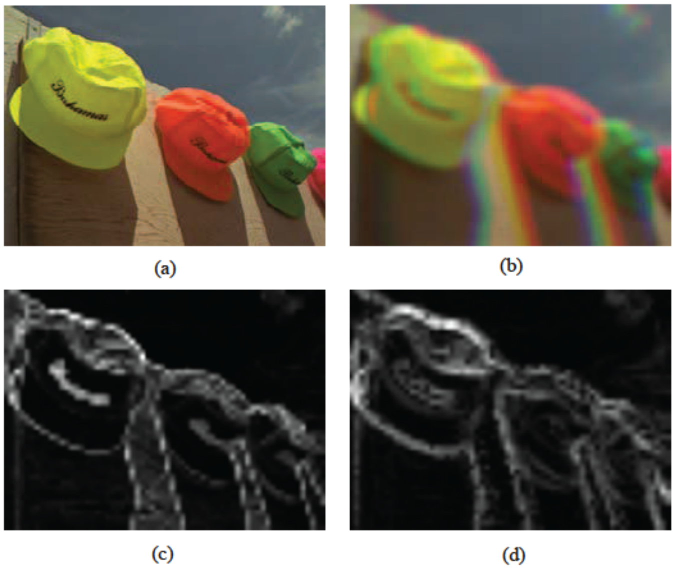


Figure 1.4: Feature maps generated by an FR-IQA metric [16]. (a) is the reference image. (b) is a distorted version of (a) with chromatic aberrations. (c) and (d) are their corresponding feature maps. The feature maps can be utilized to characterize quality degradation.

reference information [15]. A summary of different types of objective IQA approaches is shown in Fig. 1.3. When the reference images are available, FR-IQA approaches can be applied to directly quantify the disparities between distorted images and their reference versions. The peak signal-to-noise ratio (PSNR) model [17] is a classical FR-IQA method, which evaluated image quality by calculating pixel-to-pixel differences. However, this method has weak consistency with human perception in that it treats distortions in different regions equally. A renowned work called structural similarity index measure (SSIM) proposed by Wang *et al.* [18] addressed this problem through structural similarity, where local luminance and contrast deviations were combined to compute the quality scores. In VIF [19], the visual information fidelity model quantified image distortion by the amount of information lost from a reference image. Usually, FR-IQA methods would generate a map to indicate quality variations across spatial locations, as shown in Fig. 1.4. Unlike FR-IQA approaches that require the full information of reference image, in RR-IQA methods, partial information such as scaled entropies [20] extracted from the reference image is used to quantify image degradations. The RR-IQA methods provide a compromise between FR- and NR-IQA approaches in terms of both quality prediction accuracy and the amount of required information to describe the reference.

Since the pristine reference images are usually not available in real-world scenarios, the FR- and RR-IQA metrics thus have a rather limited application scope. Recent studies have put more efforts in developing NR/blind IQA metrics, where the information of corresponding reference images is not required in the quality prediction of distorted images. The blind IQA (BIQA) models are generally classified as either distortion-specific or general-purpose. Metrics in the former category are designed to quantify image quality with a presumed distortion type, such as blur, ringing, or blockiness [21, 22, 23, 24]. On the other hand, general-purpose BIQA models are able to produce quality scores without the prior information of distortion types. The focus of this thesis is on BIQA, as highlighted by the pink area in Fig. 1.3. Two methodologies have been explored in the development of effective BIQA approaches, from heuristic-based to learning-based. The relevant related work is introduced in Chapter 2.

Table 1.1: Commonly used image quality databases.

Database	Reference images	Distorted images	Distortion types	Observers	Subjective scores
LIVE	29	779	5	161	DMOS
CSIQ	30	866	6	35	MOS
TID2013	25	3000	24	971	MOS
LIVEWC	N.A.	1162	authentic	> 8100	MOS
CID2013	N.A.	474	12–14	188	MOS

1.2 Image Quality Databases and Evaluation Criteria

Image quality databases facilitate the IQA metric development and benchmarking, which are constructed through the subjective psycho-visual tests as discussed above. The associated MOS or DMOS values in the databases are served as the ground truth.

Table 1.1 lists some of the publicly available image quality databases, including three legacy databases: LIVE [18], CSIQ [25], and TID2013 [26], and two real camera image databases: LIVE in the Wild image quality Challenge database (LIVEWC) [27] and CID2013 [28]. The five distortion types in the LIVE database are white Gaussian noise, JPEG compression, JPEG2000 compression, Gaussian blur, and fast-fading transmission error. Sample images of each distortion type in this database can be found in Fig. 1.5. A single-stimulus methodology was adopted in the subjective quality ratings. The CSIQ database was generated with six distortion types, including JPEG compression, JPEG2000 compression, Gaussian blur, pink Gaussian noise, white Gaussian noise, and global contrast decrement. The largest widely-used

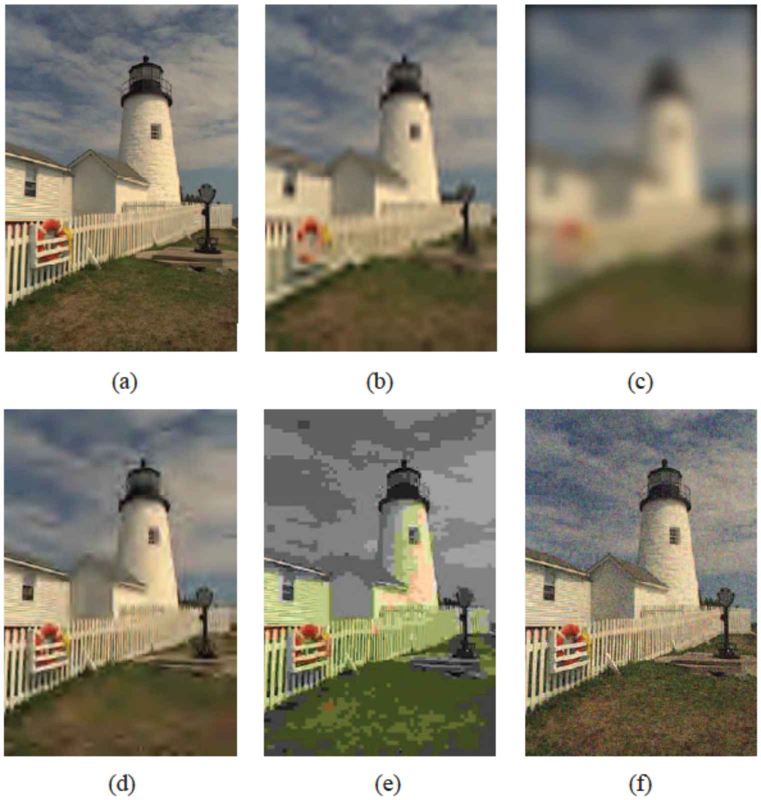
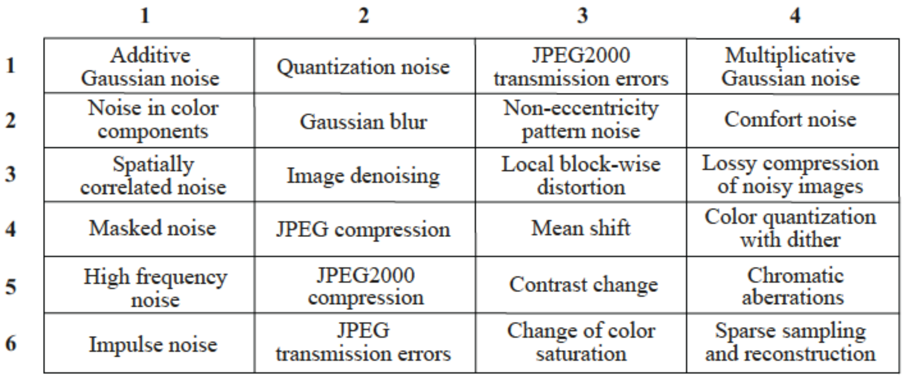


Figure 1.5: A pristine image and five distorted counterparts. (a) Pristine image. (b) Rayleigh fast fading. (c) Gaussian blur. (d) JPEG2000 compression. (e) JPEG compression. (f) White noise. Images are from the LIVE database [18].

database – TID2013 contains 24 types of distortions, as detailed in Fig. 1.6. For example, the index “$\#(5,3)$” refers to the “contrast change” distortion. In this database, a paired comparison methodology was utilized to obtain the MOS. Each distorted image participates in nine pair-wise comparisons and the scale of obtained estimations of MOS ranges from zero to nine. Distorted images in the LIVEWC database were captured using a variety of mobile devices without introducing extra artificial distortions. In the CID2013 database, images were contaminated with concurrent distortion types, and the realistic distortions were from multiple concurrent sources.



	1	2	3	4
1	Additive Gaussian noise	Quantization noise	JPEG2000 transmission errors	Multiplicative Gaussian noise
2	Noise in color components	Gaussian blur	Non-eccentricity pattern noise	Comfort noise
3	Spatially correlated noise	Image denoising	Local block-wise distortion	Lossy compression of noisy images
4	Masked noise	JPEG compression	Mean shift	Color quantization with dither
5	High frequency noise	JPEG2000 compression	Contrast change	Chromatic aberrations
6	Impulse noise	JPEG transmission errors	Change of color saturation	Sparse sampling and reconstruction

Figure 1.6: Distortion information for the TID2013 database [26].

Note that the LIVEWC and CID2013 databases were designed for BIQA only, where the reference images are not available. The legacy image databases include images with a single distortion source, whereas the two real camera image databases contain images afflicted by mixtures of multiple distortions.

Two commonly used criteria are adopted to quantitatively measure the performance of objective IQA metrics, including Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SRCC) [29]. Specifically, PLCC is used to measure the prediction accuracy and SRCC is used to measure the prediction monotonicity, which are defined as

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}, \tag{1.3}$$

$$\mathrm{SRCC} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{1.4}$$

where N is the number of test images, X_i and Y_i denote the MOS or DMOS value and the quality prediction of the i-th image respectively. $\bar{X}$ is the average value of X_i, x_i represents the corresponding rank of X_i, and $\bar{x}$ is the average value of x_i. These representations apply to $\bar{Y}$, Y_i, y_i, and $\bar{y}$ accordingly. In addition, a nonlinear fitting is usually required to map the predicted scores to the same scales of the MOS or DMOS values. In this thesis, a five-parameter logistic nonlinear mapping [12] is employed, which is described as

$$f(x) = \tau_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\tau_2 (x - \tau_3)}} \right) + \tau_4 x + \tau_5, \tag{1.5}$$

where τ_i denote the fitting parameters, $i = 1, 2, \cdot\cdot\cdot, 5$. x and $f(x)$ represent the predicted score and the mapped objective score respectively. IQA metrics with better quality evaluation performance are suppose to produce higher PLCC and SRCC values. Generally, the evaluation of a quality metric should be performed on multiple independent databases, so as to verify whether its performance is consistently reliable over generic image contents and/or distortion types.

1.3 Contributions

This thesis aims at developing efficient objective BIQA methods under various scenarios. Four BIQA metrics are presented, which cover both heuristic-based and learning-based methods. Particularly, two distortion-specific heuristic-based, one general-purpose heuristic-based, and one general-purpose learning-based BIQA metrics are introduced. The main contributions are summarized as follows:

- A novel blind image sharpness evaluator is proposed (see Chapter 3), which shows that the discrepancy measures of structural degradation between an input image and its "reblurred" version are effective indicators for sharpness evaluation. This work is published as a journal article [30]:

  Hao Cai, Mingjie Wang, Wendong Mao, and Minglun Gong. No-reference image sharpness assessment based on discrepancy measures of structural degradation. *Journal of Visual Communication and Image Representation*, 71:102861, 2020.

- An objective BIQA metric for gamut-mapped images is proposed (see Chapter 4), which is the first work that addresses the blind quality assessment of gamut-mapped images. This work is published as a journal article [31]:

  Hao Cai, Leida Li, Zili Yi, and Minglun Gong. Blind quality assessment of gamut-mapped images via local and global statistical analysis. *Journal of Visual Communication and Image Representation*, 61:250–259, 2019.

- A statistics-based general-purpose BIQA metric is proposed (see Chapter 5), which demonstrates the robustness of second-order statistical image features in quality assessment. This work is published as a journal article [32]:

  Hao Cai, Leida Li, Zili Yi, and Minglun Gong. Towards a blind image quality evaluator using multi-scale second-order statistics. *Signal Processing: Image Communication*, 71:88–99, 2019.

- A fully data-driven deep neural network is proposed for general-purpose BIQA (see Chapter 6), which demonstrates that the developed multi-scale integration strategy and multi-level supervision mechanism are able to enhance the feature representation capability for image quality prediction. This work will be submitted to a journal shortly, considering *IEEE Transactions on Multimedia.*

Co-Authorship Statement: I, Hao Cai, hold a principle author status for all the published or submitted works introduced in this thesis, as mentioned above. For all the works, I proposed the idea and the solution, conducted the experiments, wrote the manuscript, etc. The co-authors contributed on providing constructive comments, setting up experiments, and manuscript revision.

1.4 Thesis Outline

The remainder of this thesis is as follows. Chapter 2 discusses the related work in the literature. Chapters 3 – 6 introduce the four proposed methods in detail. Finally, Chapter 7 summarizes the thesis and discusses the methods, and Chapter 8 points out to promising directions for future work.

Chapter 2

Related Work

As indicated in Fig. 1.3, the current BIQA metrics can be categorized as either distortion-specific or general-purpose. The former category aims at a specific kind of distortion, whereas the latter category estimates the quality of an image without prior knowledge of the distortion types. This thesis presents two distortion-specific BIQA metrics in Chapter 3 and Chapter 4, where each chapter emphasizes one type of distortion (blur or gamut mapping), followed by two general-purpose BIQA methods introduced in Chapter 5 and Chapter 6 respectively. The following subsections discuss the relevant related work.

2.1 Distortion-Specific BIQA

2.1.1 Blur

As a key determinant in the perception of image quality, blur is typically characterized by the spread of edges [33]. The causes of blur are multifold, such as defocus, relative motion, and image compression. Generally, blur is likely to occur across the whole life-cycle of images. Note that image blurriness and sharpness are two sides of the same coin, the terms of “blur” metric and “sharpness” metric are often used interchangeably in the literature.

During the past few years, a series of image sharpness assessment methods have been proposed. Marziliano *et al.* proposed to measure the spread of image edges using the Sobel operator, and then the sharpness score was computed as the average width of edges [34]. In JNB [35], a sharpness model was derived from the measured just noticeable blur and the probability summation over space. In CPBD [36], the JNB model was further extended by calculating the cumulative probability of blur detection. In FISH [37], a fast image sharpness model was proposed by first decomposing the input image using discrete wavelet transform (DWT), and then image sharpness was measured via a weighted mean log-energies of the DWT subband coefficients. By using both the spectral and spatial properties, the spectral and spatial sharpness (S3) metric [38] quantified local perceived sharpness within and across images. In SVC [39], a simple image blur index was proposed based on the singular value curve. In LPC-SI [40], a local phase coherence (LPC) based sharpness index was introduced with the assumption that blur affects the LPC relationship near sharp image features. Bahrami *et al.* [41] utilized the maximum local variation (MLV) distribution of neighbouring image pixels as an indication of sharpness. In ARISM [42], an autoregressive image sharpness metric was proposed based on the hypothesis that image blurring increases the resemblance of locally estimated autoregressive model coefficients. Since blur affects the magnitudes of moments of an image, a blind image blur evaluation (BIBLE) algorithm [43] was proposed based on discrete

Tchebichef moments. In SPARISH [33], a blind sparse representation-based image sharpness index was proposed, where the variance-normalized energy over a set of selected high-variance blocks was computed as sharpness score. By learning multi-scale features extracted in both the spatial and spectral domains, Li *et al.* put forward with a blind and robust image sharpness evaluation (RISE) model [44]. In Zhan *et al.* [45], an efficient sharpness metric was proposed by combining the maximum gradient and the variability of gradients of blurry images, which is referred as “MVGV” in this thesis. Yu *et al.* proposed a sharpness metric [46] using a shallow convolutional neural network (CNN), where a single hidden layer CNN was employed for feature extraction. In GRNN [47], the general regression neural network was used to replace the multi-layer perception in the original CNN architecture, producing the CNN-GRNN model. Hosseini *et al.* proposed a metric called “HVS-MaxPol” [48] , which simulated the response of human visual system in a convolutional filter form. The MaxPol convolutional filters were utilized to decompose meaningful features that are related to image sharpness level. These discussed methods mainly centre on deriving effective features directly from the input image, whereas in this thesis the discrepancy measures of statistical regularity between the input image and its “reblurred” version is exploited to quantify image sharpness.

2.1.2 Gamut Mapping

Gamut mapping is a key technology to achieve high-quality cross-media color reproduction. For each device, such as a display, its color gamut is the set of colors that it can reproduce. If a color to be reproduced is beyond the device’s gamut, then it needs

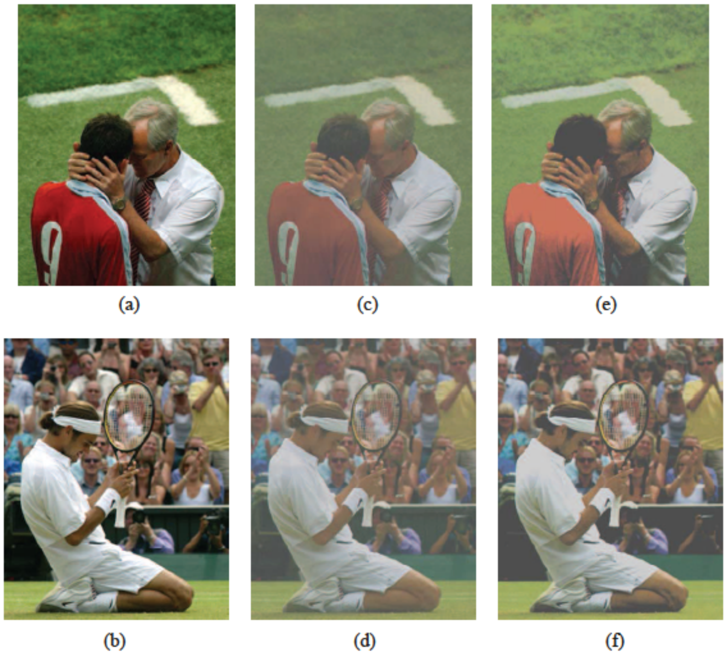


Figure 2.1: Color gamut distortion. (a) and (b) are the reference pristine images. (c) – (f) are gamut-mapped images generated by two different GMAs (i.e., SGCK and HPMinDE [56]), where (c) and (d) are generated by SGCK, and (e) and (f) are generated by HPMinDE.

a transformation, which is called “color gamut mapping” [49]. A thorough review of color gamut mapping is provided in Morovič [50].

In gamut mapping, distortions mainly appear on color while structural distortions may also exist, e.g., clipping. Another artifact of gamut mapping is lightness inversion which describes an inverse relation of two neighboring colors in the lightness channel [49]. Color distortions can have a high impact on image quality, especially when memory colors[1] (e.g., skin tones [51]) come into play. In most cases, conventional distortions operate locally, whereas gamut mapping distortions also incorporate global distortions. In gamut mapping databases, several gamut mapping algorithms (GMAs) are applied to a number of original images so as to generate the distorted images (see Fig. 2.1).

Early developed color image-difference metrics such as S-CIELAB [53] emphasized the spatial-color sensitivity of the human visual system, where the color difference at each pixel is weighted by the differences computed over a local neighborhood [54]. The iCAM framework [55] incorporated more sophisticated models of chromatic adaptation than S-CIELAB. It employs a single-scale spatial filtering and allows for the prediction of various appearance phenomena. Little work has been reported in the literature for the perceptual evaluation of gamut-mapped images, among which a large portion compares GMAs by incorporating psycho-visual tests as guided by the commission internationale de l'eclairage (CIE) technical committee [56]. In the past several years, efforts have been devoted to design objective metrics for assessing gamut-mapped images. In a related work [57], a color image quality metric was derived from the SSIM index [18], which can be used to evaluate the quality of gamut-mapped images. However, since this is an FR-IQA metric that requires ground truth pristine images available, its application scope is thus limited. In this thesis, a BIQA method for gamut-mapped images is proposed, which is able to produce quality assessment results in the absence of pristine images.

[1] *: A memory color is the typical color of an object that observers acquires through their experience with that object [52].

2.2 General-Purpose BIQA

The past decade has seen the emergence of several general-purpose BIQA methods, where the majority leverage the statistical regularities of natural images. Moorthy *et al.* proposed the blind image quality index (BIQI) that involves distortion classifying and distortion-specific quality assessment in sequence, leveraging the distorted image statistics to classify images into distortion categories [58]. BIQI was further extended to the distortion identification-based image verity and integrity evaluation index (DIIVINE) [59] by modeling the wavelet coefficients with generalized Gaussian distribution. In BLIINDS-II [60], an upgraded framework blind image integrity notator was proposed, which trains a probabilistic model based on discrete cosine transform statistics. In BRISQUE [61], the blind/referenceless image spatial quality evaluator utilizes the statistics of luminance coefficients to quantify the quality degradation. Zhang and Chandler proposed the derivative statistics-based quality evaluator (DESIQUE) [62], which uses log-derivative statistics and log-Gabor filters to extract image quality-related features in the spatial and frequency domain respectively. Follow the same framework as BIQI, Liu *et al.* proposed the spatial-spectral entropy-based quality (SSEQ) index [63], which utilizes local spatial and spectral entropy features on distorted images. By extracting gradient magnitude and Laplacian of Gaussian features, Xue *et al.* proposed a BIQA model referred as GL-BIQA [64]. Gu *et al.* proposed a metric named NR free energy-based robust metric (NFERM) [65] that utilizing free energy-based brain theory and classical human visual system inspired features. In NIQE [66], the natural image quality evaluator was proposed, which extracts features based on a multivariate Gaussian model and relates them to perceived quality in an unsupervised manner. Since NIQE uses a single global multivariate Gaussian model to describe an image, Zhang *et al.* extended NIQE to integrated local NIQE (ILNIQE) [67] by gathering multiple selected natural statistics scene features in both the spatial and frequency domain. A model named quality-aware clustering (QAC) [68] was proposed by Xue *et al.*, which is learned from a set of quality-aware centroids to compute the quality levels of image patches. These general-purpose BIQA methods mostly revolve around first-order statistics, that is, extracting quality-aware features from the characteristics of single image pixels. In this thesis, second-order statistics that emphasizing the relationship between neighbouring pixels/coefficients are examined and demonstrated to be effective to measure the extent of image distortions.

More recently, researchers have put attention on developing learning-based general-purpose BIQA methods, leveraging the advance of CNN methodologies. Kang *et al.* developed a network model with one convolutional and two fully connected layers for BIQA, which is referred to as “Kang’s CNN” [69] in this thesis. Bianco *et al.* adopted CNN features that are pretrained on the image classification task as inputs to learn a quality evaluator [70]. In Ma *et al.* [71], a discriminable image pairs inferred quality index was proposed. The input image pairs are automatically generated with the help of current available FR-IQA methods. In BIECON [72], a blind image evaluator was proposed, which utilizes FR-IQA metrics to estimate patch-wise quality maps, and the pooling process is incorporated as a single layer in the training. Pan *et al.* proposed a model called “BPSQM” [73], which is able to predict a pixel-by-pixel similar quality map from a distorted image under the guidance of similarity maps derived by FR-IQA methods. In DIQaM [74], a deep image quality measure model was introduced, which adapts a Siamese network and comprises ten convolutional layers and five pooling layers for feature extraction. After a slightly adaptation, this model can fulfill both FR- and NR-IQA. Liu *et al.* [75] proposed a model RankIQA, which learned from rankings. A Siamese network was trained to rank images in terms of image quality, leveraging synthetically generated distortions. In MEON [76], an end-to-end optimized deep neural network was designed, where the model consists of two sub-networks: a distortion identification network and a quality prediction network. Kim *et al.* [77] proposed a deep image quality assessor (DIQA), where the training process consists of two stages: regression onto objective error maps and regression onto subjective scores. A model named "DistNet-Q3" was developed in Dendi *et al.* [78], which utilizes a convolutional autoencoder for distortion map generation. The SSIM index [18] was used as a proxy for the generation of the ground-truth distortion maps. In NSSADNN [79], a multi-task natural scene statistics aided neural network model was proposed, where the auxiliary statistical feature prediction task helps the quality score prediction task to learn mapping between the input image and its quality score. Zhang *et al.* [80] proposed a deep bilinear model (DB-CNN), which treats the synthetic and authentic distortions as two-factor variations, and bilinearly pools the two pretrained feature sets into a unified representation. These learning-based general-purpose BIQA methods largely neglect the importance of multi-level supervision and multi-scale integration, while multi-granularity features have been investigated in other realms of vision tasks and shown to be effective in capturing various semantic information in images. In this thesis, a novel learning-based BIQA approach is introduced, which exhibits superior performance in the evaluation of image quality.

Chapter 3

Blind Image Sharpness Assessment

This chapter presents a novel blind image sharpness evaluator which leverages the discrepancy measures of structural degradation in both the spatial and wavelet domains. The computed discrepancies are utilized as sharpness-aware features and then a support vector regressor is employed to map the feature vectors into quality scores. Experimental results demonstrate the effectiveness of the proposed method.

3.1 Introduction

It is worth noting that the discrepancy between an image and its “reblurred” version usually indicates the extent of blur in the image. This lies in the fact that blur mainly influences the high-frequency components of an image, whereas the low-frequency components remain quite stable [22]. Particularly, blur changes the structures of a sharp image greater than that of a blurred image. Fig. 3.1 shows the impact of

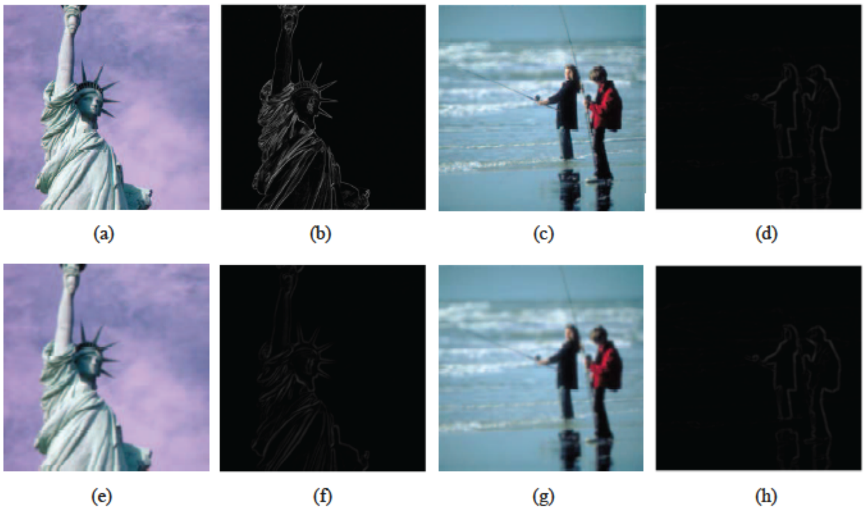


Figure 3.1: Impact of reblurring. (a) Sharp image. (b) Gradient map of (a). (c) Blurred image. (d) Gradient map of (c). (e) "Reblurring" of (a). (f) Gradient map of (e). (g) Reblurred version of (c). (h) Gradient map of (g). The gradient images are converted to the range [0, 255] for better display. This figure shows that blurring has more impact on sharp image than on blurred image.

reblurring on sharp and blurred images[1]. The reblurring processes are conducted by applying the same Gaussian low-pass filter on Fig. 3.1 (a) and (c). Since edges are better represented in gradient domain, here we propose to compute the gradients, in which case most of the low-frequency components are removed and high-frequency components become dominant which are sensitive to blur. From the comparison of Fig. 3.1 (b) and (f), and the comparison of Fig. 3.1 (d) and (h), we can see that blurring has more impact on the sharp image than on the blurred image. Generally, an image is considered as blurred if it is similar to its reblurred version, and thus the

[1] *: The gradient maps in Fig. 3.1 are normalized based on the overall maximum and minimum gradient values of all the sharp and blurred images.

discrepancy can be taken as an indicator of the extent of blur.

Based on the above observations, this chapter presents a novel blind image sharpness metric, which leverages the discrepancy measures of structural degradation. For an input image, its reblurred version is first obtained by applying a Gaussian low-pass filter. Studies in neuroscience have shown that the human visual system (HVS) exhibits orientation selectivity mechanism for perception and understanding, and the arrangement of excitatory/inhibitory cortex neuron arises orientation selectivity in a local receptive field [81, 82]. To this end, local spatial structures of an image are characterized by orientation selectivity-based visual patterns. Furthermore, it has been proved that the power spectrum of a blurred image falls faster than the sharp image [83], the reduction of the high-frequency components can then be utilized to evaluate image sharpness. Here we propose to extract sharpness-aware features by calculating the discrepancies of orientation selectivity-based visual patterns and log-Gabor filter responses between the input image and its corresponding reblurred version. Experimental results on both synthetically and real blurred image databases demonstrate that the proposed method performs consistently well across several databases and also shows good generalization ability.

3.2 Proposed Method

Fig. 3.2 shows the flowchart of the proposed sharpness metric. For an input image, its reblurred version is first built. Then sharpness-aware features are extracted between the input image and its reblurred version. Taking into account that the perceived quality of an image is greatly affected by the viewing distance [84], multiresolution

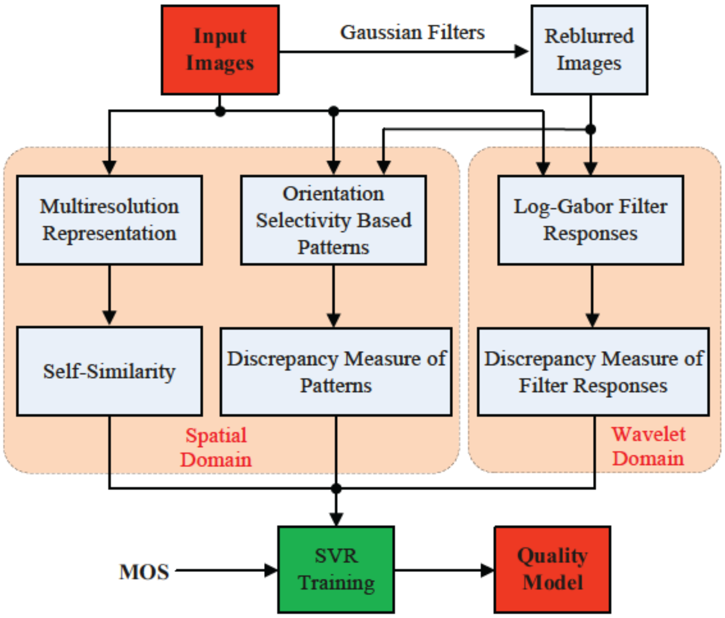

Input Images
Gaussian Filters
Reblurred Images
Multiresolution Representation
Orientation Selectivity Based Patterns
Log-Gabor Filter Responses
Self-Similarity
Discrepancy Measure of Patterns
Discrepancy Measure of Filter Responses
Spatial Domain
Wavelet Domain
MOS
SVR Training
Quality Model

Figure 3.2: Flowchart of the proposed image sharpness quality metric.

representation of the input image is constructed and then the inter-resolution self-similarities (i.e., the similarities between the original image and its multiresolution version) are calculated. Finally, to map the extracted features into an objective prediction score, a support vector regression (SVR) model [85] is utilized to train the quality model providing with the MOS values. The SVR model is then deployed for the subsequent quality prediction of test images. In the following subsections, the extraction of quality-aware features will be introduced in detail.

3.2.1 Self-Similarity

The self-similarity characteristic of natural images has been employed to applications like image compression [86]. In this work, we explore to measure the global sharpness discrepancy based on the fact that blurring would affect the global image

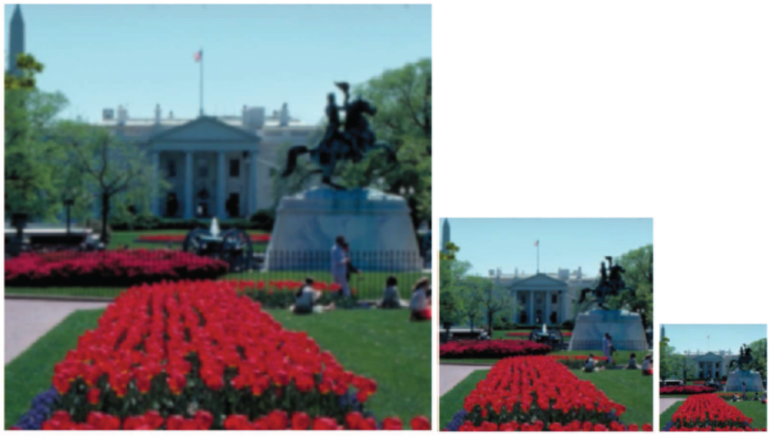


Figure 3.3: Impact of image resolution on sharpness. From left to right: Original image, images down-sampled by 2× and 4× respectively.

self-similarity attribute [87].

It has been pointed out that viewing distance and image resolution have substantial influences on image quality assessment [84]. Fig. 3.3 shows the impact of image resolution on sharpness. The original image is down-sampled by 2× and 4× in both both horizontal and vertical directions respectively, so as to generate the down-sampled images. Although blurriness can be clearly observed from the original image, the down-sampled images look sharper than the original image due to the reduction of spatial resolution. With this observation, we construct a multiresolution representation of the input image. The multiresolution representation of input image I is constructed by down-sampling it with t times, and the down-sampled images are denoted as I_1, I_2, ..., I_t, where I_t indicates the image with the lowest resolution. In this work, the scaling factor is set to 0.5, and thus I_1 represents down-sampling by 2×.

The global sharpness discrepancy is measured by computing the self-similarities between the input image and its multiresolution versions. Suppose the input image I is denoted as I_0, the self-similarity S_{0j} between images I_0 and I_j is defined as

$$S_{0j} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{\left| \left(\delta_0^{(n)}\right)^2 - \left(\delta_j^{(n)}\right)^2 \right|}, \tag{3.1}$$

where $j \in [1,\ t]$. N is the number of non-overlapping partitioned blocks. $\delta_0^{(n)}$ and $\delta_j^{(n)}$ represent the standard deviations of the nth block in I_0 and I_j respectively. The number of blocks N is computed as

$$N = \left\lfloor \frac{X}{d} \right\rfloor \cdot \left\lfloor \frac{Y}{d} \right\rfloor, \tag{3.2}$$

where $X \times Y$ represents the resolution of the original input image, d is the block size, and $\lfloor\ \rfloor$ denotes floor rounding. Note that the block sizes of down-sampled images are also processed by the down-sampling operation as with the image resolution, therefore the original image I_0 and down-sampled images I_j have the same number of blocks. In this work, we set the block size of the original image $d = 32$, and $t = 4$ which means four down-sampled images are obtained. We calculate the inter-resolution self-similarities between the original image I_0 and four down-sampled images (I_1, I_2, I_3, I_4). The self-similarities S_{01}, S_{02}, S_{03}, and S_{04} are then utilized as quality-aware features to predict image sharpness.

3.2.2 Discrepancy Measure of Orientation Selectivity-Based Patterns

Previous studies have shown that orientation selectivity reveals the inner mechanism for visual structure extraction [88]. The orientation selectivity of each local receptive field can be represented using a set of binary values which called *pattern*. Particularly, these patterns characterize the spatial structures of local image textures. Inspired by this, we propose to estimate the local spatial structural degradation in a blurred image by measuring the histogram disparity of orientation selectivity-based visual patterns between the image and its reblurred version.

To obtain the reblurred version of an input image $I(x, y)$, a Gaussian low-pass filter is employed and defined as

$$g(x, y, \rho) = \frac{1}{2\pi\rho^2}\exp\left(\frac{-(x^2+y^2)}{2\rho^2}\right), \tag{3.3}$$

where ρ is the standard deviation. In this work, we utilize a Gaussian filter with a window size 3×3 and $\rho = 5$. While generating the reblurred images, we performed this filtering operation twice iteratively. We found that for extremely blurred images, further blurring produces little effect. The reblurred version of image I is denoted as I^b.

By imitating the arrangement of the interactions among cortical neurons, the correlations between a central pixel and its neighboring pixels are binarized. Then the orientation selectivity-based patterns can be obtained according to these correlations. Specifically, the pattern P for a pixel x of image I is described as the spatial correlations with its circularly symmetric neighborhood $\Phi = \{x_1, x_2, ..., x_n\}$

$$P\left(x|\Phi\right) = A\left(\xi\left(x|\Phi\right)\right) = A\left(\xi\left(x|x_1, x_2, ..., x_n\right)\right), \tag{3.4}$$

where A represents the arrangement of correlations and n is the number of neighbours. $\xi\left(x|\Phi\right)$ denotes the spatial correlations between x and neighbor pixels in Φ. Research in neuroscience indicates that the correlations among neurons in a local receptive field are extremely complex [88]. Each neuron may connect to thousands of cortical

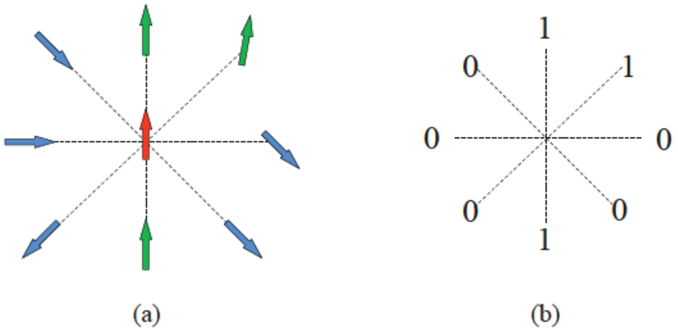


Figure 3.4: An example of orientation selectivity-based pattern. (a) shows the preferred orientations of a central pixel and its neighboring pixels, while (b) exhibits the generated pattern with respect to the central pixel. ‘1’ and ‘0’ represent excitatory and inhibitory interactions respectively.

neurons through synapses. Hubel and Wiesel [89] proposed to analyze the orientation selectivity mechanism by only utilizing the synapses between the central neuron and its neighboring neurons. To this end, here we recompute the pattern P as

$$P\left(x|\Phi\right) \approx A\left(\xi\left(x|x_1\right), \xi\left(x|x_2\right), ..., \xi\left(x|x_n\right)\right), \tag{3.5}$$

where $\xi\left(x|x_i\right)$ is the interaction between pixels x and x_i, $i \in [1, n]$.

The interaction type (excitatory/inhibitory) between cortical neuron pair is determined by the correlation of their received stimuli. Specifically, cortical neurons that have higher approximation degree with the preferred orientations would respond as excitatory, and vice versa [90]. For instance, Fig. 3.4 (a) shows the preferred orientations of a central pixel and its neighboring pixels ($n = 8$). The green arrows indicate the “excitatory interactions” since their preferred orientations are highly approximated with that of the central pixel (red arrow), while the blue arrows denote the “inhibitory interactions”. Therefore, the interaction $\xi\left(x|x_i\right)$ can be computed using the orientations of pixels x and x_i. For a pixel x of image I, the orientation $\theta(x)$ is defined as

$$\theta(x) = \arctan\frac{G_v(x)}{G_h(x)}, \tag{3.6}$$

where $\theta \in [-180°, 180°]$, G_v and G_h denote the vertical and horizontal gradients respectively. In this work, G_v and G_h are calculated as

$$G_v = [-1 \quad 1] * I, \quad G_h = [-1 \quad 1]^{\mathrm{T}} * I, \tag{3.7}$$

where $*$ denotes the convolution operation and T denotes the transpose operation.

The interaction $\xi(x|x_i)$ is estimated according to the approximation degree with the preferred orientations, which is defined as

$$\xi(x|x_i) = \begin{cases} 1, & \text{if } |\theta(x) - \theta(x_i)| < \Theta \\ 0, & \text{otherwise,} \end{cases} \tag{3.8}$$

where Θ denotes the orientation threshold, '1' and '0' represent excitatory and inhibitory interactions respectively. Campbell *et al.* [91] pointed out that nearby gratings with highly approximated orientations would cause masking effect, and the masking effect becomes marginal when the orientation difference is beyond a certain degree, e.g., 12°. Taking into account that the preferred orientations lie in two sides, Θ is set to 6°.

Finally, through Eq. 3.5 and Eq. 3.8, the correlations between a central pixel and its neighboring pixels are binarized. Fig. 3.4 (b) shows the generated pattern with respect to the orientation correlations depicted in Fig. 3.4 (a). In this work, we set $n = 8$, in which case the orientation correlations are transformed as 8-bit patterns.

For example, Fig. 3.4 (b) exhibits a pattern [00100011], which is formed starting from the right center neighbouring pixel in clockwise.

Note that the primary visual content of an image can be represented by several histograms of structural patterns [88]. In this work, the local spatial structural degradation of an image is characterized by the histogram discrepancy of orientation selectivity-based visual patterns between the image and its corresponding reblurred version. The structural patterns of all pixels in an image are computed to build a structural histogram. Generally, pixels producing the same patterns are combined. Hence, the structural histogram of an input image I is described as

$$H(k) = \sum_{c=1}^{N} w(x_c)\varphi(P(x_c), P^{(k)}), \tag{3.9}$$

where $H(k)$ represents the histogram value for the kth bin, N is the number of pixels, $P^{(k)}$ denotes the kth fundamental pattern and $w(x_c)$ is the weighting factor. The thresholding function $\varphi(\cdot)$ is defined as

$$\varphi(P(x_i), P^{(k)}) = \begin{cases} 1, & \text{if} \quad P(x_i) = P^{(k)} \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

Since blur would decrease the variance values for most of the image regions [33], here we set $w(x_c) = \mathrm{var}(x_c)$, where $\mathrm{var}(x_c)$ is the local variance with regard to pixel x_c.

It is clear that local receptive field with 8 neighbors will result in 2^8 patterns. However, patterns with the same excitatory subfield usually represent similar response [88]. Also, excitatory subfield denotes the number of excitations. For example, the pattern in Fig. 3.4 (b) denotes three excitatory interaction. To this end, these 2^8 patterns can be further divided into 9 types of fundamental patterns according to

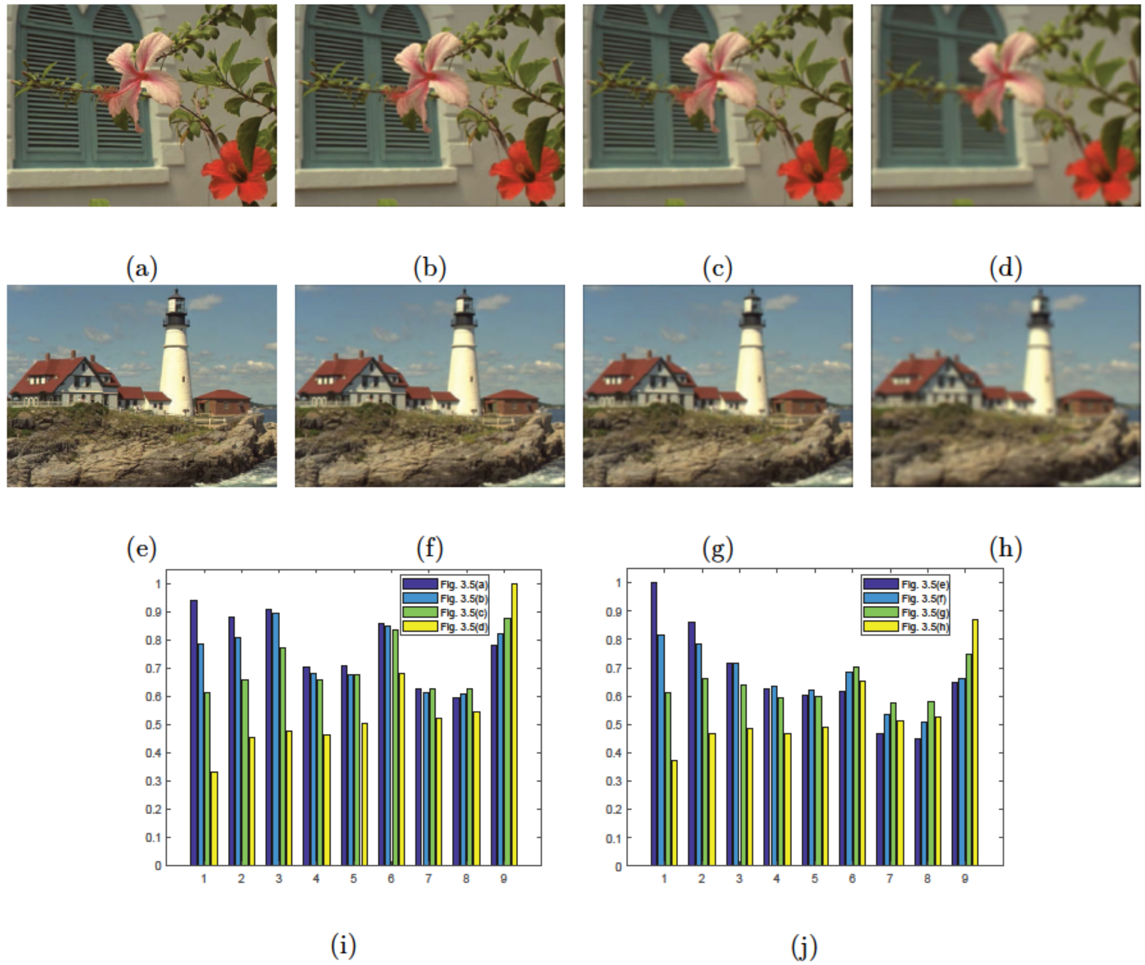


Figure 3.5: Examples of structural histogram change. (a) and (e) are sharp images. (b) – (d) and (f) – (h) are the blurred versions of (a) and (e) respectively, with increased degree of blur. (i) shows the structural histograms obtained from (a) to (d), while (j) exhibits the structural histograms featuring (e) to (h). Images are from the TID2013 database [26].

their excitatory subfields. In this case, an input image I can be mapped into a 9-bin histogram ($k = 9$). Particularly, patterns with smaller excitatory subfield are probably to appear in disorderly regions, while patterns with larger excitatory subfield are more likely to appear in orderly regions.

Fig. 3.5 shows the histograms of orientation selectivity-based visual patterns on sharp and blurred images. It can be clearly observed that the structural histograms vary between the sharp image and its blurred version. Generally, patterns correspond to edge regions (e.g., trees) are defined as disorderly patterns, while patterns correspond to plain regions (e.g., a clear sky) are defined as orderly patterns [88]. The fact that blur smooths edge regions, which may change a disorderly pattern into an orderly pattern. By comparing the histograms in Fig. 3.5 (i) and the histograms in Fig. 3.5 (j), we can see that the energies of most bins are decreased as a result of blurriness, especially for the first and second bins, while the energies of the last bin is increased. The larger bin number denotes patterns with larger excitatory subfield as illustrated above, which means the last bin corresponds to patterns appeared in orderly regions. Furthermore, the histogram comparison results of Fig. 3.5 (a) and (e) indicate that different visual contents present different structural histograms.

With these observations, we believe that local spatial structural degradation can be characterized by the histogram change of visual patterns. We measure the discrepancy of structural degradation by computing the histogram similarity S_H between the original image and its reblurred version as

$$S_H = \frac{2 \times H \cdot H^b}{(H)^2 + (H^b)^2}, \tag{3.11}$$

where H represents the histogram of patterns for the original input image, and H^b denotes the histogram for its reblurred version. It should be noted that the similarity is calculated in an element-wise manner. Since each histogram involves nine bins, the histogram similarity S_H thus produces nine quality-aware features.

3.2.3 Discrepancy Measure of Log-Gabor Filter Responses

It is widely accepted that information from both the spatial and frequency domains play important roles in characterizing image structure degradation [40]. For example, blur would lead to a reduction in the variance of edge-pixel values in the spatial domain, whereas in the frequency domain blur would result in a reduction of the high-frequency components [62]. Here we propose to extract sharpness-aware statistical features in the wavelet domain by calculating the discrepancies of log-Gabor filter responses between the input image and its corresponding reblurred version.

Cortical cells in the visual cortex are highly sensitive to frequency for scene perception [58]. Studies have proved that the response properties (local frequency responses) of these cells can be well modeled by Gabor filters [92]. Compared with classical Gabor filters, the log-Gabor filters are able to alleviate the frequency distribution problem [93]. To form bandpass responses, we employ the log-Gabor filter as defined in Zhang *et al.* [62]

$$G_{s,o}\left(\omega, \theta\right) = \exp\left\{-\frac{\left[\log(\omega/\omega_s)\right]^2}{2\left[\log(\sigma_s/\omega_s)\right]^2}\right\} \times \exp\left(-\frac{(\theta-\mu_o)^2}{2\sigma_o^2}\right), \tag{3.12}$$

where $G_{s,o}$ denotes the log-Gabor filter with scale index s and orientation index o. ω represents the normalized radial frequency and θ is the orientation. Note that the perceptual decomposition (octaves) resembles the models of bandpass responses that occur in area V1 of visual cortex [94]. To capture the magnitude differences in the high-frequency band, an input image is decomposed with three scales $s \in \{1, 2, 3\}$ and over ten frequency orientations $o \in \{0^\circ, 18^\circ, 36^\circ, 54^\circ, 72^\circ, 90^\circ, 108^\circ, 126^\circ, 144^\circ, 162^\circ\}$, thereby producing thirty subbands. The other parameter values are consistent with those used in Zhang *et al.* [62].

It is worth noting that coefficients of natural images exhibit a Gaussian-like appearance, and the presence of distortion could affect the distribution of this Gaussian model [95]. Given a collection of image patches, their statistics can be characterized by quality-aware features computed from each selected patch. Inspired by this, we propose to utilize the distribution of Log-Gabor coefficients for sharpness evaluation. Particularly, the coefficients for each subband are modeled by a zero-mean generalized Gaussian distribution (GGD)[2] as described in Gu *et al.* [65]

$$f(x;\lambda,\gamma^2) = \frac{\lambda}{2\beta\Gamma\left(\frac{1}{\lambda}\right)}\exp\left[-\left(\frac{-|x|}{\beta}\right)^{\lambda}\right], \tag{3.13}$$

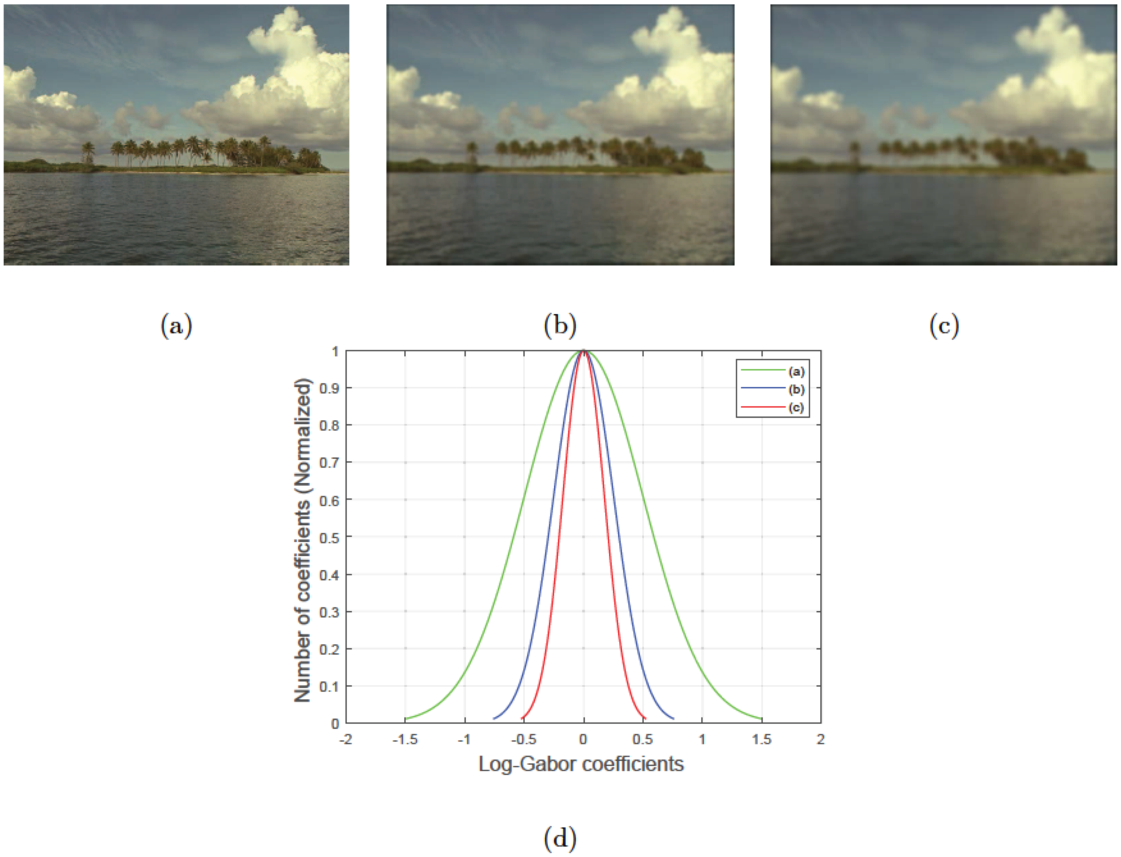
where $\beta = \gamma\sqrt{\frac{\Gamma(\frac{1}{\lambda})}{\Gamma(\frac{3}{\lambda})}}$. The function $\Gamma(\cdot)$ is defined as

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt, \quad a > 0. \tag{3.14}$$

The parameter λ controls the "shape" of the distribution, while γ^2 denotes the variance. This two parameters can be estimated by the moment-matching method illustrated in Sharifi *et al.* [96]. We deploy this parametric model to fit the distributions of Log-Gabor coefficients from both the input image and its corresponding reblurred version. Fig. 3.6 gives an example of histogram distributions of Log-Gabor coefficients for sharp and blurred images at subband ($s = 1, o = 0°$), where large MOS values indicate better image quality. It can be seen from Fig. 3.6 (d) that the histogram distributions exhibit Gaussian characteristics. Moreover, the distributions become more heavy-tailed and center-peaked as the image quality degrades. The model parameters (λ, γ^2) thus can be utilized to measure the quality degradation caused by blur.

Additionally, we find that the degradation of image sharpness usually leads to the decrease of the energy of subbands. Hence, the energy of each subband is taken

[2] *: GGD is useful when the errors around the mean or in the tails are of particular interest.


(a) (b) (c)

(d)

Figure 3.6: Example of histogram distributions of Log-Gabor coefficients. (a) is a sharp image. (b) and (c) are the blurred versions of (a), which correspond to subjective scores MOS=3.6154 and MOS=2.9250 respectively. (d) shows the histogram distribution of Log-Gabor coefficients at subband ($s = 1, o = 0°$) for the three images. Images are from the TID2013 database [26].

as another sharpness indicator. The subband energy $E_{s,o}$ is defined as the mean magnitude of the coefficients

$$E_{s,o} = \frac{1}{XY}\sum_{i=1}^{X}\sum_{j=1}^{Y}\left[G_{s,o}(i,j)\right]^2 , \tag{3.15}$$

where $X \times Y$ denotes the resolution of the input image. Fig. 3.7 shows the energy of subbands at scale ($s = 1$) for images in Fig. 3.6. It can be clearly seen that the

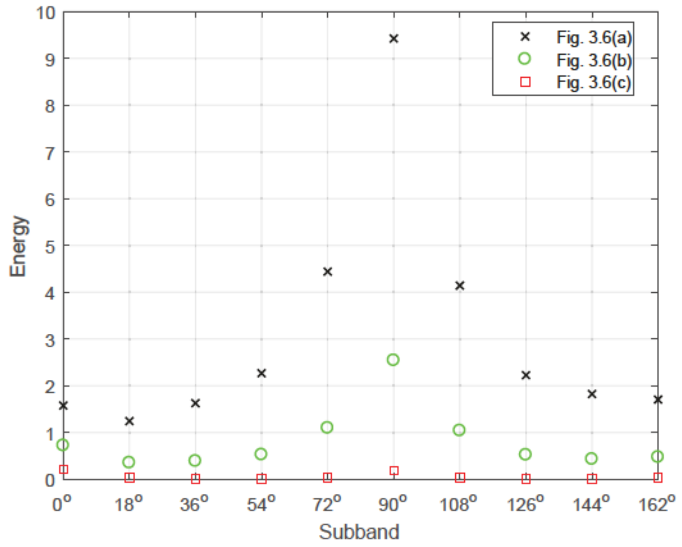


Figure 3.7: The energy of subbands at scale ($s = 1$) for images in Fig. 3.6.

energy of subbands changes accordingly to the extent of distortion, which indicates $E_{s,o}$ can be utilized for sharpness assessment.

The perceptual decomposition described in Eq. 3.12 facilitates multiple bandpass responses over frequency tuning orientations. In this work, we propose to extract multi-scale statistical features from the obtained subband coefficients to portray the local structural degradation of blurred images in the wavelet domain. Particularly, a similarity formulation [18] is adopted to compute the discrepancy (element-wise) of bandpass S_R responses between the input image and its corresponding reblurred version

$$S_R = \frac{2 \times R_z \cdot R_z^b + c_0}{(R_z)^2 + (R_z^b)^2 + c_0}, \tag{3.16}$$

where $z \in \{\lambda, \gamma^2, E\}$. R_z denotes the parameters of bandpass responses obtained from the input image, while R_z^b represents parameters from its reblurred version. $c_0 = 0.0001$ is a stabilizing constant.

As mentioned above, the input image is decomposed with three scales and over ten frequency orientations, S_R thus returns a set of ninety quality-aware features.

3.2.4 Regression Module

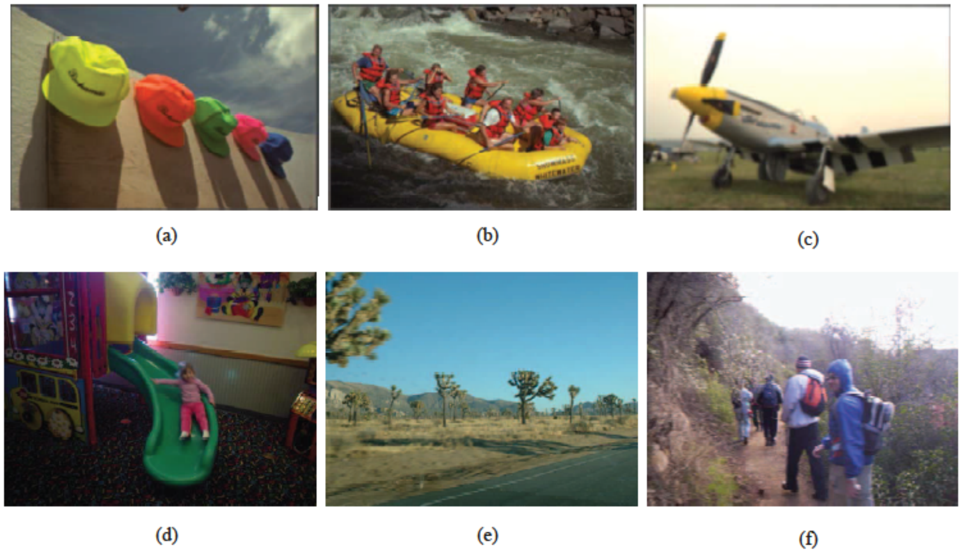
In the final stage, we utilize the support vector regression (SVR) module [85] to map the two categories of features into an overall quality score by the consideration that SVR is effective at handling high-dimensional data. Given a set of training images, quality-aware features are extracted in both wavelet domain and spatial domain respectively as described above. The extracted features and associated subjective rated scores (ground truth) are fed into the SVR for training, and then the learned model is utilized to predict the quality score of testing images.

Suppose the training set is described as $(\mathbf{x}_1, z_1), (\mathbf{x}_2, z_2), \cdots, (\mathbf{x}_i, z_i)$, where $\mathbf{x}_i$ is the feature vector and z_i is the ground truth, the SVR can be formulated as

$$\min_{\omega, b, \nu, \nu^*} \frac{1}{2}\omega^{\mathrm{T}}\omega + C\left(\sum_{i=1}^{l}\nu_i + \sum_{i}^{l}\nu_i^*\right) \quad s.t. \begin{cases} \omega^{\mathrm{T}}\phi(\mathbf{x}_i) + b - z_i \leq \epsilon + \nu_i, \\ z_i - \omega^{\mathrm{T}}\phi(\mathbf{x}_i) - b \leq \epsilon + \nu_i^*, \\ \nu_i, \nu_i^* \geq 0, \end{cases} \tag{3.17}$$

where ω denotes a high-dimensional vector variable. ν_i and ν_i^* are the slack variables. C is a hyper-parameter, ϵ is the constraint, and b is the bias. $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j)$ represents the kernel to be optimized.

We implemented the SVR using the publicly available libSVM package [97]. The radial base function (RBF) $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\eta \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right)$ was chosen as the kernel since it has fast convergence characteristic and approximates to the nonlinear function. η is the parameter of the kernel. A grid search technique with ten-fold



(a) (b) (c)

(d) (e) (f)

Figure 3.8: Sample blurred images from databases. Images (a), (b), and (c) are from the LIVE database [18]. Images (d), (e), and (f) are from the BID database [99]. Zoom in for better view.

cross-validation was employed to determine all the SVR parameters. More details about the SVR module can be found in Smola *et al.* [85].

3.3 Experimental Results and Analysis

The performance of the proposed sharpness metric was evaluated on six public image quality databases, including four synthetically blurred image databases : LIVE [18], CSIQ [25], TID2013 [26], and LIVE multiply distorted (LIVEMD) [98], as well as two real blurred image databases: BID [99] and CID2013 [28]. The numbers of blurred images in the six databases are 145, 150, 125, 450, 586, and 473 respectively. In LIVE, CSIQ, and TID2013, images were degraded by pure blur and were generated through applying Gaussian low-pass filters on pristine images. The database LIVEMD consists of images corrupted under two distortion scenarios (i.e., blur + JPEG/noise), and all the distortions were synthetically applied. In comparison, images in the real blurred image databases were captured by consumer-type cameras in uncontrolled environments, which are more challenging to evaluate. Fig. 3.8 shows several sample blurred images from the LIVE and BID databases. We can see that the synthetic blur in images (a), (b), and (c) distributes uniformly, whereas the real blur in images (d), (e), and (f) is much more complex. Subjectively rated MOS values are all provided in these databases as the ground truth. We employed two commonly used criteria PLCC and SRCC for evaluating the performance of the proposed sharpness metric, as defined in Section 1.2.

3.3.1 Performance Comparison

We first compared the performance of our proposed method against sixteen existing blind image sharpness metrics over five individual databases, namely LIVE [18], CSIQ [25], TID2013 [26], BID [99], and CID2013 [28]. The compared metrics are Marziliano [34], JNB [35], CPBD [36], FISH [37], S3 [38], SVC [39], LPC-SI [40], MLV [41], ARISM [42], BIBLE [43], SPARISH [33], RISE [44], MVGV [45], CNN-GRNN [47], Yu's CNN [46], and HVS-MaxPol [48]. In the experiment, we randomly divided the reference images along with their corresponding distorted images in each database into a training subset (80%) and a testing subset (20%). To avoid bias, this train-test partition was conducted a thousand times and the median performance of all test metrics were reported. For fair comparison, training-free metrics were evaluated on the corresponding testing subsets. Table 3.1 summarizes the experimental results on

Table 3.1: Performance comparison against blind image sharpness metrics. The two best results are marked in boldface.

	LIVE		CSIQ		TID2013		BID		CID2013	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Marziliano	0.8134	0.8113	0.7224	0.7536	0.7718	0.7617	0.2549	0.2417	0.5164	0.4278
JNB	0.8157	0.7934	0.8603	0.8270	0.7121	0.6877	0.2574	0.2290	0.5276	0.4438
CPBD	0.9016	0.9043	0.8138	0.8674	0.8590	0.8429	0.2603	0.2685	0.5146	0.4329
FISH	0.8924	0.8733	0.8655	0.8822	0.8253	0.7936	0.4682	0.4711	0.6472	0.5794
S3	0.9389	0.9314	0.9077	0.8967	0.8702	0.8528	0.4155	0.4120	0.6792	0.6373
SVC	0.9276	0.9159	0.9214	0.8975	0.8618	0.8542	0.4175	0.3485	0.3285	0.2483
LPC-SI	0.9233	0.9207	0.9251	0.8923	0.8826	0.8878	0.3866	0.3092	0.6927	0.5973
MLV	0.9351	0.9243	0.9368	0.9166	0.8724	0.8684	0.3582	0.3149	0.6743	0.6178
ARISM	0.9426	0.9411	0.9374	0.9137	0.8857	0.8802	0.1795	0.1646	0.5463	0.4673
BIBLE	0.9503	0.9513	0.9274	0.9059	0.8974	0.8865	0.3682	0.3578	0.6851	0.6772
SPARISH	0.9521	**0.9543**	0.9321	0.9067	0.8986	0.8879	0.3479	0.3408	0.6633	0.6548
RISE	0.9547	0.9402	0.9380	0.9211	0.9364	0.9228	**0.5904**	**0.5763**	**0.7832**	**0.7547**
MVGV	0.9432	0.9405	**0.9613**	**0.9481**	**0.9477**	**0.9503**	0.4377	0.4106	0.7038	0.6129
CNN-GRNN	0.9570	0.9438	0.9353	0.8990	0.9370	0.9031	0.5387	0.5260	0.7094	0.6938
Yu's CNN	0.9469	0.9486	0.9255	0.9048	0.8875	0.8376	0.5491	0.5412	0.7152	0.7039
HVS-MaxPol	**0.9762**	0.9378	0.9419	0.9124	0.8823	0.8726	0.4636	0.4483	0.7347	0.6155
Proposed	**0.9581**	**0.9547**	**0.9516**	**0.9233**	**0.9569**	**0.9409**	**0.6327**	**0.6032**	**0.8803**	**0.8739**

the five aforementioned databases. For each performance criterion and database, the two best results are highlighted in bold.

It can be seen from Table 3.1 that our proposed method achieves consistent good performance across these databases. For synthetic blur, it achieves the best prediction monotonicity in the LIVE database. In CSIQ, although not the best, our proposed method performs only slightly worse than the best metric. In TID2013, our proposed method achieves the best prediction accuracy, while the prediction monotonicity ranks the second. For real blur, it is worth noting that our proposed method outperforms all the compared metrics by a clear margin on the BID and CID2013 databases, regardless of prediction accuracy and monotonicity. Specifically, in CID2013, our proposed method achieves 0.8803 and 0.8739 for PLCC and SRCC respectively, while

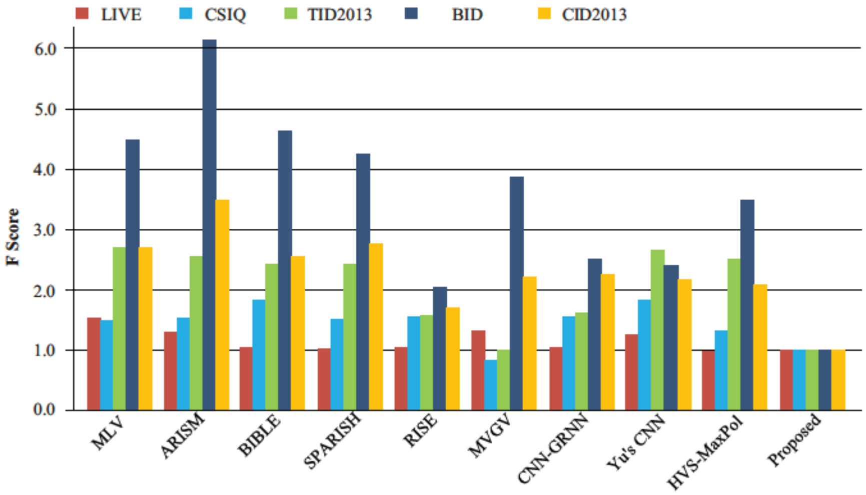


Figure 3.9: F-test results of nine compared sharpness metrics against the proposed method.

the second best results are only PLCC= 0.7832 and SRCC = 0.7547. From these results we know that our proposed method produces the state-of-the-art performance for both synthetic and real blur. Furthermore, We can notice that the current metrics usually perform better on synthetically blurred image databases than on the real blurred image databases.

To further analyze the statistical significance of the proposed method against the current leading sharpness metrics, we conducted the F-test as described in Sheikh *et al.* [12]. F-test is often used to determine if a metric has significantly larger (or smaller) prediction errors than another metric [16], which is based on an assumption of Gaussianity of the residual differences between the metric prediction and subjective rated score. Suppose the variances of prediction errors of a compared metric A and our proposed method are denoted as σ_A^2 and σ_B^2 respectively, the F-test score is then

Table 3.2: Statistical performance between the proposed method and nine compared sharpness metrics on five databases.

Metric	LIVE	CSIQ	TID2013	BID	CID2013
MLV [41]	+1	+1	+1	+1	+1
ARISM [42]	+1	+1	+1	+1	+1
BIBLE [43]	0	+1	+1	+1	+1
SPARISH [33]	0	+1	+1	+1	+1
RISE [44]	0	+1	+1	+1	+1
MVGV [45]	+1	–1	0	+1	+1
CNN-GRNN [47]	0	+1	+1	+1	+1
Yu's CNN [46]	+1	+1	+1	+1	+1
HVS-MaxPol [48]	–1	+1	+1	+1	+1

defined as

$$F = \frac{\sigma_A^2}{\sigma_B^2}. \tag{3.18}$$

Fig. 3.9 shows the F-test results between nine compared metrics and the proposed method on the five databases. It can be observed that our proposed method produces either the lowest or comparable prediction errors among the leading sharpness metrics in LIVE, CSIQ, and TID2013 databases. In addition, the prediction errors of our proposed method are smaller than all the compared blind sharpness metrics on the two real blurred image databases, which are more favorable in real-world imaging environments.

The statistical significance of the proposed method against the compared metrics are obtained by employing a threshold $F_{critical}$. $F_{critical}$ is determined by the number of prediction errors and a confidence level. If $F > F_{critical}$ (or $F < F_{critical}$), the compared metric performs worse (or better) than our proposed method in terms of statistical significance. Otherwise, their performance are comparable. A 95% confidence level was utilized to determine the threshold $F_{critical}$. Table 3.2 summarizes the

Table 3.3: Performance comparison with general-purpose BIQA metrics. The best result is marked in boldface.

	LIVE		CSIQ		TID2013		BID		CID2013	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BIQI	0.9201	0.9137	0.8463	0.7734	0.8248	0.8153	0.6039	0.5724	0.7769	0.7438
DIIVINE	0.9431	0.9355	0.8863	0.8790	0.8472	0.8417	0.5064	0.4885	0.4987	0.4766
BLIINDS-II	0.9387	0.9309	0.8864	0.8917	0.8571	0.8624	0.5583	0.5302	0.7314	0.7012
BRISQUE	0.9512	0.9430	0.9211	0.9070	0.8622	0.8608	0.6120	0.5903	0.7144	0.6823
QAC	0.9155	0.9028	0.8313	0.8306	0.8412	0.8417	0.3208	0.3177	0.1866	0.1624
NIQE	0.9387	0.9302	0.9184	0.8914	0.8155	0.8073	0.4713	0.4692	0.6930	0.6328
SSEQ	0.9468	0.9347	0.8706	0.8702	0.8628	0.8621	0.6044	0.5809	0.6891	0.6755
ILNIQE	0.9209	0.9153	0.8679	0.8578	0.8253	0.8142	0.5184	0.4867	0.6758	0.6679
Proposed	**0.9581**	**0.9547**	**0.9516**	**0.9233**	**0.9569**	**0.9409**	**0.6327**	**0.6032**	**0.8803**	**0.8739**

statistical significance results, where "+1", "0", and "-1" denote that our proposed method is significantly better, competitive, and worse than the compared metric respectively. It can be observed from Table 3.2 that our proposed method is superior to the compared metrics in most cases. In LIVE, only HVS-MaxPol [48] performs better than our proposed method. Among the 45 combinations of metrics and databases, our metric performs significantly better in 38 cases. This indicates that the proposed method performs consistently well across all databases.

Since our proposed method is designed specifically for image sharpness evaluation, it is anticipated to outperform the general-purpose BIQA metrics. We compared our proposed method with eight leading general-purpose BIQA metrics, including BIQI [58], DIIVINE [59], BLIINDS-II [60], BRISQUE [61], QAC [68], NIQE [66], SSEQ [63], and ILNIQE [67]. We employed the same train-test procedure as described above. For those training-free metrics such as NIQE [66] and QAC [68], they were evaluated on the corresponding testing subsets for fair comparison. Table 3.3 lists the experimental results on the five databases, where the best result is marked in bold

Table 3.4: Statistical performance between the proposed method and general-purpose BIQA metrics on five databases.

Metric	LIVE	CSIQ	TID2013	BID	CID2013
BIQI [58]	+1	+1	+1	+1	+1
DIIVINE [59]	+1	+1	+1	+1	+1
BLIINDS-II [60]	+1	+1	+1	+1	+1
BRISQUE [61]	0	+1	+1	+1	+1
QAC [68]	+1	+1	+1	+1	+1
NIQE [66]	+1	+1	+1	+1	+1
SSEQ [63]	+1	+1	+1	+1	+1
ILNIQE [67]	+1	+1	+1	+1	+1

for each performance criterion.

From the results shown in Table 3.3, we can see that our proposed method outperforms all the general-purpose BIQA metrics on the five databases. Particulary, our proposed method has a performance gain of about 14% ↑ and 10% ↑ in terms of prediction accuracy over the other metrics on CID2013 and TID2013 respectively. Similarly, we also conducted the F-test to analyze the statistical significance of our proposed method against the compared general-purpose BIQA metrics. Fig. 3.10 provides the F-test results, and the corresponding statistical significance results are listed in Table 3.4. It can be observed from Fig. 3.10 that our proposed method produces the lowest prediction errors in CSIQ, TID2013, BID, and CID2013 databases. Furthermore, we can see from Table 3.4 that only BRISQUE [61] is statistically competitive to our proposed method in the LIVE database. Our proposed metric performs statistically better than all the compared metrics in the other four databases. From these results, it can be concluded that the proposed method achieves the best overall statistical performance.

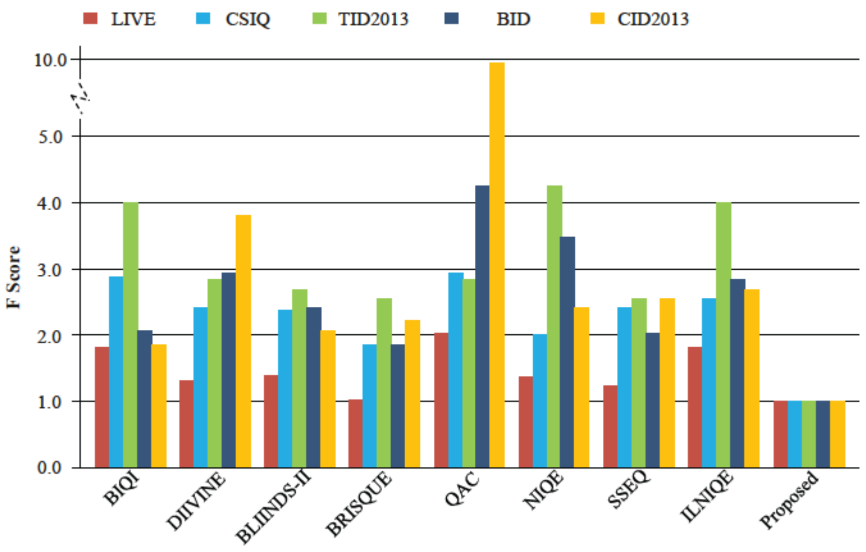


Figure 3.10: F-test results of general-purpose quality metrics against the proposed method.

Table 3.5: Cross-database performance evaluation of our proposed method.

	Testing Database					
	LIVE		CSIQ		TID2013	
Training Database	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
LIVE	–	–	0.9524	0.9460	0.9416	0.9352
CSIQ	0.9374	0.9211	–	–	0.9053	0.8951
TID2013	0.8367	0.8259	0.9178	0.9126	–	–
BID	0.9237	0.9043	0.9017	0.8913	0.8088	0.7857
CID2013	0.9310	0.9255	0.8791	0.8432	0.9015	0.8924

3.3.2 Method Analysis

Since our proposed method is based on feature learning, we conducted cross-database performance evaluation to further test its generalization capability. Specifically, for the three synthetically blurred image databases, our model was trained in one of them and then it was tested on the other two databases. As introduced in the beginning of this section, real blur is much more complex than synthetic blur in terms of formation, it is thus impractical to train a model with synthetic blur and then test it with real blur. Hence, for real blurred image databases, we trained our model in BID or CID2013 and then the trained model was employed for the sharpness evaluation on the three synthetically blurred image databases. Table 3.5 shows the cross-database performance evaluations of our proposed method. It can be observed that for the three synthetically blurred image databases, our proposed method produces quite promising results. The obtained PLCC and SRCC values are greater than 0.90 on most of the testing cases. Particularly, near state-of-the-art performance are achieved when the model is trained in the LIVE database. In addition, when our model is trained on the two real blurred image databases BID and CID2013, the prediction results are still very encouraging with most of the obtained PLCC values are larger than 0.87. Table 3.5, together with Table 3.1 and Table 3.3, indicate that our method not only achieves very good overall performance, but also shows good generalization capability.

It has been reported that adding noise may increase or decrease the sharpness of an image [38]. In this experiment, we investigated the impact of contaminated noise on the performance of our proposed method. Specifically, we tested our method on the LIVEMD database [98]. As mentioned in the beginning of this section, the LIVEMD database consists of 450 images corrupted under two multiple distortion scenarios: 1) 225 images are first blurred and then compressed by a JPEG encoder; 2) 225 images are first blurred and then corrupted by white Gaussian noise. For simplicity, the two sub-databases are denoted as LIVEMD1 (blur + JPEG) and LIVEMD2 (blur + Gaussian noise) respectively. Note that images in LIVEMD1 and LIVEMD2 were obtained by sequentially adding distortions to pristine images. Table 3.6 lists

Table 3.6: Performance comparison of our proposed method and other blind sharpness metrics on the LIVEMD database [98]. The best result is marked in boldface.

	LIVEMD1		LIVEMD2	
Metric	PLCC	SRCC	PLCC	SRCC
Marziliano [34]	0.4725	0.4235	0.3188	0.1719
JNB [35]	0.8034	0.7762	0.1934	0.0453
CPBD [36]	0.5241	0.4398	0.3215	0.2311
FISH [37]	0.8205	0.7324	0.2403	0.1908
S3 [38]	0.7329	0.5781	0.3877	0.2798
SVC [39]	0.8129	0.6488	0.2403	0.1310
LPC-SI [40]	0.8207	0.6954	0.5788	0.4329
MLV [41]	0.8466	0.7972	0.5738	0.4912
ARISM [42]	0.9084	0.8732	0.3785	0.5544
BIBLE [43]	0.8879	0.8367	0.3147	0.2386
SPARISH [33]	0.9039	0.8814	0.3217	0.5168
RISE [44]	0.9107	0.8870	0.8759	0.8603
MVGV [45]	0.8223	0.7464	0.5635	0.3169
Proposed	**0.9188**	**0.8853**	**0.9176**	**0.8954**

the experimental results of our proposed method and the compared blind sharpness metrics, where the best result for each performance criterion is marked in boldface.

It can be seen from Table 3.6 that our proposed method outperforms all the compared blind sharpness metrics on the LIVEMD database, regardless of prediction accuracy (PLCC) and monotonicity (SRCC). Specifically, our method achieves PLCC values above 0.91 on both the LIVEMD1 and LIVEMD2 sub-databases, which reveals that our method is robust to JPEG and Gaussian noises.

We further investigated how does the number of training images affect the performance of our proposed method. In the experiment, we used different proportions of images for the model training, namely, 80%, 70%, 60%, 50%, and 40% respectively. The median values of a thousand train-test procedures were reported. The experimental results are summarized in Table 3.7. From Table 3.7 we can see that

Table 3.7: Performance of the proposed metric with different percentages of images used in the training.

Traning-test	Criterion	LIVE	CSIQ	TID2013	BID	CID2013
80%-20%	PLCC	0.9581	0.9516	0.9569	0.6327	0.8803
	SRCC	0.9547	0.9223	0.9406	0.6032	0.8739
70%-30%	PLCC	0.9527	0.9490	0.9510	0.6113	0.8714
	SRCC	0.9542	0.9306	0.9385	0.5729	0.8673
60%-40%	PLCC	0.9455	0.9457	0.9437	0.5892	0.8677
	SRCC	0.9431	0.9279	0.9353	0.5362	0.8609
50%-50%	PLCC	0.9422	0.9412	0.9406	0.5574	0.8624
	SRCC	0.9407	0.9128	0.9317	0.5238	0.8588
40%-60%	PLCC	0.9358	0.9376	0.9365	0.5164	0.8526
	SRCC	0.9387	0.9054	0.9279	0.5032	0.8503

Table 3.8: Individual contributions of the two categories of features.

Database	Criterion	Spatial domain	Wavelet domain	All
LIVE	PLCC	0.9368	0.9492	0.9581
	SRCC	0.9275	0.9363	0.9547
CSIQ	PLCC	0.9371	0.9381	0.9516
	SRCC	0.9133	0.8861	0.9223
TID2013	PLCC	0.9458	0.9366	0.9569
	SRCC	0.9377	0.9312	0.9406
BID	PLCC	0.5274	0.5855	0.6327
	SRCC	0.4838	0.5345	0.6032
CID2013	PLCC	0.7424	0.8625	0.8803
	SRCC	0.7166	0.8551	0.8739

the performance of our proposed method drops slightly with the decrease of training images. Particulary, our proposed method achieves PLCC values above 0.93 on the three synthetically blurred image databases and above 0.85 on the CID2013 databases even if only 40% of images are used for training, which is quite impressive compared to the performance of other metrics reported in Table 3.1. The results shown in this table further demonstrate the robustness of our proposed method.

The proposed method involves two categories of features, including spatial-domain features and wavelet-domain features. To investigate the individual contribution of each type of feature, we conducted model training and testing using the two categories of features separately on the five databases. The train-test procedure on the databases was set as the same as introduced before. Table 3.8 summarizes the experimental results.

It can be seen from Table 3.8 that each category of features achieves good performance for synthetic blur, with the prediction accuracy (PLCC value) greater than 0.93. Nevertheless, neither of them can achieve the best results in all the three synthetically blurred image databases. For real blur, features extracted in the wavelet domain contribute more than spatial-domain features. It can be further observed that the integration of two categories of features achieves much better performance than using a single category of features on all the five databases, which indicates the importance of using features derived from both domains so as to achieve consistently good performance for sharpness evaluation.

The proposed method involves more than one hundred of features (adding up those described in Section 3.2). It would be meaningful to use dimensionally reduction techniques to limit the number of required features. We conducted an experiment which employed principal component analysis (PCA) to reduce features' redundancy. We found that our proposed method achieved comparable performance even if only fifty features were utilized in the model training. In addition, we further investigated the effectiveness of the discrepancy measures by directly using the parameters of band-pass responses (see Eq. 3.16) computed from the original input images in the model training. In the experiment, only features constructed in the wavelet domain were involved. We found that this setting produces PLCC value of 0.9188 on the TID2013 database, which is inferior to the results reported in Table 3.8. The experimental results show that additional performance gain is achieved by utilizing the discrepancy measures in our method.

In this work, we propose to evaluate image sharpness utilizing the discrepancy between an input image and its “reblurred” version which is obtained by employing a Gaussian filter. Here we investigate how does the reblurring parameter, i.e., the window size of Gaussian filter, affects the performance of our proposed method. In the experiment, we used different window sizes of Gaussian filter, and only features constructed through the reblurred version were involved in the model training. The experimental results on the three synthetically blurred image databases are summarized in Table 3.9. From Table 3.9 we can see that the best performance are achieved when window size of 3×3 is used. This setting of reblurring parameter was applied to the two real blurred image databases BID and CID2013, and from the experimental results in Table 3.1 we know that our proposed method achieves the state-of-the-art performance. However, we found that this parameter setting is not optimal for high resolution (e.g., 4K or 8K) images, considering that pixel values within a 3×3 window might be the same. Note that the highest resolution of the examined five databases is 1600×1200, thus larger window size is preferred to apply the proposed method to higher resolution images.

In addition, we investigated how does the parameters (s, o) affect the performance of our proposed method, where s and o denote scale index and orientation index respectively as described in Eq. 3.12. In the experiment, only features constructed in the wavelet domain were involved in the model training. The experimental results are summarized in Table 3.10. From Table 3.10 we can see that the best performance

Table 3.9: Impact of reblurring parameter on the performance of the proposed metric.

Kernel Size	Criterion	LIVE	CSIQ	TID2013
3×3	PLCC	0.9502	0.9419	0.9422
	SRCC	0.9413	0.9151	0.9208
5×5	PLCC	0.9453	0.9377	0.9352
	SRCC	0.9377	0.9002	0.9193
9×9	PLCC	0.9386	0.9201	0.9263
	SRCC	0.9264	0.8988	0.9095

Table 3.10: Impact of parameters (s, o) on the performance of the proposed metric.

Parameters (s, o)	Criterion	LIVE	CSIQ	TID2013
(2, 6)	PLCC	0.9125	0.8953	0.9177
	SRCC	0.9008	0.8576	0.9053
(2, 10)	PLCC	0.9322	0.9136	0.9219
	SRCC	0.9288	0.8701	0.9188
(3, 10)	PLCC	0.9492	0.9381	0.9366
	SRCC	0.9363	0.8861	0.9312
(4, 10)	PLCC	0.9501	0.9399	0.9407
	SRCC	0.9376	0.8902	0.9329

are achieved when $(4, 10)$ is employed. However, the performance with $(4, 10)$ are only slightly better than the case with $(3, 10)$. To balance between performance and computational overhead, $(3, 10)$ is chosen in this work.

We further evaluated the computational cost of our proposed method. The experiments were conducted on a PC with an Intel Xeon E5540 (four cores @ 2.53 GHz) and 12 GB RAM, running on Matlab R2017a. The compared methods include both sharpness and general-purpose BIQA metrics that presented in Table 3.1 and Table 3.3. The time cost consumed by each metric for evaluating the quality of a 512×512 image in the CSIQ database is shown in Table 3.11. It can be seen that the proposed method has a moderate computational complexity.

Table 3.11: Computational cost of each metric.

Metric	Time (s)	Metric	Time (s)	Metric	Time (s)
Marziliano [34]	0.30	ARISM [42]	21.28	BRISQUE [61]	1.11
JNB [35]	0.51	BIBLE [43]	2.13	QAC [68]	0.54
CPBD [36]	0.56	SPARISH [33]	2.95	NIQE [66]	0.98
FISH [37]	0.67	RISE [44]	0.32	SSEQ [63]	2.79
S3 [38]	15.22	MVGV [45]	0.03	ILNIQE [67]	14.28
SVC [39]	0.23	BIQI [58]	2.76	Proposed	1.56
LPC-SI [40]	1.23	DIIVINE [59]	28.53		
MLV [41]	0.08	BLIINDS-II [60]	54.38		

3.4 Summary

This chapter presents a novel blind image sharpness metric based on the observation that the discrepancy between an image and its “reblurred” version usually indicates the extent of blur in the image. Specifically, the global sharpness discrepancy is measured through inter-resolution self-similarities, while the local structural degradation of an image is characterized by the discrepancies of orientation selectivity-based visual patterns and log-Gabor filter responses between the image and its corresponding reblurred version. A regression module is employed to map the extracted features into an overall quality score. Extensive experiments and comparisons are conducted on six public blurred image databases, including both synthetic and real blur. The experimental results have demonstrated that the proposed method consistently performs well across several databases and outperforms other available metrics on the real blurred image databases by a clear margin (prediction accuracy improved from 0.7832 to 0.8803 on CID2013 database).

Chapter 4

Blind Quality Assessment of Gamut-Mapped Images

The BIQA metric introduced in Chapter 3 deals with a conventional type of distortion, namely blur. This chapter presents a BIQA metric for an unusual distortion type – gamut mapping based on natural scene statistics. To the best of our knowledge, this is the first work that addresses the blind quality assessment of gamut-mapped images.

The method described in Chapter 3 resorts to a scheme that extracts effective features from distorted images first followed by training a regression module using those features. Particularly, subjectively rated scores are required in the training stage. In this chapter, the proposed metric does not need ground-truth quality scores for training, which means it is "completely blind".

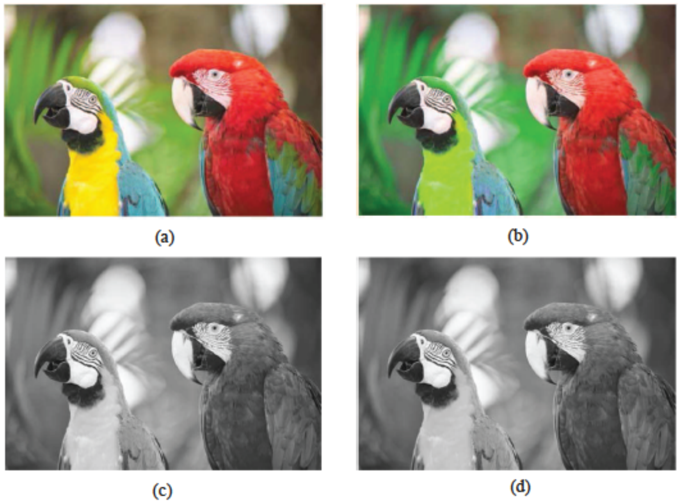


Figure 4.1: Distortion of hue change [49]. (a) and (b) are two color images where (a) is the original pristine image and (b) is its distortion counterpart. (c) and (d) are their corresponding grayscale versions. This figure shows that some distortions are hard to discriminate through the grayscale version of images.

4.1 Introduction

The rendering of a color image to device limitations, also called "gamut mapping", is often used as one of the primary control parameters for color reproduction [50, 10]. To quantify how reproduced images have been changed by the reproduction process and how much of these changes could be perceived by the human eye, a robust evaluation of the gamut-mapped images is highly needed [100].

During the past decade, a number of perceptually meaningful IQA models have been proposed while most of them do not place enough emphasis on color information. The prediction performance of some grayscale IQA models is impressive on most conventional distortions like compression, blur, or blockiness [45, 23]. However, distortions such as saturation change or gamut mapping may not be effectively detected by these grayscale IQA metrics [12]. Fig. 4.1 shows an example of this type of distortion. While the distortion can be seen clearly on Fig. 4.1 (b) (e.g., feathers of the left parrots), the grayscale versions (Fig. 4.1 (c) and Fig. 4.1 (d)) are hard to discriminate.

The main image quality factors for gamut mapping are preservation of spatial details and preservation of color [101]. It is worth noting that the statistic-based general-purpose BIQA methods mentioned in Section 2.2 often fail on the quality prediction of gamut-mapped images since it mostly involves color distortions. Motivated by this, this chapter presents a BIQA metric for gamut-mapped images. In the proposed method, images are first transformed into an opponent color space and then two categories of statistics are analyzed. In particular, the proposed metric does not need subjective quality scores for training. Quality predictions of gamut-mapped images are performed by quantifying their departure from the statistical regularities of natural undistorted images. We conducted experiments on three gamut mapping image databases to evaluate the quality evaluation performance of our method. Moreover, the proposed metric was further applied for benchmarking gamut mapping algorithms.

4.2 Proposed Method

According to the working mechanism of the human visual system (HVS), human eyes employ both global-to-local and local-to-global strategies for judging the quality of images with different extents of distortions [25]. Fig. 4.2 illustrates the flowchart of our proposed metric. It consists of two phases, a model training phase followed by

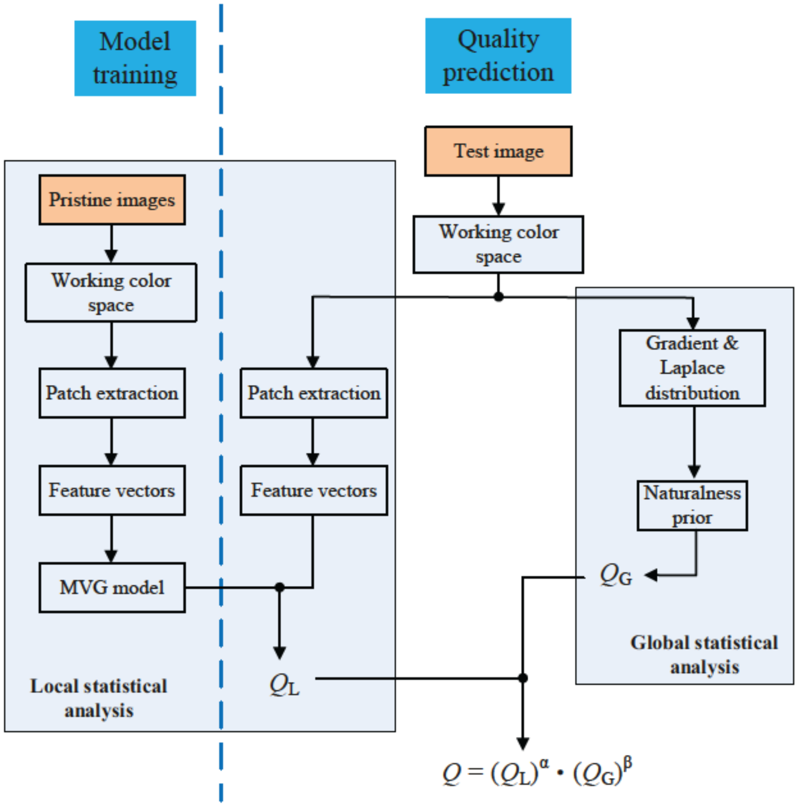


Figure 4.2: Flowchart of the proposed quality metric for gamut-mapped images.

a quality prediction phase. The images are first transformed into a working color space. In the model training phase, local statistical features are extracted from a set of pristine images and then combined to learn a multivariate Gaussian (MVG) model, which is used for the subsequent quality prediction of gamut-mapped images. The local statistical features are used to portray the structural and color distortions, while the global statistical features are utilized to characterize the loss of global naturalness. The extraction of these two types of features will be explained in detail in the following subsections.

4.2.1 Working Color Space

In digital imaging, RGB color model is widely-used to represent a color image. However, the RGB color model does not perform well with regard to perceptual uniformity, which means Euclidean distances in the space do not match perceived color differences [57].

Basically, perceptual color distortion ($D_{perceptual}$) can be computed as the mean squared error between a reference color image (I_r) and its distorted version (I_d)

$$D_{perceptual} = \frac{1}{N}||I_r - I_d||^2, \tag{4.1}$$

where N is the number of pixels in the image. In RGB color space, the set of equally distorted RGB vectors is not isotropic around the reference vector and the geometry of this set varies from different reference vectors [102]. This makes $D_{perceptual}$ fails to produce distortion values that are consistent with human perception. The same situation applies to other perceptually nonuniform color models such as YCbCr. In order to quantify perceptual color distortions effectively, distorted images are first transformed into a perceptually uniform CIELAB color space [103] which is separated into a lightness channel "L", a red-green channel "A", and a blue-yellow channel "B". The CIELAB color space corresponds to the human color perception better than the perceptually nonuniform color models and avoids cross contamination between the color attributes [49].

4.2.2 Local Statistics

Note that local statistical features extracted from image patches can effectively capture the essential statistics of natural images [66]. In this work, a number of statistical features are adopted to accomplish the quality prediction of gamut-mapped images. Specifically, features extracted from local mean subtracted and contrast normalized (MSCN) coefficients are utilized to characterize local structural distortion. On the other hand, color distortions are captured by features derived from color responses.

According to [61], characterizing normalized luminance coefficients is useful to quantify quality in the presence of distortion. It is worth noting that these coefficients follow a Gaussian model [95]. The coefficients normalization can be defined as

$$\bar{I}(i,j) = \frac{I(i,j) - \mu(i,j)}{\sigma(i,j) + C}, \tag{4.2}$$

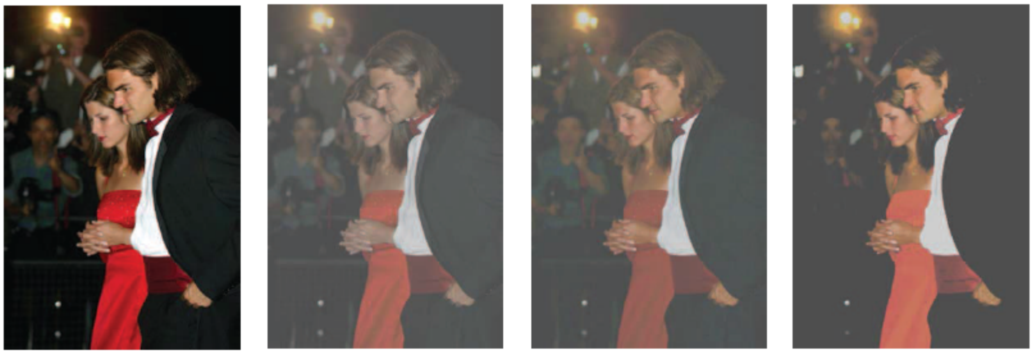
where $i \in 1, 2, \cdot\cdot\cdot, M$ and $j \in 1, 2, \cdot\cdot\cdot, N$ are coordinates (M and N are the image height and width respectively). I denotes the lightness channel “L” in the CIELAB color space. The constant $C = 1$ is used to prevent instabilities when the denominator tends to zero. $\sigma(i,j)$ and $\mu(i,j)$ are the local image contrast and mean, which can be computed as

$$\sigma(i,j) = \sqrt{\sum_{k=-K}^{K} \sum_{h=-H}^{H} \omega_{k,h} \left[I(i+k, j+h) - \mu(i,j) \right]^2}, \tag{4.3}$$

$$\mu(i,j) = \sum_{k=-K}^{K} \sum_{h=-H}^{H} \omega_{k,h} I(i+k, j+h), \tag{4.4}$$

where $\omega = \{\omega_{k,l} | k = -K, \cdot\cdot\cdot, K, h = -H, \cdot\cdot\cdot, H\}$ is a 2D circularly-symmetric Gaussian weighting function that re-scaled to unit volume.

Note that the MSCN coefficients $\bar{I}(i,j)$ conform to a Gaussian distribution on high-quality images [95]. On the other hand, the Gaussian model can be affected by the presence of distortion and quantifying the deviation will make it possible to predict perceptual quality. Specifically, the distribution of $\{\bar{I}(i,j)\}$ is modeled by a generalized Gaussian distribution (GGD) as described in Eq. 3.13. Fig. 4.3 shows



(a) Pristine image (b) MOS=0.5383 (c) MOS=0.9789 (d) MOS=1.0539

Figure 4.3: Gamut-mapped images and their subjective MOS scores. (a) is the reference pristine image, and (b) – (d) are gamut-mapped images generated by three different GMAs respectively (i.e., SGCK, Scomp, and HPMinDE [56]).

three gamut-mapped images and their subjective ratings. As shown in Fig. 4.4, pristine and distorted images show distributions differ in shape and scale, in which from Fig. 4.3 (b) to Fig. 4.3 (d) the distributions become more heavy-tailed and center-peaked. As a result, after modeling these distributions, the model parameters can be used as perceptual features for quality assessment.

In addition, the pairwise products of adjacent normalized MSCN coefficients [66] can also be used to measure perceptual quality, especially along four orientations: horizontal, vertical, main-diagonal, and secondary-diagonal. These products of neighboring coefficients are well modeled as a zero-mean Asymmetric GGD (AGGD). The density function is described as in Lasmar *et al.* [104]

$$g_a(x;\gamma,\delta_l,\delta_r) = \begin{cases} \frac{\gamma}{(\delta_l+\delta_r)\Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\left(\frac{-x}{\delta_l}\right)^{\gamma}\right), & \forall x \le 0 \\ \frac{\gamma}{(\delta_l+\delta_r)\Gamma\left(\frac{1}{\gamma}\right)} \exp\left(-\left(\frac{x}{\delta_r}\right)^{\gamma}\right), & \forall x > 0 \end{cases} \tag{4.5}$$

where the parameter γ denotes the distribution shape, while δ_l and δ_r are scale pa-

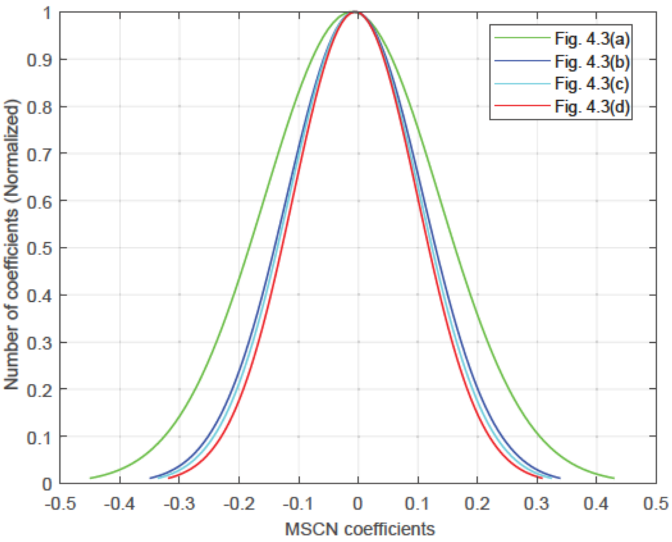


Figure 4.4: Histogram distribution of MSCN coefficients for images in Fig. 4.3.

rameters that control the spread on each side of the model. If $\delta_l = \delta_r$, then the AGGD reduces to the GGD. The mean of AGGD can also be used as a feature parameter, which is computed as

$$\eta = (\delta_r - \delta_l)\frac{\Gamma\left(\frac{2}{\gamma}\right)}{\Gamma\left(\frac{1}{\gamma}\right)}. \tag{4.6}$$

The parameters $(\gamma, \delta_l, \delta_r, \eta)$ are estimated through the moment-matching method in Sharifi *et al.* [96]. For each paired product, sixteen parameters (four parameters for each orientation) are computed, yielding the next set of perceptual features.

As pointed out in Section 2.1.2, gamut mapping largely incorporates color distortions. The statistical properties of color distortions in gamut-mapped images are then investigated. In Ruderman *et al.* [105], studies explained that the statistics of color response in natural image follow a univariate Gaussian distribution in an opponent color space. In this work, the gamut-mapped images are first transformed into the CIELAB color space which has three channels (L, A, B). The following function is

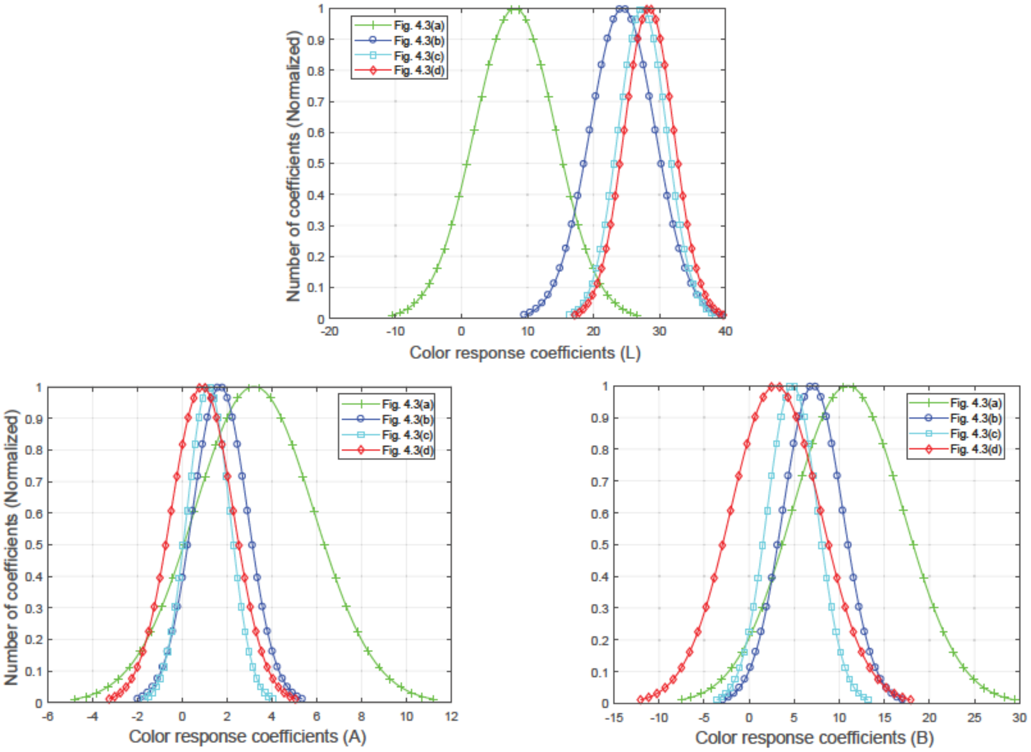


Figure 4.5: Histogram distribution of (L, A, B) coefficients for images in Fig. 4.3.

employed to fit the probability density of image data

$$f(x;\zeta,\rho^2) = \frac{1}{\sqrt{2\pi}\rho}\exp\left(\frac{-(x-\zeta)^2}{2\rho^2}\right), \tag{4.7}$$

where ζ and ρ^2 are the model parameters.

For each of the channels, ζ and ρ^2 are estimated and taken as quality features, yield another six features. Fig. 4.5 plots the histogram of (L, A, B) coefficients for images in Fig. 4.3. We can notice that in the case of color distortions, the distribution "forms" of (L, A, B) are clearly changed. This indicates that these model parameters can be used as quality-aware features to predict image quality.

To incorporate multi-scale behavior, images are down-sampled by a factor of two. All of the local statistical features are computed at two scales, yielding a set of forty-eight features. These features are then combined to learn an MVG model for the subsequent local statistic quality prediction of gamut-mapped images. The model training will be further explained in detail in Section 4.2.4.

4.2.3 Global Statistics

Unlike conventional distortions that operate locally in most cases, gamut mapping also incorporates global distortions. Many studies have been dedicated to global naturalness which has important significance to both image processing applications and the understanding of biological vision [106]. It is generally believed that natural image of high visual quality has a high degree of naturalness [107]. Inspired by this, the proposed IQA model takes into account a statistical naturalness model developed upon the naturalness prior. The naturalness prior is a linear combination of a gradient distribution prior and the consistent Laplace operator prior (i.e., the one corresponding to the divergence of the gradient).

It has been proved that the gradient distribution prior is closely related to image quality and is quite stable [108]. Specifically, the HVS mainly detects gradient information for processing and the neurons have been evolved to be adapted to the environment based on this information. In addition, different people have almost the same visual perception, so the gradient distribution of natural scene images is stable [14]. These two properties of gradient distribution prior naturally fit into the requirements of IQA.

The gradient field of an image is equivalent to the original image adding a single point constraint [109]. Given an intensity image $I(x, y)$, the gradient field $\mathbf{G}$ can be defined as

$$\mathbf{G} = (G^x, G^y) = (\nabla_x I(x,y), \nabla_y I(x,y)), \tag{4.8}$$

where G^x and G^y are the gradients in x and y directions respectively. ∇_x and ∇_y denote the finite-difference approximations, which are defined as in Eq. 3.7. It is easy to know that $G^x, G^y \in [-255, 255]$. Since G^x and G^y both satisfy the heavy-tail characteristic in log-scale, they are commonly modeled as a Gaussian or Laplacian distribution [110].

In this work, the normalized gradient histograms of G^x and G^y are modeled based on the cumulative distribution function (CDF), which is computed as

$$C(\mathbf{G}) = \int_{-255}^{G^x} \int_{-255}^{G^y} P(u,v) du dv, \tag{4.9}$$

where $P(\cdot)$ represents the probability density. Particularly, the CDF is approximated through the Cauchy distribution as in Gong *et al.* [109]

$$\widetilde{C}(\mathbf{G}) = (\frac{\text{atan}(T_1 G^x))}{\pi} + \frac{1}{2})(\frac{\text{atan}(T_1 G^y))}{\pi} + \frac{1}{2}), \tag{4.10}$$

where T_1 is the model fitting parameter.

In addition to gradient statistics, studies have shown that Laplace prior is also very powerful for image processing [111]. The Laplace field $\mathbf{L}$ is described as

$$\mathbf{L} = L(x,y) = \Delta I(x,y), \tag{4.11}$$

where Δ is the Laplace operator. $L(x,y)$ is discretized by the five-point stencil, and $L(x,y) \in [-255 \times 4, 255 \times 4]$. Similar to the gradient CDF, Laplace CDF is used to model the distribution of the Laplace operator response, which is defined as

$$L(t) = \int_{-\infty}^{t} P(\Delta I(x)) d\Delta I(x). \tag{4.12}$$

Table 4.1: Estimated global statistical parameters for the gamut-mapped images shown in Fig. 4.3.

Parameter	Fig. 4.3(b)	Fig. 4.3(c)	Fig. 4.3(d)
MOS	0.5383	0.9789	1.0539
T_1	2.2326	2.4462	2.5149
T_2	3.3120	4.4140	6.9153
N_f	2.7723	3.4301	4.7151

To approximate the Laplace CDF, the following parametric model is utilized

$$\widetilde{L}(t) = \frac{\mathrm{atan}(T_2 t))}{\pi} + \frac{1}{2}, \tag{4.13}$$

where T_2 is the model fitting parameter.

The gradient and Laplace distributions are both of great importance for characterizing natural images [14]. The model parameters T_1 and T_2 are combined to define the image naturalness factor N_f as

$$N_f = \frac{1}{2}\left(\frac{T_1}{T_1^{pr}} + \frac{T_2}{T_2^{pr}}\right), \tag{4.14}$$

where T_1^{pr} and T_2^{pr} are the gradient distribution and Laplace distribution priors respectively, which are obtained based on the aggregation results on seven datasets of natural scene images [109]. In this work, $T_1^{pr} = 0.380$ and $T_2^{pr} = 0.145$. The disparity between the gradient/Laplace distributions of a test image and the prior distributions represents the naturalness extent of the test image. For high-quality natural images, the N_f value should be close to 1.

Note that Fig. 4.3 shows three gamut-mapped images and their subjective MOS scores. Table 4.1 lists the naturalness factors N_f of these images and their associated statistical model parameters T_1, T_2. From the table we can see that T_1 and T_2 increase monotonically with the decreased qualities from Fig. 4.3 (b) to Fig. 4.3 (d).

Moreover, the values of N_f are becoming more and more larger than 1, which indicate that the images from Fig. 4.3 (b) to Fig. 4.3 (d) are perceived with degraded quality.

4.2.4 Model Training and Quality Prediction

The proposed IQA metric computes quality scores based on both local and global statistics. Since the naturalness level is an indication of global distortion severity, we take the naturalness factor N_f as the global statistic score Q_{G}, which means $Q_{\mathrm{G}} = N_f$. To compute the local statistic quality score, an MVG model of the local statistical features was first learned from a set of pristine images. The MVG model can be computed by fitting features from patches with an MVG density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}}(|\Sigma|)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\nu)^T\Sigma^{-1}(\mathbf{x}-\nu)\right), \tag{4.15}$$

where $\mathbf{x} = (x_1, \cdot\cdot\cdot, x_m)$ is the feature vector. ν and Σ are the mean and covariance matrix of the MVG model respectively. The values of ν and Σ were estimated using a standard maximum likelihood estimation technique [112]. Note that an MVG model is fully described by the pair (ν, Σ).

A set of original pristine color images in the AlgMix databases [113] were collected to learn the pristine MVG model, which do not overlap with those discussed in Table 4.2. The pristine images were first divided into patches and then features described in Section 4.2.2 were extracted from these patches. The patch size is $p \times p$. We selected patches whose contrast (see Eq. 4.3) are bigger than a threshold θ to get more meaningful feature extraction. The local statistic quality of the distorted image is expressed as the distance between the learned pristine MVG model and the distorted image's MVG model which fitted to the features extracted from the distorted image. The local statistic quality score Q_{L} is computed referring to the Bhattacharyya distance [114] as

$$Q_{\mathrm{L}} = \sqrt{\left((\nu_0 - \nu')^T \left(\frac{\Sigma_0 + \Sigma'}{2}\right)^{-1} (\nu_0 - \nu')\right)}, \tag{4.16}$$

where $(\nu_0,\, \Sigma_0)$ and $(\nu',\, \Sigma')$ denote the learned pristine MVG model and the distorted image's MVG model respectively.

So far, we have obtained the local statistic quality score Q_{L} and global statistic quality score Q_{G}. The overall quality score is then calculated as

$$Q = (Q_{\mathrm{L}})^{\alpha} \cdot (Q_{\mathrm{G}})^{\beta}, \tag{4.17}$$

where parameters $\alpha, \beta \in [0, 1]$ are used to balance the relative contributions of different components.

4.3 Experimental Results and Analysis

The performance of the proposed IQA metric was evaluated based on three gamut mapping databases, namely Image Gamut [115], Basic Study [101], and Local Contrast [116]. We refer to these three databases by their initials simply as IG, BS, and LC. MOS values are provided as the ground truth of image quality, which were derived from a choice distribution model Thurstone's Law [117] with the raw data (i.e., the observers' paired comparison choices). The databases are detailed in Table 4.2. In the IG database, algorithms that either uses a linear or sigmoidal mapping have three possible source gamuts, then these six combinations were compared to HPminDE and SGCK [56], resulting in eight GMAs. More description about these databases can be found in Barańczuk *et al.* [101].

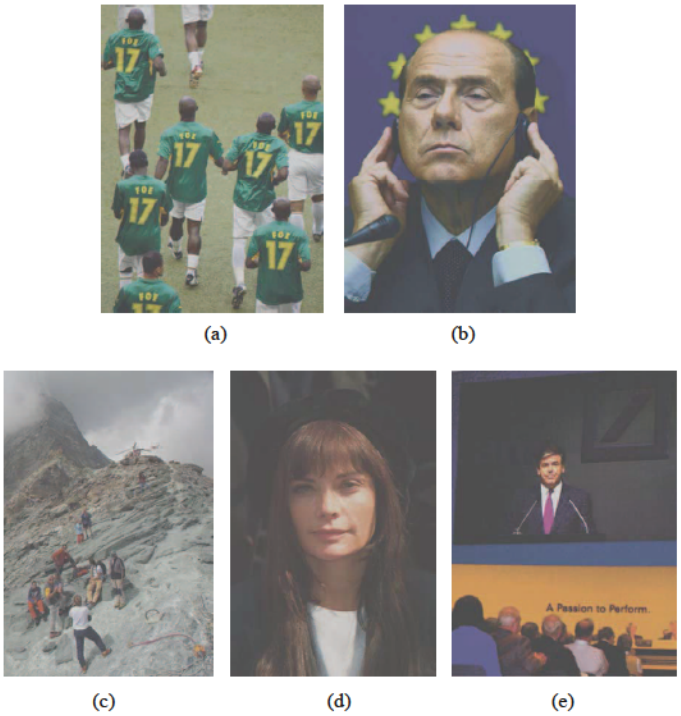
Table 4.2: Gamut mapping image databases.

Database	Reference images	Distorted images	Subjects	GMAs
IG	65	520	42	8
BS	97	1067	9–12	11
LC	72	576	21	8

PLCC and SRCC were employed for evaluating the performance of image quality metrics as defined in Eq. 1.3 and Eq. 1.4. In our implementation, the patch size p was set to 96 as suggested in Mittal *et al.* [66]. Stable performance was observed across p ranging from 64×64 to 128×128 in the training of the MVG model. The threshold θ was computed as 75% of the image's maximum patch contrast empirically. There are two free parameters α and β in Eq. 4.17, which were determined based on the IG database. Specifically, these parameters were first tuned by maximizing the SRCC and PLCC values in the IG database, and then they were used in the performance evaluation of the other two databases, namely BS and LC. In this work, we set $\alpha = 0.72$, $\beta = 0.32$, which can achieve consistently good performance across all databases.

4.3.1 Performance Comparison

The performance of the proposed metric was compared with several existing general-purpose BIQA metrics, namely BIQI [58], DIIVINE [59], BLIINDS-II [60], BRISQUE [61], DESIQUE [62], GL-BIQA [64], NFERM [65], NIQE [66], ILNIQE [67], and QAC [68]. Note that a majority of these general-purpose BIQA metrics require subjective quality scores to calibrate the learned support vector regression (SVR) module [85]. We implemented the SVR as described in Section 3.2.4. To get fair comparison, for



(a) (b)

(c) (d) (e)

Figure 4.6: Five images with different extents of gamut distortion. (a) MOS=0.1750. (b) MOS=0.2984. (c) MOS=0.4906. (d) MOS=0.8107. (e) MOS=1.1327.

metrics like NIQE [66] and the proposed metric that do not need subjective scores to train a model, the performance were evaluated on the corresponding testing subsets. Likewise, we employed the train-test procedure that introduced in Section 3.3.1.

To show how the proposed metric performs on real images, we tested it using several gamut-mapped images. Fig. 4.6 shows five images in the IG database. They have different extents of gamut distortion and their subjective qualities are indicated by the MOS values. From Fig. 4.6(a)–(e), the MOS values are monotonically increased. From this perspective, a good quality metric should produce monotonically decreased/increased scores. Table 4.3 summarizes the quality scores predicted by different metrics. It is observed from the table that the proposed method can produce

Table 4.3: Quality scores predicted by different metrics on the images shown in Fig. 4.6.

Image	(a)	(b)	(c)	(d)	(e)
MOS	0.1750	0.2984	0.4906	0.8107	1.1327
BIQI [58]	0.3252	2.5993	0.2448	1.7573	0.9508
DIIVINE [59]	2.2832	1.4810	2.9830	4.3668	3.2062
BLIINDS-II [60]	1.0178	2.2231	1.3299	2.3807	0.8861
BRISQUE [61]	1.4116	0.2731	1.0728	0.5091	0.8992
DESIQUE [62]	2.1690	0.4767	2.4472	0.6817	1.4283
GL-BIQA [64]	1.2103	1.3521	1.3068	0.9956	1.9890
NFERM [65]	1.9165	1.7856	2.1444	1.0183	2.0063
NIQE [66]	3.6236	3.1368	4.3637	4.7277	5.1013
ILNIQE [67]	25.6355	29.1973	22.1943	18.5658	19.0343
QAC [68]	0.6322	0.6480	0.7747	0.6984	0.7543
Proposed	**2.3151**	**3.5770**	**4.3392**	**5.3979**	**6.7507**

Table 4.4: Performance comparison of the proposed metric and the leading general-purpose BIQA metrics on three gamut mapping databases.

	BS		IG		LC	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BIQI [58]	0.5937	0.5225	0.6098	0.5493	0.5861	0.5284
DIIVINE [59]	0.6538	0.5911	0.6926	0.6342	0.6387	0.5726
BLIINDS-II [60]	0.5865	0.5187	0.5880	0.5322	0.5990	0.5280
BRISQUE [61]	0.5958	0.5330	0.5779	0.5284	0.6262	0.5579
DESIQUE [62]	0.6554	0.5811	0.6435	0.5833	0.6264	0.5639
GL-BIQA [64]	0.6298	0.5684	0.6551	0.5994	0.6241	0.5624
NFERM [65]	0.6285	0.5509	0.5800	0.5230	0.5915	0.5266
NIQE [66]	0.5840	0.5280	0.5256	0.4520	0.6974	0.6288
ILNIQE [67]	0.5545	0.4849	0.5085	0.4520	0.5447	0.4940
QAC [68]	0.5143	0.4790	0.5241	0.3766	0.5634	0.4577
Proposed	**0.7865**	**0.7275**	**0.7464**	**0.7165**	**0.7573**	**0.7074**

monotonically increased quality scores for the five images, which are consistent with their subjective scores. By contrast, the quality scores produced by the compared metrics do not satisfy the monotonicity very well.

Table 4.4 summarizes the experimental results on the BS, IG, and LC databases. It is observed from Table 4.4 that the proposed metric outperforms all the compared

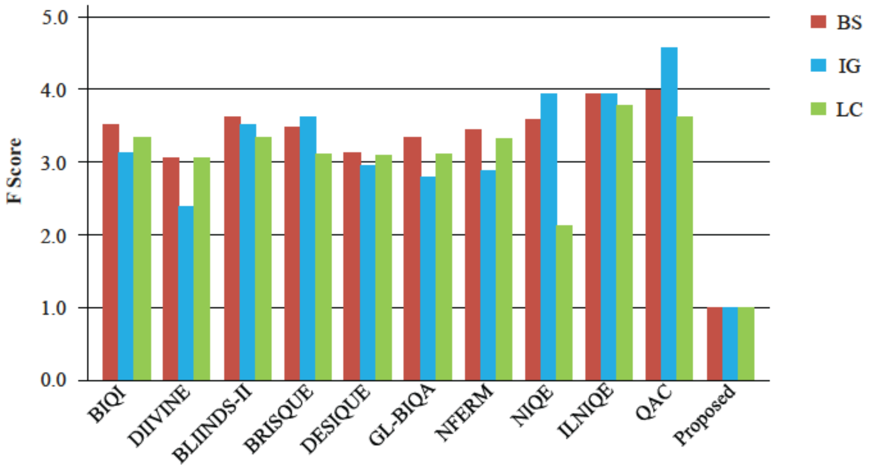


Figure 4.7: F-test results of the compared metrics against the proposed method.

Table 4.5: Summary of statistical performance between the proposed method and the compared metrics on three databases.

Metric	BS	IG	LC	Metric	BS	IG	LC
BIQI [58]	+1	+1	+1	GL-BIQA [64]	+1	+1	+1
DIIVINE [59]	+1	0	+1	NFERM [65]	+1	+1	+1
BLIINDS-II [60]	+1	+1	+1	NIQE [66]	+1	+1	0
BRISQUE [61]	+1	+1	+1	ILNIQE [67]	+1	+1	+1
DESIQUE [62]	+1	+1	+1	QAC [68]	+1	+1	+1

metrics by a clear margin. The general-purpose BIQA metrics can only produce quality scores that slightly correlate with subjective ratings. It appears that distortions of gamut mapping are indeed very different from conventional ones, only measuring grayscale of image is not sufficient for the quality assessment of gamut-mapped images.

The statistical significance of the proposed model was further analyzed using the F-test [12] as introduced in Section 3.3.1. It is observed from Fig. 4.7 that in the three gamut mapping databases (i.e., BS, IG, and LC), the proposed method produces the

Table 4.6: Performance of each feature category.

Database	Criterion	Local	Global	All
BS	PLCC	0.7465	0.7033	0.7865
	SRCC	0.6979	0.6575	0.7275
IG	PLCC	0.7269	0.6954	0.7464
	SRCC	0.6738	0.6565	0.7165
LC	PLCC	0.7273	0.7073	0.7573
	SRCC	0.6865	0.6474	0.7074

smallest prediction errors among all compared metrics. Table 4.5 lists the statistical significance between the proposed metric and ten compared metrics. In Table 4.5, it is clear that our method is superior to the existing metrics in most cases. Among the 30 cases, our metric performs significantly better in 28 cases and comparable in 2 cases. Specifically, the proposed metric outperforms all the compared metrics in the BS database. In the IG database, only DIIVINE is comparable to our method, and the proposed metric outperforms all other metrics. In LC, only NIQE performs comparably, and our method is superior to the remaining nine metrics. This indicates that the proposed method performs consistently well across all databases, which is desired in real applications.

4.3.2 Method Analysis

Two categories of statistical features are employed in the proposed metric, i.e., the local statistics and the global statistics. In order to understand the relative contributions of the two components, the performance of each feature category was evaluated separately on all the databases. The results are reported in Table 4.6. From the results shown in Table 4.6, we can see that both local statistical and global statistical features exhibit very good performance, which are already better than those of the

Table 4.7: Computational overhead of each metric.

Metric	Time (s)	Metric	Time (s)
BIQI [58]	2.39	GL-BIQA [64]	1.88
DIIVINE [59]	27.38	NFERM [65]	70.67
BLIINDS-II [60]	76.51	NIQE [66]	1.25
BRISQUE [61]	1.63	ILNIQE [67]	13.22
DESIQUE [62]	1.14	QAC [68]	0.38
Proposed	1.42		

compared metrics in Table 4.4. Moreover, much better results are obtained when they are used together. This indicates that the two categories of statistical features can work cooperatively for image quality evaluation.

We further compared the computational overhead of each competing BIQA metric. The experiments were conducted using the same platform as described in Section 3.3.2. The average runtime consumed by each metric for evaluating the quality of a 610 × 400 image in the IG database is shown in Table 4.7. We observe that NFERM and BLIINDS-II are the slowest, and QAC is the fastest. However, the quality prediction performance of QAC is worse than other metrics (see Fig. 4.7). It can be seen from Table 4.7 that the proposed model only needs less than 1.5 seconds to process an image, which exhibits relatively low computational complexity.

In the next experiment, the proposed quality metric was applied for benchmarking gamut mapping algorithms (GMAs). Since MOS scores are served as the ground truth of image quality, we first ranked the GMAs according to the MOS values. This MOS-based ranking was taken as the indicator of the relative performance. We then did the same rankings on different quality metrics with their predicted scores and finally the metric score-based rankings were compared with MOS-based ranking. For a good quality metric, its score-based ranking should be consistent with the MOS-based

Table 4.8: Performance rankings of eight GMAs based on the MOS values and predicted scores by different image quality metrics. Number "8" represents the worse performance and smaller number indicates better performance. "Statistics" shows the number of objective rankings that are consistent with the subjective rankings. Correct objective rankings are marked in boldface.

GMAs	MOS	BIQI	BRISQUE	NFERM	NIQE	ILNIQE	QAC	Proposed
		[58]	[61]	[65]	[66]	[67]	[68]	
Img00	1	6	2	3	3	4	3	**1**
Img01	2	4	1	**2**	2	5	1	**2**
Img10	3	3	5	4	1	1	2	4
Img11	4	1	**4**	1	**4**	3	8	3
LComp	5	**5**	7	6	7	2	**5**	**5**
SGCK	6	7	6	8	**6**	8	7	**6**
SComp	7	2	3	**7**	5	**7**	4	**7**
HPMinDE	8	**8**	**8**	5	**8**	6	6	**8**
Statistics	–	2	2	2	3	1	1	6

ranking [14]. Hence, in terms of benchmarking GMAs, the performance of different quality metrics can be easily determined through checking the number of consistent rankings. Table 4.8 summarizes the experimental results on the eight GMAs [115] which are used to build the IG database. Six general-purpose BIQA metrics were included in the performance comparison.

From Table 4.8 we can see that our proposed quality metric shows the best consistent ranking performance. Specifically, the proposed metric produces six consistent rankings among the eight considered GMAs. In comparison, the compared metrics produces at most three consistent rankings, which are significantly fewer than that of our proposed metric. This further validates the superiority of the proposed metric in benchmarking GMAs, which is quite useful in image processing systems.

4.4 Summary

This chapter presents a “completely blind” quality assessment metric for gamut-mapped images. A multivariate Gaussian model is pre-learned from local statistical features extracted from a set of pristine natural images. The local features are used to portray color and structural distortions, while features extracted from global statistics are utilized to characterize the global naturalness of image. The performance of the proposed metric has been tested on three gamut mapping databases. Compared to the relevant general-purpose BIQA metrics, the proposed metric produces predicted quality scores that are more consistent with human subjective perception. The proposed metric can also be used for benchmarking gamut mapping algorithms.

Chapter 5

Blind Image Quality Assessment Based on Multi-Scale Second-Order Statistics

Chapter 3 and Chapter 4 each introduces a distortion-specific BIQA metric. This chapter presents a general-purpose BIQA method based on multi-scale second-order statistics. Statistical features are extracted in both the wavelet domain and spatial domain on natural images. A regression module is employed to map the extracted features into quality scores.

5.1 Introduction

The majority of current statistics-based general-purpose BIQA methods, as discussed in Section 2.2, focus on first-order statistics, which try to compute the characteristics of single pixels overlooking the spatial relationship with other pixels. For example, the BRISQUE model [61] utilized the statistics of locally normalized luminance coefficients to measure image degradation. In recent years, research has found that higher-order statistics are also of great importance in the quality evaluation of images. Xu *et al.* [118] presented a BIQA metric incorporating higher-order image statistics for document image quality prediction. This metric showed superior performance in comparison with their previously proposed neural network-based metric [119] which considered only zero-order statistics from images. Liu *et al.* [120] utilized high-level statistical features to capture the quality degradation of real camera images. Both of these two methods employ features derived from cumulative higher-order moments such as skewness (third-order) and kurtosis (fourth-order). However, the third- and fourth-order image statistics are less robust than first- and second-order statistics due to the high power of those functions that employed to produce the statistics.

Second-order statistics have proved to be effective in various image processing tasks [109]. Huang *et al.* [121] pointed out that the second-order descriptor – HSOG provides more discriminative ability than first-order descriptors such as SIFT [122]. The first-order information usually can not describe second-order properties of an image such as variations of contrast or orientation whose detection requires the comparison of neighbouring pixel points [123]. Moreover, second-order image dependencies such as inner-orientation dependency are often disturbed by distortions, which can be utilized to quantify deviations caused by image impairments [124]. While there are extensive work on first-order statistical metrics, application of second-order statistics to BIQA remains largely under-investigated.

In this chapter, a blind image quality evaluator based on multi-scale second-order statistics (BQEMSS) is proposed. The statistical features extracted in the wavelet domain are used to model the joint distribution of adjacent subband coefficients, while features derived from the histogram of Gaussian derivative pattern in the spatial domain are employed to capture structural degradation. To quantify the statistical regularities between subband coefficients, three types of image dependencies are explored, which include spatially adjacent dependency, subband orientation dependency, and subband scale dependency. Particularly, a bivariate generalized Gaussian distribution is utilized to model the spatially adjacent bandpass responses. Moreover, the distortion effects on spatial-oriented correlations between adjacent subband coefficients are examined by deploying an exponentiated cosine model. Experimental results on several publicly available image quality databases demonstrate the good performance of our approach.

5.2 Proposed Method

Fig. 5.1 shows the flowchart of the proposed BQEMSS metric. Before feature extraction, images are first transformed into a working color space – CIELAB [103], so as to obtain better perceptual uniformity (see Section 4.2.1). Since the human visual system (HVS) perceives image structures in a coarse-to-fine strategy, the spatial-domain and wavelet-domain features are extracted at multiple scales. The two categories of features are stacked to form a feature vector, then feature vectors obtained from a set of training images are used to learn a support vector regression (SVR) model [85] providing with the subjectively rated scores. The SVR model is then utilized for the subsequent quality prediction of test image. The following subsections explain the

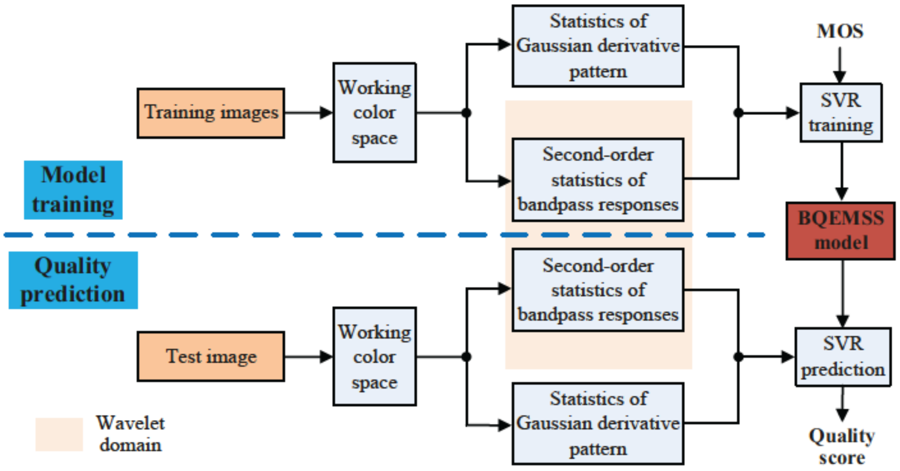

Training images
Working color space
Statistics of Gaussian derivative pattern
Second-order statistics of bandpass responses
MOS
SVR training
BQEMSS model
Model training
Quality prediction
Test image
Working color space
Second-order statistics of bandpass responses
Statistics of Gaussian derivative pattern
SVR prediction
Quality score
Wavelet domain

Figure 5.1: Flowchart of the proposed BQEMSS model.

extraction of these two categories of features.

5.2.1 Second-Order Statistics of Bandpass Responses

Studies have demonstrated that there exist strong statistical relationships between co-located bandpass coefficients [125]. Previous primary realizations mainly focus on extracting first-order univariate statistical features to capture the marginal descriptions of image bandpass responses, thus to create single pixels based statistical models (e.g., BRISQUE [61]). Note that first-order statistical features and second-order statistical features in wavelet domain differ in the numbers of pixels required for defining local features, namely one pixel for first-order and pair of pixels for second-order [126]. In this work, second-order image dependencies between adjacent subband coefficients (i.e., pair of coefficients) are studied. Particularly, the dependency between spatially neighbouring pixel points is captured by exploiting bivariate statistical features.

To form bandpass responses, distorted images are first decomposed using a wavelet transform model, i.e., the steerable pyramid [127]. Through the perceptual decomposition, we are able to extract multi-scale statistical features from the obtained subband coefficients. The use of steerable filters increases orientation selectivity and avoids aliasing in subbands [128]. Given a frequency tuning orientation θ, a steerable filter can be defined as

$$F(\theta) = \cos(\theta)F_x + \sin(\theta)F_y, \tag{5.1}$$

where F_x and F_y are the gradient component of two-dimensional bivariate isotropic Gaussian function along the horizontal and vertical directions respectively [129, 130]. The Gaussian function is described as in Eq. 3.3 where ρ is the scale parameter. By altering ρ, we are able to perform the multi-scale bandpass image decomposition. In this work, images are decomposed with three scales and over ten frequency tuning orientations.

Then a perceptually significant process (i.e., divisive normalization [131]) is applied to the subband coefficients. Note that a cortical neuron's response is normalized based upon the responses of its neighboring neurons in the HVS [125]. The use of divisive normalization is designed to de-couple subband responses, which can be computed as in Lyu *et al.* [132]

$$\hat{s}(x_i, y_i) = \frac{s(x_i, y_i)}{\sqrt{c_1 + \sum_j w(x_j, y_j)s(x_j, y_j)^2}}, \tag{5.2}$$

where (x_i, y_i) are spatial indices, s denotes the subband coefficients and $\hat{s}$ represents the coefficients after divisive normalization. c_1 is a saturation constant. The summation (Σ) is calculated over the neighboring pixels in the same subband which are indexed by j, and $w(x_j, y_j)$ is the associated Gaussian weighting function. After applying divisive normalization, the subband statistics of high quality images become more Gaussianized.

Univariate generalized Gaussian distribution has been widely-used to model the locally mean subtracted and contrast normalized coefficients which are often utilized to investigate first-order statistics [61, 67]. Here we study the bivariate statistics of spatially adjacent subband responses. Bivariate analysis examines the correlation between two sets of values, which can be used to explore second-order statistics [133]. In this work, A zero-mean bivariate generalized Gaussian distribution (BGGD) is employed to model the joint distribution of spatially adjacent subband coefficients $\hat{s}(x_i, y_i)$. The corresponding density function of BGGD is defined as

$$d(\mathbf{x}|\mathbf{M};\lambda,\delta) = \frac{1}{|\mathbf{M}|^{\frac{1}{2}}} g_{\lambda,\delta}\left(\mathbf{x}^T\mathbf{M}^{-1}\mathbf{x}\right), \tag{5.3}$$

where $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{M}$ is a 2×2 symmetric real scatter matrix. The parameter δ controls the "shape" of the distribution while λ controls the scale. $g_{\lambda,\delta}$ is the density generator which can be computed as

$$g_{\lambda,\delta}(z) = \frac{\delta}{2^{\frac{1}{\delta}}\pi\lambda\Gamma\left(\frac{1}{\delta}\right)} \exp\left(-\frac{1}{2}\left(\frac{z}{\lambda}\right)^{\delta}\right), \tag{5.4}$$

where $z \in \mathbb{R}^+$ and $\Gamma(\cdot)$ is the gamma function as defined in Eq. 3.14. In the case of $\delta \to \infty$, Eq. 5.3 converges to a bivariate uniform distribution; while $\delta = 0.5$, Eq. 5.3 corresponds the bivariate Laplacian distribution [125]. Note that when $\delta = 1$, Eq. 5.3 becomes the similar form of the distribution shown in Eq. 4.15 with $m = 2$. λ and δ are "quality-aware" features that can be estimated using the maximum likelihood estimator algorithm [112].

To quantify the statistical regularities between subband responses, we extract features from joint distributions of adjacent subband coefficients. Three types of

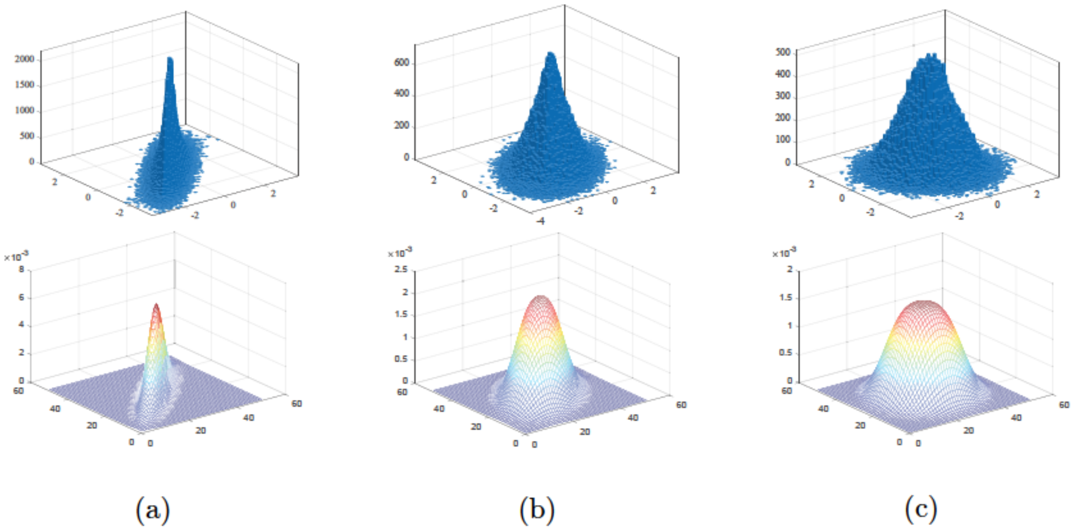


Figure 5.2: Bivariate distributions and BGGD fits of vertically adjacent subband coefficients derived from the pristine image Fig. 1.5 (a). Column (a), (b), and (c) correspond to tuning orientations 0°, 36°, and 90° respectively. The first row shows the bivariate distributions, where the blue bars denote the histograms. The colored 3D meshes in the second row represent the corresponding BGGD fits.

image dependencies are explored and the fact that the presence of distortion will alter these dependencies. We will use images in Fig. 1.5 to illustrate the different feature behaviors in pristine and distorted images.

1) *Spatially Adjacent Dependency*[1]: In the first image dependency, we investigate the statistics of spatially adjacent bandpass responses. Specifically, we emphasize the bivariate statistics of adjacent pixels along the horizontal and vertical directions at each subband. For an image $I(x, y)$, coefficient pairs are collected from adjacent locations $(x, y)(x + 1, y)$ and $(x, y)(x, y + 1)$ which correspond to spatial orientations 90° and 0° respectively. We represent the spatial orientation as θ_s. A number of bins are created using the subband coefficients and then bivariate histograms are obtained

[125].

The joint empirical distributions of spatially adjacent subband coefficients are well fitted by the BGGD model [124]. Fig. 5.2 shows the bivariate distributions and BGGD fits of vertically adjacent (i.e., $\theta_s = 0°$) subband coefficients derived from the lightness channel "L" of the pristine image Fig. 1.5 (a) at scale $\rho = 1$. Here we only plot figures with tuning orientations $\theta \leq 90°$, since the distribution across different θ is nearly symmetric around 90°. It can be seen that the shapes of both the bivariate distributions and BGGD fits change along with the subband tuning orientations. The three-dimensional illustrations of bivariate distributions exhibit close fits of BGGD. As θ_s matches θ (see Fig. 5.2 (a)), the distribution height reaches peak and becomes elliptical, which indicates highest dependency between vertically adjacent subband coefficients. On the other hand, the bivariate distribution becomes a circular Gaussian when θ_s and θ are orthogonal (see Fig. 5.2 (c)), which means almost no dependency exists at this tuning orientation. Generally, the bivariate distribution becomes more circular as θ increases and would exhibit an opposite trend in the case of $\theta \geq 90°$. Interestingly, the spatially adjacent subband coefficients derived from the red-green channel "A" and blue-yellow channel "B" of images also exhibit this kind of bivariate distribution. For simplicity, here we only discuss the statistics in lightness channel "L".

To further explain this spatially adjacent dependency, we plot the two BGGD model parameters λ and δ based on the difference between θ_s and θ on the images in Fig. 1.5. The results are shown in Fig. 5.3. The spatial and tuning orientation difference is defined as $\Delta_\theta = \theta - \theta_s$. We can see from Fig. 5.3 that both λ and δ

[1] *: Here means dependency between spatially adjacent subband coefficients in *wavelet* domain.

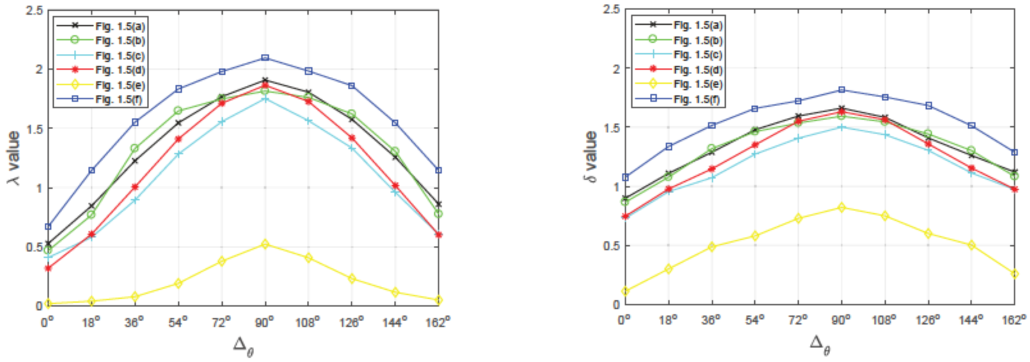


Figure 5.3: BGGD model parameter values across spatial and tuning orientation differences for images in Fig. 1.5.

exhibits strong dependencies. The values of λ and δ reach maximum when $\Delta_\theta = 90°$, i.e., θ_s and θ are orthogonal. Note that different types of distortion exhibit different extend of dependency, which proves that λ and δ can be used as quality-aware features. Since there are thirty subbands, a set of sixty features are then obtained.

2) *Subband Orientation Dependency*: It has been demonstrated that natural images exhibit statistical correlations across orientations [59]. Here we examine the orientation dependencies between adjacent subband coefficients in two perspectives.

First, adjacent subband responses within the same scale are investigated by utilizing a windowed structural correlation. Subbands are filtered using an 11×11 circular symmetric Gaussian weighting function with standard deviation of 1.5 [18]. With a similar formulation in Eq. 3.16, the structural correlation is calculated as

$$\beta_{ij} = \frac{2\alpha_{ij} + c_2}{\alpha_i^2 + \alpha_j^2 + c_2}, \tag{5.5}$$

where β_{ij} denotes the structural correlation map, α_i and α_j are windowed standard variances in adjacent subbands. α_{ij} is the cross-covariance within the local window between adjacent subband coefficients. c_2 is a stabilizing constant. By using the win-

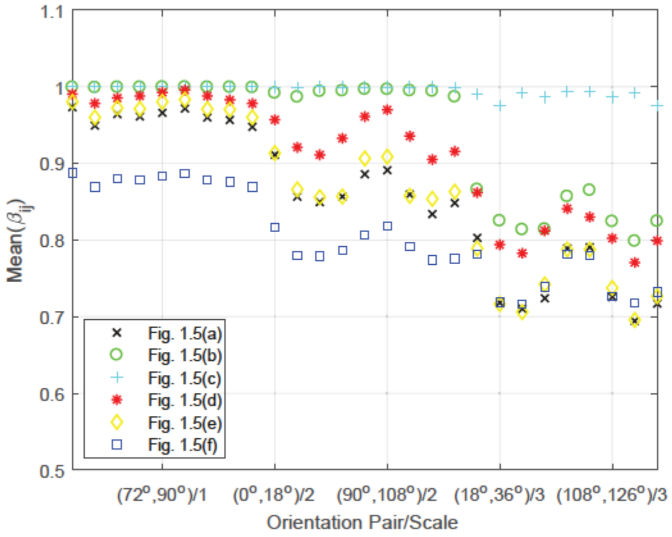


Figure 5.4: Adjacent orientation correlation statistics for images in Fig. 1.5.

dowing strategy, the orientation correlation maps exhibit a locally isotropic property. The mean of β_{ij} is taken as a quality feature. Fig. 5.4 shows the feature values of each adjacent orientation pair over three scales for images in Fig. 1.5. We can observe that features from one scale are not enough to distinguish all the orientation correlations. Here we propose to investigate the nine adjacent orientation pairs at three scales, yield a total of twenty-seven perceptual features. It can be seen from Fig. 5.4 that the orientation correlation between adjacent subband responses does exist in natural images, and this correlation is affected differently by different types of distortion.

On the other hand, studies have shown that the correlation between spatially adjacent subband coefficients exhibits periodic behavior across orientations [124]. An exponentiated cosine function can be utilized to model this periodic behavior, which is defined as

$$\phi = A\left[\cos(2\Delta_\theta)\right]^{2\sigma} + \epsilon, \tag{5.6}$$

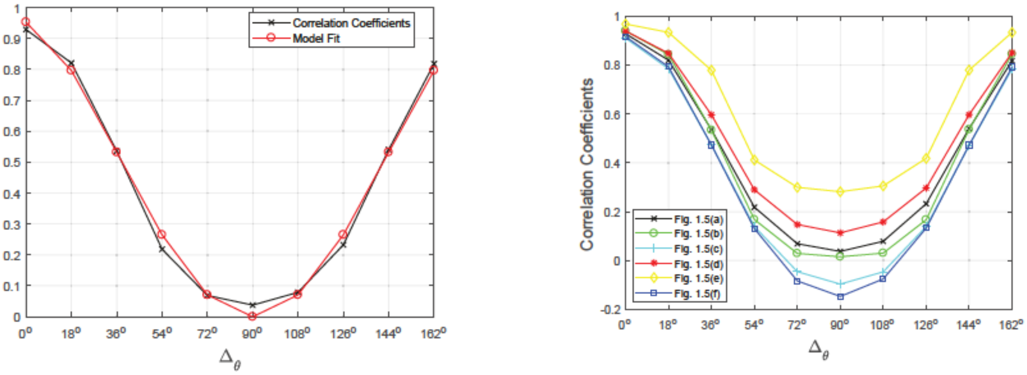


Figure 5.5: Correlations between vertically adjacent subband coefficients with respect to relative orientation for images in Fig. 1.5. The left image shows the distribution of correlation coefficients and its corresponding model fit for Fig. 1.5 (a), while the right image indicates that the distribution of correlation coefficients varys with different distortion types.

where ϕ represents the correlation coefficients, A denotes the amplitude. σ and ϵ are the shape exponent and offset respectively. This spatial-oriented correlations serve as inner-orientation dependencies of image bandpass responses. Fig. 5.5 shows the correlations between vertically adjacent subband coefficients with respect to relative orientation Δ_θ for images in Fig. 1.5 at scale $\rho = 1$. It can be seen that the correlations between vertically adjacent subband coefficients are well fitted by the exponentiated cosine model. The value of ϕ reaches minimum when $\Delta_\theta = 90°$, as θ_s and θ are orthogonal. The occurrence of distortions change the distribution of the correlation coefficients from Fig. 1.5 (a) in terms of amplitude and/or shape. Moreover, we can notice that distributions from different types of distortion exhibit different extent of deviations. The parameters A, σ and ϵ are then utilized as quality-aware features. For each scale, three parameters are computed, constituting the next set of nine perceptual features.

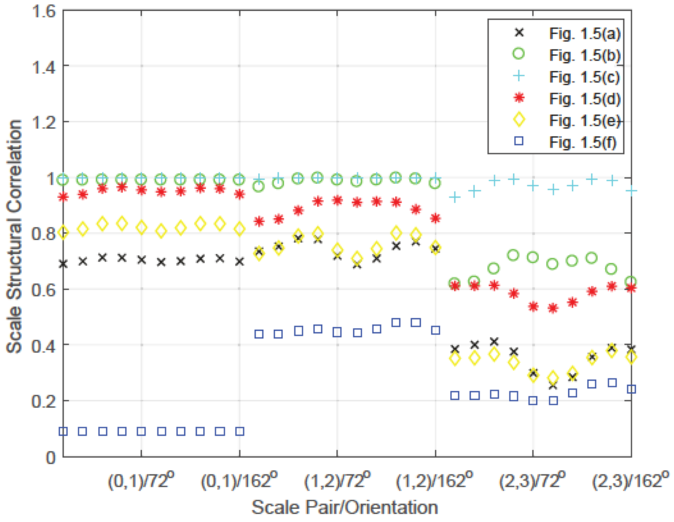


Figure 5.6: Adjacent scale correlation statistics for images in Fig. 1.5.

3) *Subband Scale Dependency*: Previous study found that the responses of retinal ganglion cells are closely related to enhancement of features like edges where the correlations exist across scales [59]. In this work, we investigate the statistical properties of subband responses between adjacent scales by utilizing a windowed structural correlation (see Eq. 5.5). Specifically, subband coefficients from adjacent scales are employed to compute the structural correlation map across orientations. Again, the mean of this correlation map is used as a quality feature. Fig. 5.6 shows the feature values of each adjacent scale pair over ten orientations for images in Fig. 1.5. Note that the high-pass subband is taken as "scale '0'" bandpass response. It can be seen that the scale-structural correlations between adjacent subband responses are influenced differently across distortion types. Since there are three scale pairs for each orientation, a group of thirty features are then extracted.

5.2.2 Statistics of Gaussian Derivative Pattern

First-order pattern features have been utilized for background modeling and face recognition [134]. However, the first-order features usually fail to capture more detailed discriminative information compared to that of second-order features. It has been demonstrated that gradient is a powerful descriptor of local image structure, and distortions in an image would change the distributions of its Gaussian smoothed gradient magnitudes [67]. Here we calculate the gradient magnitude map by convolving image I (lightness channel "L") with two Gaussian derivative filters along the horizontal and vertical directions respectively. The gradient magnitude g_m is computed as

$$g_m = \sqrt{\left(I_x * G^x\right)^2 + \left(I_y * G^y\right)^2}, \tag{5.7}$$

where G^x and G^y are the Gaussian derivative filters.

In the next step, spatially varying patterns in local regions of image I are encoded by applying the local binary pattern operator [135] on the gradient magnitude map. Note that applying the local binary operator on the original image would extract the first-order structural information [136]. Here we apply the local binary operator on the gradient magnitude map, thus to obtain second-order features. The Gaussian derivative pattern (GDP) code is defined as

$$\mathrm{GDP}_{N,R} = \sum_{k=0}^{N-1} \varphi\left(\tilde{g_m}\right) 2^k, \tag{5.8}$$

where $\tilde{g_m} = g_m^{(k)} - g_m^{(c)}$, R is the radius of neighbourhood circle and N is the number of neighbours. c and k denote the center and neighbouring locations respectively. If the coordinates of center c are $(0, 0)$, then the coordinates of its neighbours are given by $(-R\mathrm{sin}\frac{2\pi k}{N}, R\mathrm{cos}\frac{2\pi k}{N})$. $\varphi(\cdot)$ is the thresholding function

$$\varphi\left(\tilde{g_m}\right) = \begin{cases} 0, & \tilde{g_m} \leq 0 \\ \\ 1, & \tilde{g_m} > 0. \end{cases} \tag{5.9}$$

The GDP code characterizes the spatial structure of the local image texture. Ojala *et al.* [135] pointed out that "rotation invariant" codes can measure the occurrence statistics of individual patterns corresponding to certain micro-features in image, and the histogram of "uniform" patterns provides better discrimination compared to that of all individual patterns. In order to get a rotation invariant uniform texture descriptor, the GDP code is revised as

$$\overline{\mathrm{GDP}_{N,R}} = \begin{cases} \sum_{k=0}^{N-1} \varphi\left(\tilde{g_m}\right), & \text{if } \Psi\left(\mathrm{GDP}_{N,R}\right) \leq 2 \\ \\ N+1, & \text{otherwise} \end{cases} \tag{5.10}$$

where $\overline{\mathrm{GDP}_{N,R}}$ denotes the locally rotation invariant uniform GDP operator. $\Psi(\cdot)$ is the uniformity measure function, which is defined as

$$\begin{aligned} \Psi\left(\mathrm{GDP}_{N,R}\right) = &\left|\varphi\left(g_m^{(N-1)} - g_m^{(c)}\right) - \varphi\left(g_m^{(0)} - g_m^{(c)}\right)\right| \\ &+ \sum_{k=0}^{N-1} \left|\varphi\left(g_m^{(k)} - g_m^{(c)}\right) - \varphi\left(g_m^{(k-1)} - g_m^{(c)}\right)\right|. \end{aligned} \tag{5.11}$$

This uniformity measure corresponds to the number of bitwise transitions in patterns. The choice that restricts the bitwise transition number to no larger than 2 is proved to provide the GDP operator with better discriminative ability [135]. The uniform operator $\overline{\mathrm{GDP}_{N,R}}$ outputs $N+2$ patterns where one of them is labeled as "non-uniform pattern" and the rest are grouped as "uniform pattern".

In this work, we set $R = 1$ and $N = 8$, which means the uniform GDP operator $\overline{\mathrm{GDP}_{N,R}}$ has ten distinct output values. The obtained nine "uniform" patterns correspond to primitive micro-structures such as edges and spots, which can be taken as feature detectors [135]. Specifically, “0” represents bright spot, “1–7” stands for edges of different negative and positive curvature, “8” denotes dark spot or flat area. Note that the presence of distortion could change the pattern type, for example, blurring can change an edge pattern to flat pattern, which makes the uniform GDP operator an effective measure to describe different distortions [137].

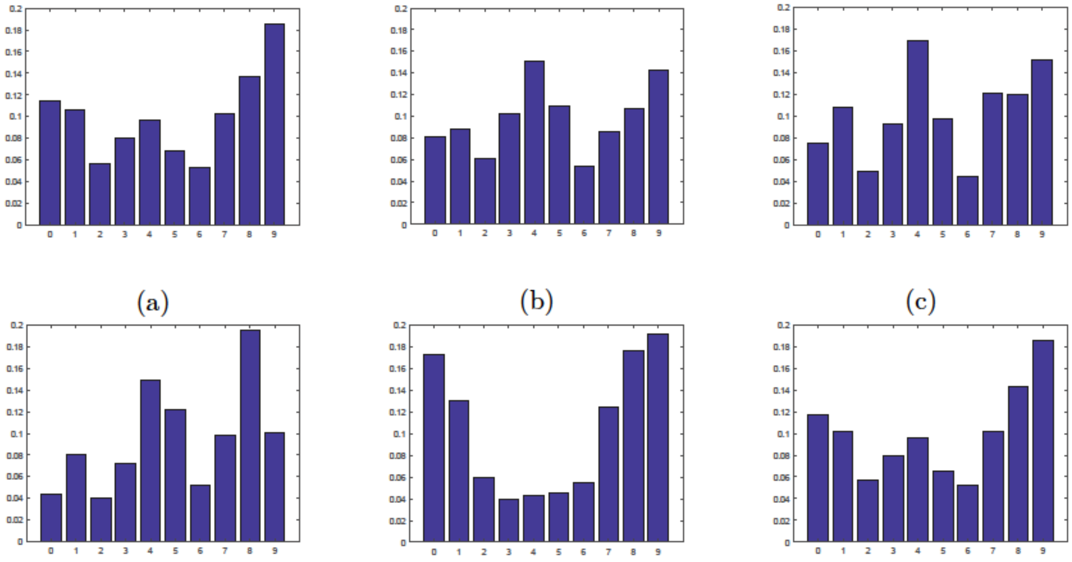
The uniform GDP operator is applied on each pixel to extract discriminative features from its neighborhood. We model the distribution of obtained patterns by spatial histogram. Suppose the size of an input image is $X \times Y$ and a pixel location is denoted as (i, j), the histogram of the uniform GDP patterns is computed as

$$H(p) = \frac{1}{XY} \sum_{i=1}^{X} \sum_{j=1}^{Y} \varphi' \left(\overline{\mathrm{GDP}_{N,R}}(i, j), p \right), \tag{5.12}$$

where $p \in [0, N+1]$ are the possible patterns, $\varphi'(\cdot)$ is a thresholding function as $\varphi(\cdot)$ (see Eq. 5.9). If $\overline{\mathrm{GDP}_{N,R}}(i, j) = p$ then $\varphi'(\cdot) = 1$, else $\varphi'(\cdot) = 0$.

Fig. 5.7 shows the histograms of the uniform GDP patterns on images in the CSIQ database [25]. We calculate the average histograms of images in the same category. It can be seen that the structural histograms vary by distortion type except for the contrast change distortion. Since the uniform GDP operator excludes magnitudes of the difference between the center pixel and its neighbours in encoding, it fails to capture local contrast change (cause blur between neighbouring pixels) in images which is important in the human perception [138]. To this end, we modify H, leveraging gradient magnitude g_m (see Eq. 5.7) as a weighting map. The improved histogram H' is then calculated as

$$H'(p) = \frac{1}{XY} \sum_{i=1}^{X} \sum_{j=1}^{Y} g_m(i, j) \varphi' \left(\overline{\mathrm{GDP}_{N,R}}(i, j), p \right). \tag{5.13}$$



(a) (b) (c)

(d) (e) (f)

Figure 5.7: Histograms of the uniform GDP patterns on images in the CSIQ database [25]. (a) Histogram from pristine images. (b–f) correspond to histograms from distorted images where the distortion type is Gaussian blur, JPEG2000 compression, JPEG compression, white noise, and contrast change respectively.

The histogram H' incorporates the structural and gradient information, which makes it effective to describe the impact of distortions. We extract ten quality-aware features from each image as H' has ten bins. To capture multi-scale behavior, features are computed at six scales by down-sampling the images, yielding a set of sixty features.

5.2.3 Regression

In the final stage, a classic regression module SVR [85] is utilized to learn the proposed quality prediction model. Details about the regression procedure are given in Section 3.2.4.

5.3 Experimental Results and Analysis

The performance of the proposed BQEMSS model was evaluated on five public image quality databases. The databases are detailed in Table 1.1, including the numbers of distorted images, distortion types, etc. A total of 6281 images were used in the experiments. We compared our proposed BQEMSS model with nine leading statistics-based general-purpose BIQA metrics, including BIQI [58], DIIVINE [59], BLIINDS-II [60], BRISQUE [61], QAC [68], BHOD [136], NIQE [66], NFERM [65], and ILNIQE [67]. Two evaluation criteria were adopted as described in Section 1.2. In our proposed BQEMSS model, there are three free parameters, including the number of scales and orientations in wavelet decomposition and the number of scales in spatial-domain feature extraction. These parameters were tuned based on the CSIQ database [25] by maximizing the PLCC and SRCC values, and then they were applied in the performance evaluation of the other four databases.

5.3.1 Performance Comparison

We first evaluated the IQA models on each individual database using the train-test procedure introduced in Section 3.3.1. To get fair comparison, metrics that do not incorporate training process (as described in Section 2.2) were evaluated on the partitioned testing subsets. The results are summarized in Table 5.1. It can be seen from Table 5.1 that the proposed BQEMSS model performs consistently well in the five databases. Specifically, in CSIQ, BQEMSS produces the best prediction accuracy (PLCC) and monotonicity (SRCC). In LIVE, although not the best, BQEMSS performs only slightly worse than the best metrics. In the largest database TID2013 and

Table 5.1: Performance comparison on each individual database. The best result is marked in boldface.

	LIVE		CSIQ		TID2013		LIVEWC		CID2013	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BIQI	0.7405	0.7581	0.5348	0.5209	0.3664	0.3458	0.5479	0.5324	0.6757	0.6569
BLIINDS-II	0.9129	0.9165	0.6983	0.6961	0.6278	0.5307	0.5064	0.4885	0.4987	0.4766
DIIVINE	0.8923	0.8822	0.7154	0.7150	0.6595	0.5471	0.5283	0.5148	0.5124	0.4972
BRISQUE	0.9275	0.9296	0.6950	0.7077	0.5422	0.5296	0.5864	0.5685	0.4783	0.4309
QAC	0.8612	0.8645	0.6323	0.5161	0.4950	0.3914	0.2835	0.2764	0.2244	0.2063
BHOD	**0.9344**	0.9282	0.7654	0.7483	0.6137	0.5432	0.5361	0.4755	0.6092	0.5934
NIQE	0.9037	0.8993	0.7275	0.6214	0.4363	0.3236	0.4848	0.4292	0.6136	0.6007
NFERM	0.9321	**0.9369**	0.8030	0.7861	0.6709	0.6389	0.6055	0.5908	0.6322	0.6281
ILNIQE	0.8904	0.8936	0.8335	0.8143	0.6456	0.5278	0.5127	0.5033	0.4634	0.4540
BQEMSS	0.9305	0.9212	**0.8464**	**0.8225**	**0.6973**	**0.6674**	**0.6335**	**0.6212**	**0.7214**	**0.7053**

two real camera image databases LIVEWC and CID2013, BQEMSS achieves notably better prediction performance than the compared methods. From these results we then draw the conclusion that BQEMSS produces the state-of-the-art performance. It is worth noting that the competing methods usually perform better on legacy databases than on the real camera image databases, which indicates that quality assessment for real camera images reserves great potential for further study.

Besides direct comparisons with numerous IQA metrics, we further analyzed the statistical significance of the proposed model using F-test [12]. Details about the F-test can be found in Section 3.3.1. Fig. 5.8 shows the F-test results between the compared metrics and the proposed BQEMSS model on the five databases. It can be observed that BQEMSS produces the smallest prediction errors among all the compared metrics in CSIQ, TID2013, LIVEWC, and CID2013. In LIVE, only NFERM produces slightly smaller prediction errors than BQEMSS. Table 5.2 lists the statistical significance between the proposed metric and ten compared metrics. It

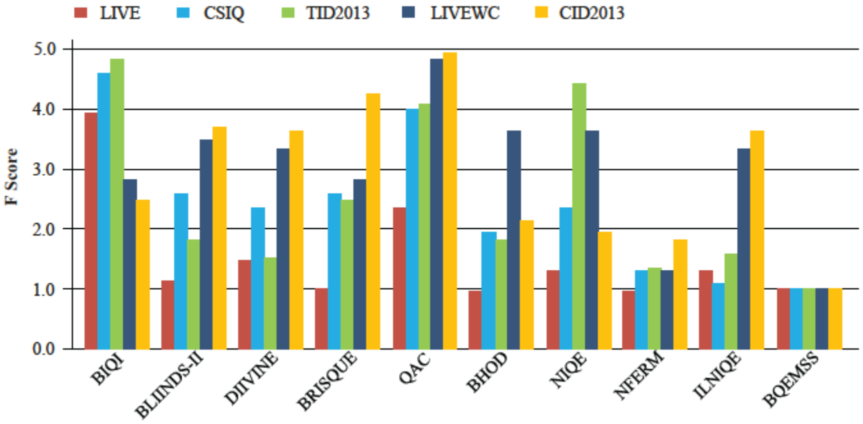


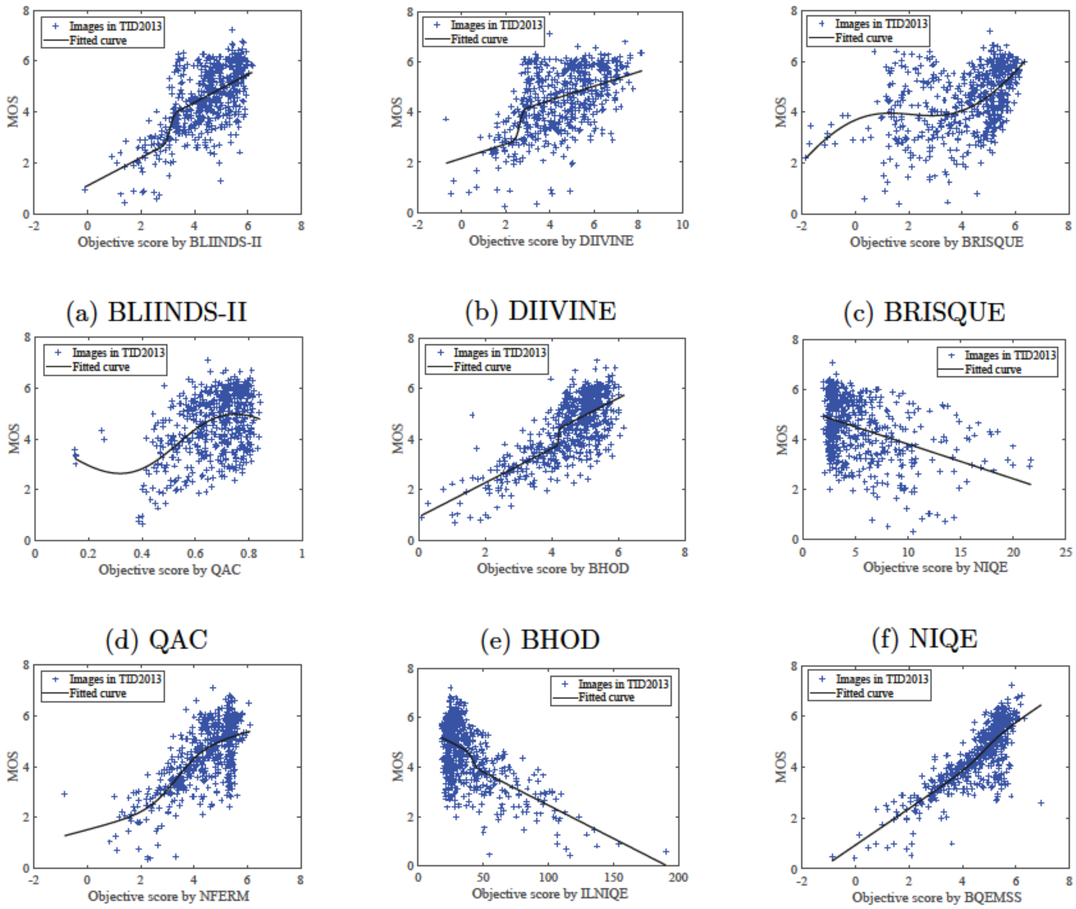
Figure 5.8: F-test results of the compared metrics against the proposed BQEMSS method.

Table 5.2: Statistical performance between the proposed BQEMSS method and the compared metrics on five databases.

Metric	LIVE	CSIQ	TID2013	LIVEWC	CID2013
BIQI	+1	+1	+1	+1	+1
BLIINDS-II	0	+1	+1	+1	+1
DIIVINE	+1	+1	+1	+1	+1
BRISQUE	0	+1	+1	+1	+1
QAC	+1	+1	+1	+1	+1
BHOD	–1	+1	+1	+1	+1
NIQE	+1	+1	+1	+1	+1
NFERM	–1	+1	+1	+1	+1
ILNIQE	+1	0	+1	+1	+1

is clear from Table 5.2 that the proposed BQEMSS model is superior to the existing metrics in most cases. Among the 45 combinations of metrics and databases, our metric performs significantly better in 40 cases, comparable in 3 cases and worse in only 2 cases. This indicates that the proposed method performs consistently well across all databases, which is desired in practical applications.

For visualization, we provide the scatter plots of the subjective MOS values against



(a) BLIINDS-II (b) DIIVINE (c) BRISQUE

(d) QAC (e) BHOD (f) NIQE

(g) NFERM (h) ILNIQE (i) BQEMSS

Figure 5.9: Scatter plots of subjective MOS values against predicted scores by different metrics on the TID2013 database.

the objective scores predicted using different quality metrics in the largest TID2013 database (see Fig. 5.9). A good metric is expected to produce scatter plots with better convergence and monotonicity. It can be observed from Fig. 5.9 that the quality scores predicted by the proposed BQEMSS model are more consistent with subjective evaluations.

In the next experiment, we compared the performance of BQEMSS and five leading general-purpose BIQA metrics (BLIINDS-II, DIIVINE, BRISQUE, NFERM, and

Table 5.3: SRCC results on each individual distortion type in the full TID2013 database. The indexes represent the distortion types explained in Fig. 1.6.

Metric	#(1,1)	#(2,1)	#(3,1)	#(4,1)	#(5,1)	#(6,1)	#(1,2)	#(2,2)
BLIINDS-II	0.7148	0.6642	0.5269	0.4118	0.7362	0.6096	0.5830	0.8165
DIIVINE	0.7537	0.4676	0.4684	0.3766	0.6084	**0.7468**	0.5656	0.8075
BRISQUE	0.8284	0.5547	0.4948	0.3459	0.6204	0.6055	0.6786	**0.8426**
NFERM	0.8262	**0.8393**	0.3127	0.2473	**0.8958**	0.5525	0.6684	0.8187
ILNIQE	**0.8737**	**0.8155**	**0.8703**	**0.5107**	**0.8653**	0.7047	**0.8733**	0.8132
BQEMSS	**0.8643**	0.7266	**0.8458**	**0.6047**	0.8385	**0.7106**	**0.8869**	**0.8221**

Metric	#(3,2)	#(4,2)	#(5,2)	#(6,2)	#(1,3)	#(2,3)	#(3,3)	#(4,3)
BLIINDS-II	0.6882	0.8684	0.8894	0.2424	0.2818	0.0982	**0.3235**	0.1250
DIIVINE	**0.7516**	0.7927	0.8533	**0.3365**	0.2609	**0.1389**	**0.1924**	0.0914
BRISQUE	0.5821	0.9055	**0.9135**	0.2878	0.3136	0.0829	0.1717	0.1122
NFERM	0.6554	**0.9277**	**0.9264**	0.2140	0.2804	0.0629	0.0718	**0.1842**
ILNIQE	0.7257	0.8386	0.8536	0.2828	**0.4217**	0.0820	0.1331	0.1558
BQEMSS	**0.7952**	**0.9135**	0.9075	**0.3488**	**0.3167**	**0.2023**	0.1876	**0.1764**

Metric	#(5,3)	#(6,3)	#(1,4)	#(2,4)	#(3,4)	#(4,4)	#(5,4)	#(6,4)
BLIINDS-II	0.0286	0.0146	0.3075	**0.4610**	0.5169	**0.7882**	0.5363	0.8159
DIIVINE	0.1686	0.1632	0.7463	0.2157	0.6336	0.6384	**0.6657**	0.8353
BRISQUE	0.0488	0.1053	0.6263	0.2068	0.5889	**0.7747**	0.6586	0.7654
NFERM	**0.2624**	**0.3335**	**0.8053**	0.1197	**0.7886**	0.6635	0.5779	0.6585
ILNIQE	0.1125	0.1696	0.6932	0.3535	0.7558	0.7462	**0.6777**	**0.8636**
BQEMSS	**0.6532**	**0.3826**	**0.7886**	**0.4179**	**0.7797**	0.7065	0.6211	**0.8582**

ILNIQE) on each distortion type in TID2013. The competing models were trained on the 80% of images of various distortion types and then tested on the left 20% of images with the specific distortion type. Their performance are reported in Table 5.3. Without losing the generality, only the SRCC results are shown. For each performance criterion and database, the two best results are highlighted in bold. Note that the TID2013 database has 24 distortion types, where the detailed distortion information can be found in Fig. 1.6.

From the results shown in Table 5.3, we can see that BQEMSS has won 18 times in the first two place, whereas the compared metrics at most have 9 times. This further validates the effectiveness of the proposed BQEMSS model across various distortion types compared to the leading general-purpose BIQA metrics. Note that for distortion type "$\#(5,3)$", the proposed BQEMSS model outperforms the other metrics by a large margin, which owes to the improved histogram of Gaussian derivative patterns (see Section 5.2.2). Moreover, it is observed that for distortion types "$\#(1,3)$", "$\#(2,3)$", "$\#(2,4)$", "$\#(3,3)$", "$\#(4,3)$", "$\#(6,2)$", and "$\#(6,3)$", none of the evaluated models is able to obtain satisfying results. This indicates that more investigations are needed to deal with these "sophisticated" distortion types.

5.3.2 Method Analysis

To test the generalization capability of the proposed BQEMSS model, we conducted cross-database performance evaluation. Five general-purpose BIQA metrics that incorporate training process were included in the comparison, namely BIQI, BLIINDS-II, DIIVINE, BRISQUE, and NFERM. The competing metrics were trained on one database and then tested on other two databases. The results are shown in Table 5.4.

It is clear in Table 5.4 that the proposed BQEMSS model exhibits the best generalization capability. Specifically, when trained on TID2013 and then applied to other databases, BQEMSS outperforms the compared metrics in terms of both prediction accuracy and monotonicity in CSIQ. In LIVE, only NFERM is comparable to our proposed BQEMSS model. When trained on LIVE, the proposed BQEMSS metric outperforms all the compared metrics by a large margin. Note that in the case of training on LIVE, the competing methods deliver worse performance than that of training on TID2013. This is not surprising since the TID2013 database contains far

Table 5.4: Cross-database performance evaluation. Models are trained on full LIVE or TID2013 respectively, and tested on the other two databases. The best result is marked in boldface.

	Trained on LIVE				Trained on TID2013			
	CSIQ		TID2013		LIVE		CSIQ	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BIQI	0.3650	0.2725	0.0685	0.0608	0.6031	0.5858	0.3802	0.2817
BLIINDS-II	0.4342	0.3460	0.0945	0.0833	0.8088	0.7857	0.4991	0.4208
DIIVINE	0.4113	0.3332	0.1284	0.1089	0.6421	0.6294	0.4780	0.4528
BRISQUE	0.3925	0.3679	0.1453	0.1370	0.5877	0.5247	0.4517	0.4167
NFERM	0.5585	0.5209	0.1965	0.1832	**0.8283**	0.7963	0.6567	0.5644
BQEMSS	**0.6843**	**0.6275**	**0.4165**	**0.3211**	0.8211	**0.8054**	**0.7398**	**0.7193**

more distortion types than the other two databases, and a majority of distortion types in TID2013 are not included in the other two databases. The results indicate that BQEMSS not only produces very good overall performance, but also shows superior generalization capability.

In BQEMSS, we use two categories of features, including wavelet-domain features and spatial-domain features. To explore the contributions of two components, we evaluated the performance of each feature category separately on all databases. The results are reported in Table 5.5.

It can be seen from Table 5.5 that by using a single type of features, the performance of BQEMSS is much worse than using the integrated features. The wavelet-domain features are more capable of quantifying distortions in the LIVE database, while spatial-domain features are superior on the quality prediction in CSIQ and CID2013. In TID2013 and LIVEWC, the two categories of features exhibit comparable performance. This indicates that both two categories of features are needed in the proposed metric, and they have complementary contributions to the overall

Table 5.5: Relative contributions of the categories of features.

Database	Criterion	Wavelet domain	Spatial domain	All
LIVE	PLCC	0.9138	0.8927	0.9305
	SRCC	0.9046	0.8904	0.9212
CSIQ	PLCC	0.7894	0.8037	0.8464
	SRCC	0.7355	0.7835	0.8225
TID2013	PLCC	0.6418	0.6329	0.6973
	SRCC	0.6354	0.6032	0.6674
LIVEWC	PLCC	0.6094	0.6138	0.6335
	SRCC	0.5994	0.6057	0.6212
CID2013	PLCC	0.6388	0.6947	0.7214
	SRCC	0.6229	0.6791	0.7053

performance.

5.4 Summary

In this chapter, a general-purpose BIQA metric based on multi-scale second-order statistics is presented, namely BQEMSS. The proposed BQEMSS model investigates second-order statistics in both the wavelet domain and spatial domain on natural images. Specifically, the statistical features extracted in the wavelet domain are used to model the joint distribution of adjacent subband coefficients, while features derived from the histogram of Gaussian derivative pattern are employed to capture structural degradation in the spatial domain. We have conducted extensive experiments and comparisons on five subjectively rated image quality databases. The experimental results have demonstrated that BQEMSS is superior over the state-of-the-art statistics-based general-purpose BIQA models.

Chapter 6

Deep Neural Network for Blind Image Quality Assessment

The BIQA metrics presented in Chapters 3 – 5 are all heuristic-based, which require highly manually engineered feature designs. In addition, these frameworks usually involve separate feature extraction stage and quality prediction stage. In this chapter, a fully data-driven learning-based model is presented, which provides an end-to-end solution to general-purpose BIQA.

6.1 Introduction

During the past few years, convolutional neural networks (CNNs) have an enormous impact on computer vision research and witnessed the great development of data-driven algorithms for various vision tasks. Instead of carefully designing hand-crafted features, deep CNNs models are able to automatically discover feature representations from raw image data. With continuous efforts of boosting the abilities of CNNs, a series of popular architectures have been developed, such as AlexNet [139], VGGNet [140], GoogLeNet [141], ResNet [142], and DenseNet [143], etc., which have been widely used in many vision applications [144, 145, 146, 147, 148, 149, 150]. Inspired by the huge success of CNNs and their higher representational ability, a range of CNNs-based approaches (as described in Section 2.2) have been progressively proposed to promote the performance of IQA.

Analogous to the applications in other vision tasks, the CNNs-based IQA frameworks also conform to the basic principles of CNNs architecture design, which generally consist of a feature extractor for characterizing the rich features of input images followed by a regression head for predicting the final quality assessment scores. For example, Kang *et al.* [69] developed a network with the first several convolutional layers producing feature maps and the last layer is a simple linear regression with an one-dimensional output that predicts the score. In Bosse *et al.* [74], features were extracted from the distorted image patches and the reference image patches by CNNs and then the feature vector was regressed to a patch-wise quality estimate. Although impressive prediction accuracies have been achieved, these prior approaches somewhat ignore the importance of multi-level supervision and multi-scale integration, which have become the study focuses in other realms of vision problems, e.g., saliency detection [151, 152], crowd counting [153, 154], and classification [155, 156, 157]. The works [158, 159] have demonstrated that shallow layers are beneficial for extracting informative localization information, whereas higher layers usually feature the more abstract and high-level semantic cues. Moreover, the inception module in GoogLeNet [141] with different kernel sizes proves the effectiveness of multi-scale features aggregation.

Recently, investigation on delicate CNNs structures with higher learning ability and finer feature fusion has drawn attention in designing IQA models. Gao *et al.* [160] analyzed the deficiencies of shallow networks and proposed a deep structure with multi-level feature aggregation to reduce the sensitivity to local feature degradations, where the potential of utilizing supervision of intermediate features to boost the prediction performance was discussed. Even though the usage of multi-level features obtained certain performance improvement, they densely reused the previous layers and inevitably increased the input dimension for the final score predictor, thereby impeding the model to well adapt to testing datasets. Considering the limited training samples in available IQA datasets, it is meaningful to sparsely sample the feature levels for final prediction with the demands of necessary and fundamental multi-level semantics. Furthermore, scale-invariance was neglected in the underlying network of this pioneer work, whereas Fu *et al.* [161] explained that multi-scale features are of great importance for the robustness of IQA methods, as different receptive fields can implicitly capture distinct noise levels.

To mine the diverse semantic levels of representations and to extract scale-invariant features while avoiding drastic increase of model's parameters, a novel CNNs-based multi-scale integration network (MSINet) is put forward for general-purpose BIQA in this chapter. The proposed MSINet combines the advantages of multi-level and multi-scale aggregations at the same time. In specific, a VGG-16 model pretrained on ImageNet database [162] is attached at the beginning of the network to extract general features, and then a group of multi-scale integration modules (MSIs) are successively connected to exploit high-level fine-grained representations. To facilitate the convergence of the network and help MSIs better learn useful information, residual connections among deep MSI modules and two-level pretrained features are incorporated. Finally, features at multiple levels (low, middle, and high) are concatenated and fed into the regression head to utilize the discriminative hints for score estimation. Apart from the discriminative features, this multi-level strategy also plays a vital role in directly passing the gradient flows to deep layers, and thus avoids the issues of gradient vanishing.

It is worth noting that attention mechanisms have become the prevailing construction units in CNNs architectures and obtained promising improvements. For example, the SENet [163] presented an attention protocol along the convolution channels to emphasize informative kernels while suppressing useless ones. The spatial attention mechanism [164] was proposed to reweigh the spatial pixel-wise values. Benefiting from the intrinsic merits of attention mechanisms, these two strategies have been used in many CNNs models for performance boosting [165, 166, 167]. To take advantages of the attention mechanisms and better recalibrate the array of kernels from multi-level features, in our proposed MSINet, channel-wise multi-level adaptation is deployed on the combined multi-level feature maps before entering into the regression head.

Experimental results on five public available datasets shows that our proposed MSINet attains the state-of-the-art performance among the existing learning-based general-purpose BIQA approaches. Ablation studies also shed light on the effectiveness of each component in our model.

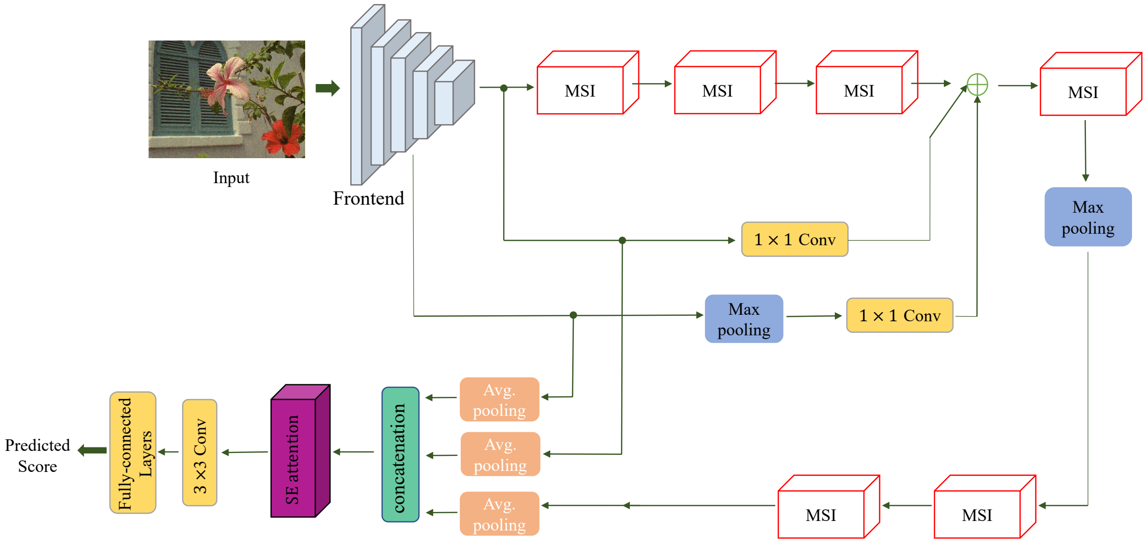


Figure 6.1: The overall architecture of our proposed MSINet.

6.2 Proposed Method

6.2.1 The Overall Pipeline of MSINet

As shown in Fig. 6.1, the proposed MSINet for general-purpose BIQA is comprised of four primary components: the pretrained frontend, a stack of multi-scale integration modules (MSIs), multi-level semantics reuse, and channel-wise multi-level adaptation. In the rest of this section, we will elaborate each component accordingly.

6.2.2 The Frontend

The pretrained VGGNet features the simplification and good generalization ability of the model and has dominated various downstream tasks (e.g., object detection [144] and visual tracking [168]) by fine-tuning its parameters on datasets in unseen domains. Although ResNet [142] and DenseNet [143] are also well-known pretrained

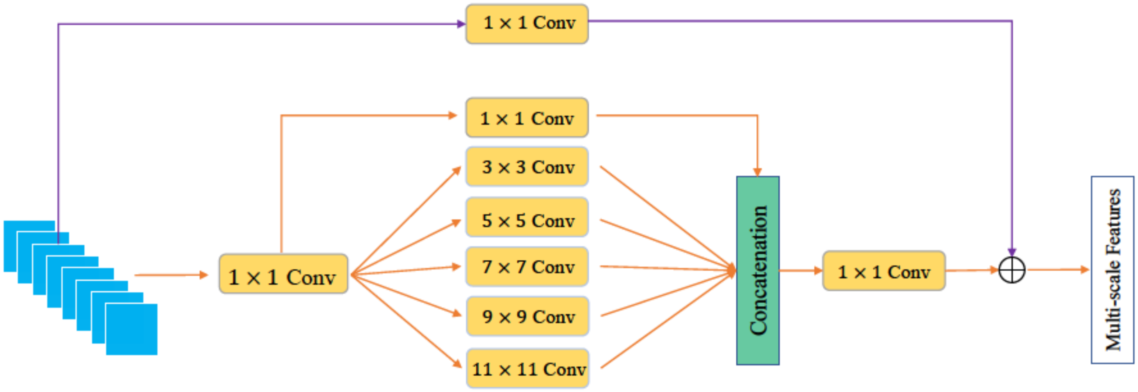


Figure 6.2: Multi-scale integration module (MSI). It consists of multi-branch convolutions with kernel sizes from 1×1 to 11×11 to extract fine-grained features with the property of scale-invariance. The residual learning [142] is leveraged to facilitate the convergence of the whole network.

models, they have relatively higher scales and are easily prone to overfitting. Hence, in our method, the first ten layers of VGGNet-16 pretrained on ImageNet database [162] are selected to build the frontend stem for extracting low-level features, which is capable of remedying the issues of limited training samples. The use of pretrained model not only speeds up the convergence of the network, but also aims at providing low- and middle-level features for enriching the clues of distortion extent inference. Note that the frontend is fine-tuned along with the backend in an end-to-end manner.

6.2.3 Multi-Scale Integration Module

Inspired by the inherent merits of inception module in GoogLeNet [141], we propose a multi-scale integration (MSI) module to extract characteristics at varied scales, see Fig. 6.2. Different from the original inception module, we remove the pooling sub-branch from the cardinality dimension and move a further step to broaden the diversity of scale branches. The six branches of convolutions C_i ($i \in \{1,2,3,4,5,6\}$) with different kernel sizes from 1×1 to 11×11 are delicately designed to capture abundant receptive fields. Taking the model complexity into account, the plain convolutions with real larger kernel sizes are evaded as they easily incur the explosion of computational overheads. Instead, the atrous convolutions [169] with a number of dilation rates (1,2,3,4,5) are leveraged in the MSI module to enlarge the kernel sizes.

As shown in Fig. 6.2, an 1×1 convolution layer (R_f) is first plugged in the front of a multi-branch structure to readjust and fuse the input feature maps. The outputs of all branches are concatenated together along the channel dimension to form the input of the rear 1×1 convolution (R_r) with the purpose of dimension reduction. Notably, residual learning [142], which uses identity mapping for shortcut connections, is employed in this module to accelerate the learning procedure by incorporating residual connection. Given the input features F, the computation of the MSI module can be formulated as

$$\begin{aligned}
&T_1 = R_f(F, W_{R_f})\\
&B_1 = C_1(T_1, W_{C_1}), B_2 = C_2(T_1, W_{C_2}), B_3 = C_3(T_1, W_{C_3})\\
&B_4 = C_4(T_1, W_{C_4}), B_5 = C_5(T_1, W_{C_5}), B_6 = C_6(T_1, W_{C_6})\\
&T_2 = B_1 \circ B_2 \circ B_3 \circ B_4 \circ B_5 \circ B_6\\
&T = R_r(T_2, W_{R_r}) + T_1,
\end{aligned} \tag{6.1}$$

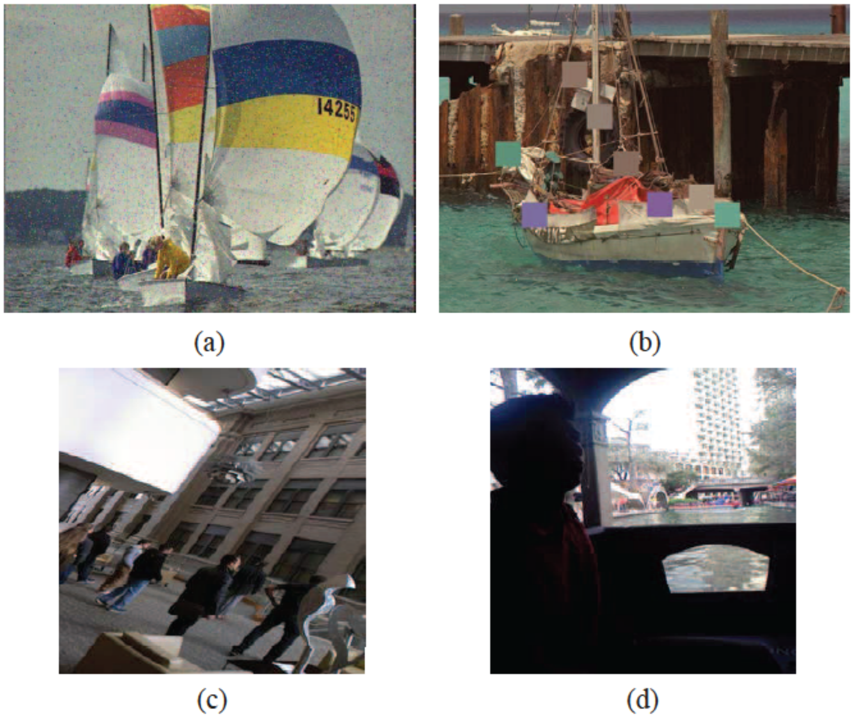
where $W_{R_f}, W_{C_1}, W_{C_2}, W_{C_3}, W_{C_4}, W_{C_5}, W_{C_6}, W_{R_r}$ denote the learnable parameters of corresponding convolutions and "$\circ$" represents the concatenation operation. T_i and B_i represent the intermediate features and T is the output. All convolution layers are followed by batch normalization and ReLU activation function to reduce the internal covariate shift and to add non-linearity. This multi-scale convolution aggregation module is believed to substantially reduce the number of parameters while preserving the performance.

As for the configuration of the MSINet, six MSI modules are stacked sequentially to process the general features from the frontend and to produce the high-level representations with more abstract hints, as shown in Fig. 6.1. To delve deep into previous multi-scale information encoded by the frontend, we devise side shortcut connections among intermediate layer and final layer of the frontend, and the output of the third MSI module. Then three the groups of feature maps are combined through max pooling, 1×1 convolution and pixel-wise summation.

6.2.4 Multi-Level Semantics Reuse

As described in Section 6.1, features extracted at multiple levels are vital for the development of IQA metrics since different levels focus on various semantics of granularities. The categories and distributions of distortions dispersing across natural images are drastically varied, from global to local [26]. Therefore, it is reasonable and meaningful to capture multi-granularity features for multi-level noises that appeared in images. CNNs naturally learn hierarchical features with the depth of layers from shallow to deep [160]. For example, the assessment of global distortion may heavily rely on the shallow layers with the behaviour of low-pass filtering (e.g., low-level features as described in Chapter 3), whereas modelling local distortions demands more abstract features provided by deep layers[1].

[1] *: In IQA, global distortion means the distortion operates "globally", which differs from the terminology in CNNs architectures such as global features. For example, impulse noise and Gaussian



(a) (b)

(c) (d)

Figure 6.3: Images with conventional or realistic distortions. (a) and (b) are conventional distortions where (a) is contaminated with impulse noise and (b) comes with local block-wise distortions (e.g., inpainting situation). (c) and (d) are realistic distortions. (a) and (b) are from the TID2013 database [26]. (c) and (d) are from the LIVEWC database [27].

Fig. 6.3 depicts images with several types of distortions, which explains that multi-level semantics are needed in the assessment of distortions. Specifically, Fig. (a) and (b) are conventional distortions, whereas Fig. (c) and (d) are realistic distortions as described in Chapter 1. Intuitionally, global objects-agnostic distortions like Fig. (a) may be measured through the correlation differences from the gradient and contrast information that usually captured by shallow features in CNNs, whereas for local

noise are considered as global distortion.

distortions like Fig. (b), the objects that dispersing across the whole scene need to be recognized by resorting to higher-level representations with larger receptive fields in CNNs models. Particularly, assessing the distortions in Fig. (c) and (d) needs more rich and abstract semantic information provided by deeper CNN layers. For example, in Fig. (c), observers need to recognize that the building and the persons are not in their upright positions, and in Fig. (d), it is vital to be aware of there is a person in the left part of the image before predicting the image quality.

To avoid the dense sampling of levels, we predefine three levels of representations to be reused by the decoder head, namely low-, middle-, and high-level semantics (denoted as L_1, L_2, and L_3), which cover diverse noises levels. Specifically, the low-level semantics are delivered from the fourth layer of the frontend, and the output of the frontend is regarded as the middle-level cues, while the final MSI module produces the high-level semantics, see Fig. 6.1. Three adaptive average pooling operations are imposed on the three-level semantics respectively so as to reduce dimensions of the feature maps, and then the results are concatenated together to be utilized by the subsequent module. Hence, the output H_{in} of multi-level semantics reuse can be denoted as

$$H_{in} = avg(L_1) \circ avg(L_2) \circ avg(L_3), \tag{6.2}$$

where "avg" means adaptive average pooling. The concatenation operations perform along the channel dimension. By this means, the discriminative features at multiple levels of granularities are fused and passed through the regressor for adapting to a wide range types of distortions.

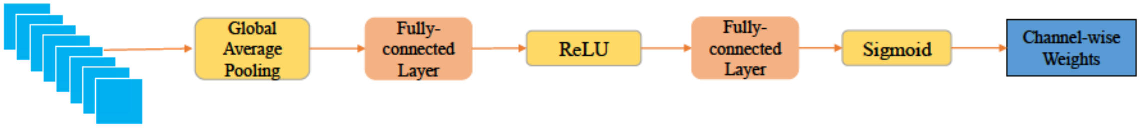


Figure 6.4: Channel-wise attention mechanism. The learned weights are imposed on input features along the channel dimension. The output width of sigmoid function is identical to that of input features.

6.2.5 Channel-Wise Multi-Level Adaptation

Once the multi-level features are extracted, it is natural to combine them and then take them as the input of regression head. An intuitive idea is to treat them equally and directly convolve them. However, this integration method inevitably introduces noises into the subsequence convolution layer as there are redundant or information-less channels, thereby increasing burdens of learning at the next layer. To tackle this issue, a channel-wise attention ($Attn$) strategy is implemented to achieve the multi-level adaptation, see Fig. 6.4. The attention subnetwork takes the multi-level characteristics H_{in} as input and computes a set of channel-wise weights through a sequence of convolution, average pooling, fully-connected layers, and sigmoid function. Then, the input representations are recalibrated using the learned channel-wise weights to emphasize informative maps while suppressing the impacts of useless ones on the regressor, which is formulated as

$$\hat{H}_{in} = Attn(H_{in}, W_{attn}) \times H_{in}, \tag{6.3}$$

where W_{attn} indicates the parameters of the attention subnetwork.

$\hat{H}_{in}$ is further fed into the regression head RH to generate the final predicted score value P. The regression head is comprised of one 3× 3 convolution and three fully-connected layers. Likewise, batch normalization, ReLU and dropout techniques are also employed in the regressor. The predicted score P is calculated as

$$P = RH(\hat{H}_{in}, W_{RH}), \tag{6.4}$$

where W_{RH} are the learnable weights of the regression head. Hence, the object function ($Loss$) of the whole network is formulated as:

$$Loss = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} ||P_i - GT_i||_2, \tag{6.5}$$

where N is the batch size and GT_i represents the ground truth of i_{th} input image.

6.3 Experimental Results and Analysis

The proposed MSINet for general-purpose BIQA was evaluated on the commonly-used popular image quality datasets LIVE [18], CSIQ [25], TID2013 [26], LIVEMD [98], and LIVEWC [27]. The datasets are introduced in Chapter 1 and Chapter 3. The PLCC and SRCC were selected as the evaluation criteria as described in Chapter 1.2.

In the implementation, the Adam [170] with the initial fixed learning rate of 0.0001 served as the optimizer for our MSINet, thanks to its higher convergence ability of learning. Since deep CNNs models usually require large amount of training samples, whereas most of the current image quality databases only contains a few hundreds annotated images, which is prone to overfitting. Random data augmentations of horizontal and vertical flip were performed. For each input image, a 336×336 patch was randomly cropped in an online manner. We train our model with Pytorch [171] on a single *GTX TITAN XP* GPU with a mini-batch size of 20. All experiments

Table 6.1: Performance comparison on five databases. The best results on each database are marked in boldface.

	LIVE		CSIQ		TID2013		LIVEMD		LIVEWC	
Metric	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Kang's CNN	0.9528	0.9507	0.7537	0.6845	0.6538	0.5592	0.9259	0.9304	0.5321	0.5118
BIECON	0.9602	0.9573	0.8235	0.8157	0.7654	0.7170	0.9322	0.9008	0.6177	0.5937
BPSQM	0.9665	0.9603	0.9142	0.8733	0.8852	0.8644	0.9052	0.8847	–	–
DIQaM-NR	0.9708	0.9625	0.8977	0.8634	0.8536	0.8359	0.9357	0.9022	0.6053	0.6009
RankIQA	0.9814	0.9736	0.9157	0.8944	0.7966	0.7813	0.9247	0.9088	0.6748	0.6442
MEON	0.9562	0.9487	0.9422	0.9310	0.8223	0.8114	0.9344	0.9243	0.6955	0.6871
DIQA	0.9703	0.9618	0.9125	0.8842	0.8572	0.8266	0.9318	0.9206	0.7029	0.7006
DistNet-Q3	0.9547	0.9520	0.9133	0.8874	0.8253	0.7946	0.8926	0.8472	0.6034	0.5728
NSSADNN	**0.9834**	**0.9817**	0.9257	0.8944	**0.9175**	0.8466	0.9487	0.9413	0.8122	0.7455
DB-CNN	0.9688	0.9658	0.9572	0.9437	0.8653	0.8155	0.9327	0.9211	**0.8603**	**0.8517**
MSINet (ours)	0.9743	0.9769	**0.9613**	**0.9623**	0.9051	**0.9024**	**0.9641**	**0.9688**	0.8438	0.8331

were conducted ten times to avoid the bias of randomness and the average results of evaluation criteria were reported.

6.3.1 Quantitative Evaluation

We compared our proposed MSINet with ten leading learning-based general-purpose BIQA approaches on five datasets, see Table 6.1, including Kang's CNN [69], BIECON [72], BPSQM [73], DIQaM-NR [74], RankIQA [75], MEON [76], DIQA [77], DistNet-Q3 [78], NSSADNN [79], and DB-CNN [80]. We employed the train-test procedure that introduced in Section 3.3.1.

It can be observed from Table 6.1 that our proposed method outperforms all the competing approaches on the CSIQ and LIVEMD datasets as indicated by the correlation coefficients PLCC and SRCC. Besides, the best SRCC is also produced on TID2013 while the PLCC value is only somewhat lower than that of NSSADNN. Although the performance of our MSINet on the LIVE and LIVEWC datasets are slightly behind NSSADNN and DB-CNN respectively, the overall performance of our method is still the best. In comparison, NSSADNN only obtains the best performance on the LIVE dataset and the same situation goes for DB-CNN on the LIVEWC dataset. Although NSSADNN attains the best PLCC on the TID2013 dataset, it produce rather inferior SRCC results compared with the state-of-the-arts. Moreover, both NSSADNN and DB-CNN report "noticeable worse" performance over some datasets, such as NSSADNN on the CSIQ and LIVEWC datasets, and DB-CNN on the TID2013 and LIVEMD datasets. In particular, the SRCC value on TID2013 (4.4%) and the PLCC value (1.6%) and SRCC value (2.9%) on LIVEMD are improved by our method with a clear margin compared to the second-ranked approaches. Furthermore, through Table 6.1 and Table 5.1, we can notice that the learning-based approaches generally perform much better than the traditional heuristic-based ones, which is expected and demonstrates the high capability of CNNs in representation and feature learning. More discussions about the two methodologies are given in Chapter 7.2.

As we described above, in the experimental setting, the input images were randomly cropped into patches. Here we examine the influence of the form of input samples (i.e., cropping or no cropping) on the performance of our MSINet. It can be seen from Table 6.2 that the input images under the "no cropping" setting generally produce a slightly worse performance across the databases, which is understandable since the cropping operation provides more samples. We can further notice that, the performance on the LIVE database under the "cropping" setting is inferior than the "no cropping" setting. Our hypothesis is that the LIVE database contains imbalance image samples, e.g., there are 175 and 145 images under the JPEG compression

Table 6.2: Impact of the form of input samples. SRCC results on databases are reported.

	LIVE	CSIQ	TID2013	LIVEMD	LIVEWC
Proposed	0.9769	0.9623	0.9024	0.9688	0.8331
w/o cropping	0.9797	0.9617	0.9005	0.9662	0.8233

"w/o M" means our proposed model without component M

and Gaussian blur types of distortion respectively. Nevertheless, our proposed BIQA method consistently performs well without the cropping operation, which indicates the robustness of the MSINet model.

As in the previous chapter, we further analyzed the statistical significance of the proposed model using F-test [12]. Description about the F-test can be found in Section 3.3.1. Fig. 6.5 shows the F-test results between nine compared metrics and the proposed MSINet model on the aforementioned five databases. Table 6.3 lists the statistical significance between the proposed method and the compared metrics.

It can be observed from Fig. 6.5 that our MSINet produces the smallest prediction errors among all the compared metrics in CSIQ and LIVEMD. In LIVE and TID2013, only NSSADNN produces slightly smaller prediction errors than MSINet. In LIVEWC, DB-CNN reports smaller prediction errors than the proposed model. As described in Chapter 3 and Chapter 1, LIVEMD [98] contains images under multiple distortions, and LIVEWC [27] is a real camera image database which involves realistic distortions. Our proposed MSINet produces the smallest prediction errors than most of the compared models on these two databases, which is quite impressive.

It is clear from Table 6.3 that the proposed MSINet model is superior to the existing approaches in most cases. Among the 44 combinations of methods and databases,

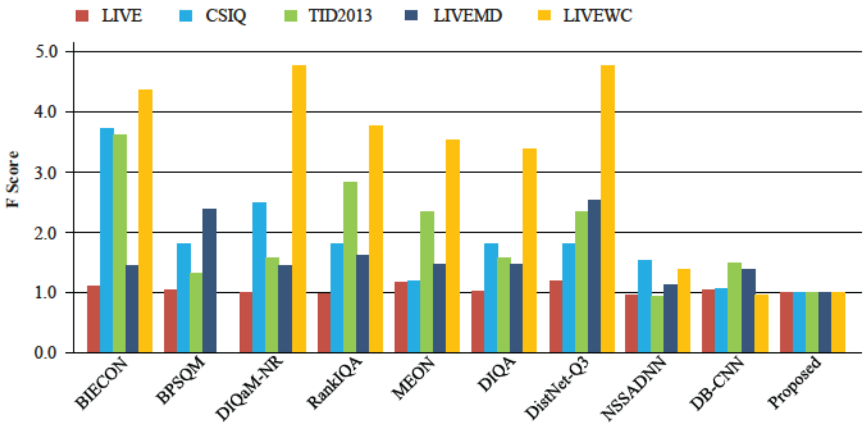


Figure 6.5: F-test results of nine existing approaches and our proposed MSINet.

Table 6.3: Statistical performance between the proposed MSINet and nine compared approaches on five databases.

Metric	LIVE	CSIQ	TID2013	LIVEMD	LIVEWC
BIECON	+1	+1	+1	+1	+1
BPSQM	0	+1	+1	+1	–
DIQaM-NR	0	+1	+1	+1	+1
RankIQA	0	+1	+1	+1	+1
MEON	+1	+1	+1	+1	+1
DIQA	0	+1	+1	+1	+1
DistNet-Q3	+1	+1	+1	+1	+1
NSSADNN	–1	+1	–1	+1	+1
DB-CNN	0	0	+1	+1	-1

our model performs significantly better in 35 cases, comparable in 6 cases and worse in only 3 cases. This indicates that the proposed method performs consistently well across all the five databases.

6.3.2 Ablation Study

In order to demonstrate the effectiveness of each component in the proposed MSINet and to better understand the impacts of these components, we conducted a series of

Table 6.4: Ablation study results of MSI modules.

Kernels	1×1	3×3	5×5	7×7	9×9	11×11	Proposed
PLCC	0.9260	0.9285	0.9053	0.9053	0.9144	0.9077	0.9613
SRCC	0.9167	0.9244	0.9160	0.9037	0.9033	0.9219	0.9623

ablation studies on the CSIQ database to empirically single out each contributor in this BIQA model.

First, we ablated the impacts of the proposed multi-scale integration on the final performance. The groups of multi-branch convolutions in each MSI module were configured with the same fixed kernel size in the set of $\{1 \times 1, 3 \times 3, \cdots, 11 \times 11\}$. By conducting such experiments, we can not only be aware of the effectiveness of multi-scale features aggregation, but also take a closer look at the importance of different scales on the prediction performance.

The empirical results shown in Table 6.4 demonstrate that when considering diverse scale information, our model achieves much better performance owing to the parallel multi-branch convolutions with distinct kernel sizes (see Fig. 6.2), which force the model to extract fine-grained multi-scale representations with varying receptive fields. This is consistent with the observation in the prior work [161]. It is proved beneficial to use different kernel sizes, and we attribute the reason to the aggregation of multi-scale information. Furthermore, we can find that the performance of networks configured by pure 1×1 or 3×3 kernels outperforms those with larger kernel sizes as large kernels may cause the failure of localizing local-oriented distortions. This observation coincides with the phenomenon that a wide range of the prevailing CNNs (i.e., VGGNet) are inclined to adopt vast kernels with small sizes instead of larger ones.

Table 6.5: Ablation study results of multi-level semantics reuse.

Level(s)	Low+High	Middle+High	High	Proposed
PLCC	0.9198	0.9271	0.9217	0.9613
SRCC	0.9197	0.9187	0.9168	0.9623

Table 6.6: Ablation study results of channel-wise attention.

	w/o attention	Proposed
PLCC	0.9382	0.9613
SRCC	0.9227	0.9623

Then, the ablation study on multi-level semantics reuse was also carried out to have an insight into the impacts of features at multiple levels of granularities. The ablative results shown in Table 6.5 illustrate that the multi-level strategy (low-, middle- and high-level features) indeed plays a pivotal role for the outstanding performance of the model. It can be seen that the removal of features at low- or middle-level seriously degrades the performance of the whole network. Additionally, it seems that the middle-level representations are vital for the overall impact of this multi-level strategy since removing them brings slightly worse results than only utilizing high-level features. This might be the reason that middle layers in our model allows producing more discriminative clues which are beneficial for combating overfitting. The experimental results further verify our hypothesis that multi-level semantics are important in capturing multi-granularity cues for the assessment of various distortion types.

Finally, the impacts of channel-wise recalibration on the regression accuracy were investigated to demonstrate the necessity of attention mechanism in our model design. The experimental results in Table 6.6 demonstrate the effectiveness of “attention along channel dimension”, as the channel-wise attention is capable of enhancing the model to learn more discriminative multi-level characteristics.

6.4 Summary

In this chapter, a multi-scale integration network (MSINet) is introduced, which provides an end-to-end solution for general-purpose BIQA. The proposed MSINet features the pretrained VGGNet-based frontend, multi-scale integration module (MSI), multi-level semantics reuse, and channel-wise attention mechanism. Specifically, the MSI module is devised to characterize fine-grained information at various scales, and allows the network to be equipped with scale-invariant property. A pattern of multi-level semantics reuse is proposed to make full use of features from previous layers. Furthermore, the channel-wise attention mechanism is designed, so as to adaptively recalibrate the channels from multiple levels to learn more discriminative features. Experimental results on several image quality datasets demonstrate the superiority of our proposed MSINet compared with the state-of-the-art learning-based BIQA methods.

Chapter 7

Conclusion and Discussion

7.1 Conclusion

Subjective quality ratings for digital images cannot be performed in real-time applications and are usually quite expensive and inefficient. This thesis has focused on designing novel and effective BIQA metrics for natural images. Specifically, Chapter 3 and Chapter 4 each presents a distortion-specific BIQA metric, one for blur (conventional) and the other one for gamut mapping (unconventional) distortions, whereas Chapter 5 and Chapter 6 aim at developing general-purpose BIQA methods.

In Chapter 3, sharpness-aware features are extracted based on the discrepancies of orientation selectivity-based visual patterns and log-Gabor filter responses between the input image and its reblurred version. Considering the influence of viewing distance on image quality, global sharpness discrepancy is measured through inter-resolution self-similarities. The proposed blind image sharpness evaluator demonstrates that these discrepancy measures are effective indicators for quantifying image blurriness.

In Chapter 4, two categories of statistics are analyzed in the design of BIQA metric for gamut-mapped images. Specifically, the local statistical features are used to portray structural and color distortions, and features extracted from global statistics are utilized to characterize the naturalness of image. This chapter provides the first attempt to blindly quantify the gamut mapping distortion. To further validate its effectiveness, the proposed metric has been applied for benchmarking gamut mapping algorithms.

In Chapter 5, second-order statistical features are investigated in multiple scales from the joint distribution of adjacent subband coefficients in the wavelet domain and the histogram of Gaussian derivative pattern in the spatial domain respectively. Experimental results of the proposed general-purpose BIQA metric over several datasets demonstrate that second-order image statistics are more robust in the measurement of various image distortions compared with those first-order statistics-based approaches.

Unlike these three heuristics-based methods, Chapter 6 introduces a learning-based BIQA method which does not involve engineered design of quality-aware features. In the proposed MSINet network model, multi-scale integration modules and multi-level supervision mechanism are delicately structured. This fully data-driven method provides an end-to-end framework for general-purpose BIQA, which demonstrates the strong capability of CNNs in feature representations.

7.2 Discussion

The BIQA models presented in Chapter 3 and Chapter 5 share a same two-stage structure: 1) extraction of quality-aware features; 2) a nonlinear regression function is learned through machine learning tools (e.g., support vector regression (SVR) [85]) from training images with the ground truth. The metric proposed in Chapter 4 does not need ground-truth subjective scores in the model training, thus is an unsupervised BIQA method. These three BIQA metrics are all heuristic-based, which means their performance greatly depends on the relevance of utilized features to visual quality perception. Hence, identifying a set of quality-aware features that are able to properly mimic the characteristics of the human visual system is of particularly importance. In the thesis, statistics-based features are largely employed, with the fact that high quality natural images obey some kind of statistical regularities while quality degradations can be deviated from these statistics. The proposed MSINet model in Chapter 6 provides a learning-based general-purpose BIQA method, which jointly optimizes feature representation and quality prediction. From the experimental results shown in Chapter 6 and Chapter 5, we can observe that MSINet produces much better performance than the heuristic-based method, which demonstrates the high capability of deep neural networks in feature learning.

Objective IQA of natural images is an ill-posed problem. The development effective BIQA metrics can be quite challenging since the problem itself is multi-disciplinary. It involves physiology, psychology, vision science and engineering [15]. In addition, the understanding of the HVS mechanism is still limited and there are large variability of image contents and distortions. In this thesis, two methodologies have been explored in the development of effective BIQA approaches, both heuristic-based and learning-based. The heuristic-based methods have the advantages of low computational cost, small data scale required in the training, etc., but the development of this type of methods needs prior knowledge about image distortions and highly engineered feature designs. On the other hand, the learning-based approaches benefit from fully data-driven, and those neural networks could perform feature learning automatically and deliver powerful feature representations. However, this kind of methods are mostly computation intensive and in the need of vast amount of training data. Moreover, the deep neural network-based methods are generally considered as “black box”, which makes them quite difficult in terms of explicit feature analysis.

In recent years, deep neural networks have gained much attention in the research community and achieved great success on various computer vision tasks. While this learning-based methodology is powerful, the requirement of high computational overhead may limit their application scope, e.g., running on low-cost mobile devices. In addition, for quality assessment of some unconventional distortions with limited training samples, it is still meaningful to investigate heuristic-based methods.

Although four BIQA metrics have been proposed and their performance has been validated over several public databases in this thesis, they are still vulnerable in the case of customer-facing applications. As we discussed in Section 1.2 and throughout the thesis, the authentic distortions that occupied often in practical scenarios are quite complex and there might be multiple types of distortion occurred in an image at the same time, which makes them very hard to measure. More balanced and comprehensive image quality datasets are needed in this regard. Since setting up a physical psycho-visual test environment and recruiting a large group of people performing the subjective ratings would be very costly, one possible solution is resorting to the crowdsourcing technique, e.g., Amazon Mechanical Turk[1]. In this case, proper guidelines and instructions about the “labelling” as suggested by the ITU recommendation (see Section 1.1) should be given to the workers.

[1] *: https://www.mturk.com/.

Chapter 8

Future Work

There are several possible directions where the current work can be extended. The related research topics are listed as follows for further study.

Blind Video Quality Assessment. Compared to IQA, the development of video quality assessment metrics needs to take the perceptual spatiotemporal characteristics into consideration. One simple solution is applying a current IQA metric on a frame-to-frame basis and then averaging the scores. However, this strategy may not work well when the video contents incorporate large motion [172]. It would be interesting to extend the proposed BIQA methods to blind video quality assessment by investigating the temporal structure and temporal features.

Image Aesthetics Assessment. Although the assessment of image aesthetics can be viewed as a classification or regression problem [173], it is quite different from IQA tasks. One crucial distinction is that the aesthetic quality of an image is greatly influenced by common photographic rules like "rule of thirds" and visual balance [174]. Potential applications of this technique include image recommendation, personalized photo album management and aesthetics-based image cropping [175]. While directly applying the proposed BIQA approaches to image aesthetic assessment may produce misleading results, it is believed that the multi-scale integration modules and multi-level supervision mechanism introduced in Chapter 6 would benefit the network model design of this task.

IQA Guided Image Synthesis. In recent years, the most prominent approach to image synthesis is based on generative adversarial network (GAN) [176] which consists of a generative network and a discriminative network. Related representative work in this field includes StackGAN [177] (text-to-image synthesis) and DualGAN [178] (image-to-image translation). However, the current GANs for image generation usually suffer from one critical issue, that is, the learned distribution is prone to poor quality samples. Attempts have been made to tackle this issue, such as building prior for underlying data distribution [179].

It is natural to think about employing IQA metrics as cost functions or regularizers in GANs' objective functions, since the generated images should preserve the local structural and statistical characteristics. Nevertheless, the mathematic formulation of these metrics are mostly not suitable being directly applied in the optimization framework due to their inherent properties such as non-convexity. In a pioneer work [180], valid regularizers were derived from the popular SSIM [18] and NIQE [66] indexes, and the experimental results proved their effectiveness. One promising future direction is to modify the proposed heuristic-based metrics, so that it can be utilized in guiding the generative network towards high-quality images.

Bibliography

[1] H. Liu, K.-K. Huang, C.-X. Ren, Y.-F. Yu, and Z.-R. Lai. Quadtree coding with adaptive scanning order for space-borne image compression. *Signal Process. Image Commun.*, 55:1–9, 2017.

[2] A. Kadaikar, G. Dauphin, and A. Mokraoui. Joint disparity and variable size-block optimization algorithm for stereoscopic image compression. *Signal Process. Image Commun.*, 61:1–8, 2018.

[3] L. Ma, D. Zhao, and W. Gao. Learning-based image restoration for compressed images. *Signal Process. Image Commun.*, 27(1):54–65, 2012.

[4] H. Wang, A.T.S. Ho, and S. Li. A novel image restoration scheme based on structured side information and its application to image watermarking. *Signal Process. Image Commun.*, 29(7):773–787, 2014.

[5] P. Ye. *Feature learning and active learning for image quality assessment.* Diss. Univ. Maryland, 2014.

[6] A. C. Bovik. Automatic prediction of perceptual image and video quality. *Proceeding of the IEEE*, 101(9):2008–2024, 2013.

[7] F. Porikli *et al.*. Multimedia Quality Assessment [DSP Forum]. *IEEE Signal Process. Mag.*, 28(6):164–177, 2011.

[8] ITU-R. Methodology for the subjective assessment of the quality of television pictures. Recommendation BT.500–11, International Telecommunication Union, 2002.

[9] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk. Comparison of four subjective methods for image quality assessment. *Computer Graphics Forum*, 31(8):2478–2491, 2012.

[10] P. Zolliker, Z. Barańczuk, I. Sprow, and J. Giesen. Conjoint analysis for evaluating parameterized gamut mapping algorithms. *IEEE Trans. Image Process.*, 19(3):758–769, 2010.

[11] ITU-R. Subjective video quality assessment methods for multimedia applications. Recommendation P.910, International Telecommunication Union, 2008.

[12] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11):3441–3452, 2006.

[13] C. Cheadle *et al.*. Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics*, 5(2):73–81, 2003.

[14] L. Li, Y. Yan, Z. Lu, J. Wu, K. Gu, and S. Wang. No-reference quality assessment of deblurred images based on natural scene statistics. *IEEE Access*, 5:2163–2171, 2017.

[15] D. M. Chandler. Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Process.*, 2013:1–53, 2013.

[16] L. Li, H. Cai, Y. Zhang, W. Lin, A. C. Kot, and X. Sun. Sparse representation-based image quality index with adaptive sub-dictionaries. *IEEE Trans. Image Process.*, 25(8):3775–3786, 2016.

[17] Z. Wang and A. C. Bovik. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.*, 26(1):98–117, 2009.

[18] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[19] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Trans. Image Process.*, 15(2):430–444, 2006.

[20] R. Soundararajan and A. C. Bovik. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Trans. Image Process.*, 21(2):517–526, 2012.

[21] L. Li, Y. Yan, Y. Fang, S. Wang, L. Tang, and J. Qian. Perceptual quality evaluation for image defocus deblurring. *Signal Process. Image Commun.*, 48:81–91, 2016.

[22] P. Zhao, L. Li, and H. Cai. Saliency guided gradient similarity for fast perceptual blur assessment. *IEICE Trans. Inf. Syst.*, E98-D(8):1613–1616, 2015.

[23] L. Li, Y. Zhou, J. Wu, W. Lin, and H. Li. GridSAR: grid strength and regularity for robust evaluation of blocking artifacts in JPEG images. *J. Vis. Commun. Image Represent.*, 30:153–163, 2015.

[24] S. Lee and S. J. Park. A new image quality assessment method to detect and measure strength of blocking artifacts. *Signal Process. Image Commun.*, 27(1):31–38, 2012.

[25] E. C. Larson and D. M. Chandler. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imag.*, 19(1):011006, 2010.

[26] N. Ponomarenko *et al.*. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.*, 30:57–77, 2015.

[27] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.*, 25(1):372–387, 2016.

[28] T. Virtanen, M. Nuutinen, M. Vaahteranoksa, P. Oittinen, and J. Häkkinen. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 24(1):390–402, 2015.

[29] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Trans. Image Process.*, 20(5):1185–1198, 2011.

[30] H. Cai, M. Wang, W. Mao, and M. Gong. No-reference image sharpness assessment based on discrepancy measures of structural degradation. *J. Vis. Commun. Image Represent.*, 71:102861, 2020.

[31] H. Cai, L. Li, Z. Yi, and M. Gong. Blind quality assessment of gamut-mapped images via local and global statistical analysis. *J. Vis. Commun. Image Represent.*, 61:250–259, 2019.

[32] H. Cai, L. Li, Z. Yi, and M. Gong. Towards a blind image quality evaluator using multi-scale second-order statistics. *Signal Process. Image Commun.*, 71:88–99, 2019.

[33] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot. Image sharpness assessment by sparse representation. *IEEE Trans. Multimedia*, 18(6):1085–1097, 2016.

[34] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi. Perceptual blur and ringing metrics: Application to JPEG2000. *Signal Process. Image Commun.*, 19(2):163–172, 2004.

[35] R. Ferzli and L. J. Karam. A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *IEEE Trans. Image Process.*, 18(4):717–728, 2009.

[36] N. D. Narvekar and L. J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Trans. Image Process.*, 20(9):2678–2683, 2011.

[37] P. V. Vu and D. M. Chandler. A fast wavelet-based algorithm for global and local image sharpness estimation. *IEEE Signal Process. Lett.*, 19(7):423–426, 2012.

[38] C. T. Vu, T. D. Phan, and D. M. Chandler. S3: A spectral and spatial measure of local perceived sharpness in natural images. *IEEE Trans. Image Process.*, 21(3):934–945, 2013.

[39] Q. B. Sang, H. X. Qi, X. J. Wu, C. F. Li, and A. C. Bovik. No-reference image blur index based on singular value curve. *J. Vis. Commun. Image Represent.*, 25(7):1625–1630, 2014.

[40] R. Hassen, Z. Wang, and M. Salama. Image sharpness assessment based on local phase coherence. *IEEE Trans. Image Process.*, 22(7):2798–2810, 2013.

[41] K. Bahrami and A. C. Kot. A fast approach for no-reference image sharpness assessment based on maximum local variation. *IEEE Signal Process. Lett.*, 21(6):751–755, 2014.

[42] K. Gu, G. T. Zhai, W. S. Lin, X. K. Yang, and W. J. Zhang. No-reference image sharpness assessment in autoregressive parameter space. *IEEE Trans. Image Process.*, 24(10):3218–3231, 2015.

[43] L. Li, W. Lin, X. Wang, G. Yang, K. Bahrami, and A. C. Kot. No-reference image blur assessment based on discrete orthogonal moments. *IEEE Trans. Cybernetics*, 46(1):39–50, 2015.

[44] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang. No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features. *IEEE Trans. Multimedia*, 19(5):1030–1040, 2016.

[45] Y. Zhan and R. Zhang. No-reference image sharpness assessment based on maximum gradient and variability of gradients. *IEEE Trans. Multimedia*, 20(7):1796–1808, 2018.

[46] S. Yu, S. Wu, L. Wang, F. Jiang, Y. Xie, and L. Li. A shallow convolutional neural network for blind image sharpness assessment. *PloS One*, 12(5):e0176632, 2017.

[47] S. Yu, F. Jiang, L. Li, and Y. Xie. CNN-GRNN for image sharpness assessment. In: *Proc. Asian Conf. Comput. Vis. (ACCV)*, pp. 50–61, 2016.

[48] M. S. Hosseini, Y. Zhang, and K. N. Plataniotis. Encoding visual sensitivity by MaxPol convolution filters for image sharpness assessment. *IEEE Trans. Image Process.*, 28(9):4510–4525, 2019.

[49] Jens. Preiss. *Color-image quality assessment: From metric to application.* Diss. Technische Universität, 2015.

[50] J. Morovič. *Color gamut mapping.* John Wiley &Sons, 2008.

[51] L. A. Taplin and G. M. Johnson. When good hues go bad. In *Conference on Colour in Graphics, Imaging, and Vision*, pp. 348–352, 2004.

[52] C. Witzeland K. Gegenfurtner. Memory color. In *Luo, R. (eds) Encyclopedia of Color Science and Technology*, Springer, New York, pp. 1–7, 2013.

[53] X. Zhang and B. A. Wandell. A spatial extension of CIELAB for digital color image reproduction. *Society for Information Display*, 5(1):61–63, 1997.

[54] J. Y. Hardeberg, E. Bando, and M. Pedersen. Evaluating colour image difference metrics for gamut mapping images. *Coloration Technology*, 124(4):243–253, 2008.

[55] M. D. Fairchild and G. M. Johnson. The iCAM framework for image appearance, image differences, and image quality. *J. Electron. Imaging*, 13:126–138, 2004.

[56] J. Morovič. Guidelines for the evaluation of gamut mapping algorithms. *Commission Internationale de l'Eclairage (CIE)*, 153(D8-6), 2003.

[57] I. Lissner, J. Preiss, P. Urban, M. S. Lichtenauer, and P. Zolliker. Image-difference prediction: From grayscale to color. *IEEE Trans. Image Process.*, 22(2):435–446, 2013.

[58] A. K. Moorthy and A. C. Bovik. A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.*, 17(5):513–516, 2010.

[59] A. K. Moorthy and A. C. Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.*, 20(12):3350–3364, 2011.

[60] M. A. Saad, A. C. Bovik, and C. Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.*, 21(8):3339–3352, 2012.

[61] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

[62] Y. Zhang and D. M. Chandler. No-reference image quality assessment based on log-derivative statistics of natural scenes. *J. Electron. Imag.*, 22(4):043025-1, 2013.

[63] L. Liu, B. Liu, H. Huang, and A. C. Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Signal Process., Image Commun.*, 29(8):856–863, 2014.

[64] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.*, 23(11):4850–4862, 2014.

[65] K. Gu, G. Zhai, X. Yang, and W. Zhang. Using free energy principle for blind image quality assessment. *IEEE Trans. Multimedia*, 17(1):50–63, 2015.

[66] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[67] L. Zhang, L. Zhang, and A. C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Trans. Image Process.*, 24(8):2579–2591, 2015.

[68] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 995–1002, 2013.

[69] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1733–1740, 2014.

[70] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2016.

[71] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans. Image Process.*, 26(8):3951–3964, 2017.

[72] J. Kim and S. Lee. Fully deep blind image quality predictor. *IEEE J. Select. Topics Signal Process.*, 11(1):206–220, 2017.

[73] D. Pan, P. Shi, M. Hou, Z. Ying, S. Fu, and Y. Zhang. Blind predicting similar quality map for image quality assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6373–6382, 2018.

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

[75] X. Liu, J. van Weijer, and A. Bagdanov. RankIQA: Learning from rankings for no-reference image quality assessment. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1040–1049, 2017.

[76] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Process.*, 27(3):1202–1213, 2017.

[77] J. Kim, A.-D. Nguyen, and S. Lee. Deep CNN-based blind image quality predictor. *IEEE Trans. Neural Networks Learning Syst.*, 30(1):11–24, 2018.

[78] S. V. Dendi, C. Dev, N. Kothari, and S. S. Channappayya. Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment. *IEEE Signal Process. Lett.*, 26(1):89–93, 2018.

[79] B. Yan, B. Bare, and W. Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Trans. Multimedia*, 21(10):2603–2615, 2019.

[80] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.*, 30(1):36–47, 2020.

[81] T. D. Albright. Direction and orientation selectivity of neurons in visual area MT of the macaque. *J. Neurophysiol.*, 52(6):1106–1130, 1984.

[82] J. Wu, W. Lin, G. Shi, L. Li, and Y. Fang. Orientation selectivity based visual pattern for reduced-reference image quality assessment. *Info. Sci.*, 351:18–29, 2016.

[83] D. Field and N. Brady. Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. *Vis. Res.*, 37(23):3367–3383, 1997.

[84] K. Gu, M. Liu, G. T. Zhai, X. K. Yang, and W. J. Zhang. Quality assessment considering viewing distance and image resolution. *IEEE Trans. Broadcast.*, 61(3):520–531, 2015.

[85] A. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[86] A. Skodras, C. Christopoulos, and T. Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal Process. Mag.*, 18(5):36–58, 2001.

[87] K. Gu, J. Qiao, S. Lee, H. Liu, W. Lin, and P. Le Callet. Multiscale natural scene statistical analysis for no-reference quality evaluation of DIBR-synthesized views. *IEEE Trans. Broadcast.*, 66(1):127–139, 2019.

[88] J. Wu, W. Lin, G. Shi, Y. Zhang, W. Dong, and Z. Chen. Visual orientation selectivity based structure description. *IEEE Trans. Image Process.*, 24(11):4602–4613, 2015.

[89] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160(1):106–154, 1962.

[90] D. Hansel and C. van Vreeswijk. The mechanism of orientation selectivity in primary visual cortex without a functional map. *J. Neurosci.*, 32(12):4049–4064, 2012.

[91] F. W. Campbell and J. J. Kulikowski. Orientational selectivity of the human visual system. *J. Physiol.*, 187(2):437–445, 1966.

[92] S. Marcelja. Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Amer.*, 70(11):1297–1300, 1980.

[93] D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Amer.*, 4(12):2379–2394, 1987.

[94] B. A. Olshausen and D. J. Field. How close are we to understanding V1? *Neural Comput.*, 17(8):1665–1699, 2005.

[95] D. L. Ruderman. The statistics of natural images. *Network: Comput. Neural Syst.*, 5(4):517–548, 1994.

[96] K. Sharifi and A. Leon-Garcia. Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Trans. Circuits Syst. Video Technol.*, 5(1):52–56, 1995.

[97] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Symp. Technol.*, 2(3):27, 2011.

[98] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik. Objective quality assessment of multiply distorted images. In: *Asilomar Conference on Signals, Systems and Computers*, pp. 1693–1697, 2012.

[99] A. Ciancio *et al.*. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process.*, 20(1):64–75, 2011.

[100] E. Bando, J. Hardeberg, and D. Connah. Can gamut mapping quality be predicted by colour image difference formulae? In *Proc. of the IS &T/SPIE Electronic Imaging: Human Vision and Electronic Imaging X*, pp. 180–191, 2005.

[101] Z. Barańczuk, P. Zolliker, and J. Giesen. Image-individualized gamut mapping algorithms. *J. Imag. Sci. Technol.*, 54(3):030201-1, 2010.

[102] U. Rajashekar, Z. Wang, and E. P. Simoncelli. Perceptual quality assessment of color images using adaptive signal representation. In *Proc. of the SPIE: Human Vision and Electronic Imaging XV*, pp. 75271L, 2010.

[103] G. J. Braun and M. D. Fairchild. General-purpose gamut-mapping algorithms: Evaluation of contrast-preserving rescaling functions for color gamut mapping. In *Color and Imaging Conference*, pp. 167–172, 1999.

[104] N.-E. Lasmar, Y. Stitou, and Y. Berthoumieu. Multiscale skewed heavy tailed model for texture analysis. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 2281–2284, 2009.

[105] D. L. Ruderman, T. W. Cronin, and C.-C. Chiao. Statistics of cone responses to natural images: Implications for visual coding. *J. Opt. Soc. Amer. A*, 15(8):2036–2045, 1998.

[106] M. Čadík and P. Slavík. The naturalness of reproduced high dynamic range images. In *Proc. 9th Int. Conf. Inf. Visual.*, pp. 920–925, 2005.

[107] S. N. Yendrikhovskij, F. J. J. Blommaert, and H. de Ridder. Color reproduction and the naturalness constraint. *Color Research* & *Application*, 24(1):52–67, 1999.

[108] Y. Gong and I. F. Sbalzarini. Gradient distribution priors for biomedical image processing. *arXiv preprint arXiv:1408.3300*, 2014.

[109] Y. Gong and I. F. Sbalzarini. Image enhancement by gradient distribution specification. In *Proc. Asian Conf Comput. Vis. (ACCV)*, pp. 47–62, 2014.

[110] D. Krishnan and R. Fergus. Fast image deconvolution using hyper-Laplacian priors. In *Advances in Neural Information Processing Systems*, pp. 1033–1041, 2009.

[111] Y. Gong and I. F. Sbalzarini. Local weighted Gaussian curvature for image processing. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 534–538, 2013.

[112] L. Bombrun, F. Pascal, J.-Y. Tourneret, and Y. Berthoumieu. Performance of the maximum likelihood estimators for the parameters of multivariate generalized gaussian distributions. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 3525–3528, 2012.

[113] P. Zolliker, Z. Barańczuk and J. Giesen. Image fusion for optimizing gamut mapping. In *Color and Imaging Conference*, pp. 109–114, 2011.

[114] K. Fukunaga. *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

[115] J. Giesen, E. Schuberth, K. Simon, P. Zolliker, and O. Zweifel. Image-dependent gamut mapping as optimization problem. *IEEE Trans. Image Process.*, 16(10):2401–2410, 2007.

[116] P. Zolliker and K. Simon. Retaining local image information in gamut mapping algorithms. *IEEE Trans. Image Process.*, 16(3):664–672, 2007.

[117] L. L. Thurstone. A law of comparative judgment. *Psychol. Rev.*, 34(4):273–286, 1927.

[118] J. Xu, P. Ye, Q. Li, Y. Liu, and D. Doermann. No-reference document image quality assessment based on high order image statistics. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 3289–3293, 2016.

[119] L. Kang, P. Ye, Y. Li and D. Doermann. A deep learning approach to document image quality assessment. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 2570–2574, 2014.

[120] Y. Liu, K. Gu, S. Wang, D. zhao, and W. Gao. Blind quality assessment of camera images based on low-level and high-level statistical features. *IEEE Trans. Multimedia*, 21(1):135–146, 2018.

[121] D. Huang, C. Zhu, Y. Wang, and L. Chen. HSOG: A novel local image descriptor based on histograms of the second-order gradients. *IEEE Trans. Image Process.*, 23(11):4680–4695, 2012.

[122] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[123] A. P. Johnson and C. L. Baker. First- and second-order information in natural images: A filter-based approach to image statistics. *Journal of the Optical Society of America A*, 21(6):913–925, 2004.

[124] C.-C. Su, L. K. Cormack, and A. C. Bovik. Oriented correlation models of distorted natural images with application to natural stereopair quality evaluation. *IEEE Trans. Image Process.*, 24(5):1685–1699, 2015.

[125] C.-C. Su, L. K. Cormack, and A. C. Bovik. Bivariate statistical modeling of color and range in natural scenes. In *Proc. of the SPIE: Human Vision and Electronic Imaging X*, pp. 90141G, 2014.

[126] G. S. Raghtate and S. S. Salankar. Comparison of second order statistical analysis and wavelet transform method for texture image classification. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 318–323, 2015.

[127] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multiscale transforms. *IEEE Trans. Inf. Theory*, 38(2):587–607, 1992.

[128] Q. Li and Z. Wang. Reduced-reference image quality assessment using divisive normalization-based image representation. *IEEE J. Select. Topics Signal Process.*, 3(2):202–211, 2009.

[129] Z. Sinno and A. C. Bovik. Generalizing a closed-form correlation model of oriented bandpass natural images. In *IEEE Glob. Conf. on Sig. and Info. Proc.*, 2015, pp. 373–377.

[130] Z. Sinno, C. Caramanis, and A. C. Bovik. Towards a closed form second-order natural scene statistics model. *IEEE Trans. Image Process.*, 27(7):3194–3209, 2018.

[131] M. J. Wainwright, O. Schwartz, and E. P. Simoncelli. Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. *Statist. Theories of the Brain*, pp. 203–222, 2002.

[132] S. Lyu. Dependency reduction with divisive normalization: Justification and effectiveness. *Neural Comput.*, 23(11):2942–2973, 2011.

[133] I. Area, E. Godoy, A. Ronveaux, and A. Zarzo. Bivariate second-order linear partial differential equations and orthogonal polynomial solutions. *J. Math. Anal. Appl.*, 387:1188–1208, 2012.

[134] B. Zhang, Y. Gao, S. Zhao, and J. Liu. Local derivative pattern versus local binary pattern: Face recognition with higher-order local pattern descriptor. *IEEE Trans. Image Process.*, 19(2):533–544, 2010.

[135] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[136] Q. Li, W. Lin, and Y. Fang. No-reference image quality assessment based on high order derivatives. In *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2016.

[137] Q. Li, W. Lin, J. Xu, and Y. Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Trans. Multimedia*, 18(12):2457–2469, 2016.

[138] Q. Li, W. Lin, and Y. Fang. No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Process. Lett.*, 23(4):541–545, 2016.

[139] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NIPS)*, 2012.

[140] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[141] C. Szegedy *et al.*. Going deeper with convolutions. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–9, 2015.

[142] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[143] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4700–4708, 2017.

[144] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards realtime object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*, pp. 91–99, 2015.

[145] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 770–778, 2016.

[146] U. Muhammad, W. Wang, S. P. Chattha, and S. Ali. Pre-trained VGGNet architecture for remote-sensing image scene classification. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pp. 3904–3908, 2017.

[147] P. Ballester and R. M. Araujo. On the performance of GoogLeNet and AlexNet applied to sketches. In *Proc. Thirtieth AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 1124–1128, 2016.

[148] P. Tang, H. Wang, and S. Kwong. G-MS2F: GoogLeNet based multi-stage feature fusion of deep CNN for scene recognition. *Neurocomputing*, 225:188–197, 2017.

[149] Z. Lu, X. Jiang, and A. Kot. Deep coupled resnet for low-resolution face recognition. *IEEE Signal Process. Lett.*, 25(4):526–530, 2016.

[150] W. Liu and K. Zeng. SparseNet: A sparse DenseNet for image classification. *arXiv preprint arXiv:1804.05340*, 2018.

[151] G. Li and Y. Yu. Visual saliency detection based on multiscale deep CNN features. *IEEE Trans. Image Process.*, 25(11):5012–5024, 2016.

[152] M. Huang, Z. Liu, L. Ye, X. Zhou, and Y. Wang. Saliency detection via multi-level integration and multi-scale fusion neural networks. *Neurocomputing*, 364:310–321, 2019.

[153] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang. Multi-scale convolutional neural networks for crowd counting. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 465–469, 2017.

[154] Z. Zou, Y. Cheng, X. Qu, S. Ji, X. Guo, and P. Zhou. Attend to count: Crowd counting with adaptive capacity multi-scale CNNs. *Neurocomputing*, 367:75–83, 2019.

[155] G. Huang *et al.*. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844*, 2017.

[156] T. Rao, X. Li, H. Zhang, and M. Xu. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing*, 333:429–439, 2019.

[157] M. He, B. Li, and H. Chen. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 3904–3908, 2017.

[158] Sindagi, Vishwanath A and Patel, Vishal M. Multi-level bottom-top and top-bottom feature fusion for crowd counting. *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

[159] M. Wang, J. Zhou, W. Mao, and M. Gong. Multi-scale convolution aggregation and stochastic feature reuse for DenseNets. *arXiv preprint arXiv:1810.01373*, 2018.

[160] F. Gao, J. Yu, Q. Huang, and Q. Tian. Blind image quality prediction by exploiting multi-level deep representations. *Pattern Recognition*, 81:432–442, 2018.

[161] Y. Fu *et al.*. Screen content image quality assessment using multi-scale difference of Gaussian. *IEEE Trans. Circuits Syst. Video Technol.*, 28(9):2428–2432, 2018.

[162] J. Deng *et al.*. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 248–255, 2009.

[163] J. Hu, L. Shen, and G. Sun Squeeze-and-excitation networks. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7132–7141, 2018.

[164] F. Wang *et al.*. Residual attention network for image classification. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3156–3164, 2017.

[165] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. Cbam: Convolutional block attention module. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 3–19, 2018.

[166] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8697–8710, 2018.

[167] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6848–6856, 2018.

[168] C. Ma, Y. Xu, B. Ni, and X. Yang. When correlation filters meet convolutional neural networks for visual tracking. *IEEE Signal Process. Lett.*, 23(10):1454–1458, 2016.

[169] L.-C. Chen, G. Papandreou, I. Kokkinos, K Murphy, and A. L. Yuille. MDeeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017.

[170] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Computer Science*, 2014.

[171] A. Paszke *et al.*. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (NIPS) workshop*, 2017.

[172] T. J. Liu, Y. C. Lin, W. Lin, and C.-C. Kuo. Visual quality assessment: recent developments, coding applications and future trends. *APSIPA Trans. Sig. Info. Process.*, 2, 2013.

[173] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 662–679, 2016.

[174] J. T. Lee, H. U. Kim, C. Lee, and C. S. Kim. Semantic line detection and its applications. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 3229–3237, 2017.

[175] Y. Deng, C. C. Loy, and X. Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Process. Mag.*, 34(4):80–106, 2017.

[176] I. Goodfellow *et al.*. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[177] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 5907–5915, 2017.

[178] Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2849–2857, 2017.

[179] S. Gu, J. Bao, D. Chen, and F. Wen. PriorGAN: Real data prior for generative adversarial nets. *arXiv preprint arXiv:2006.16990*, 2020.

[180] P. Kancharla and S. C. Channappayya. Quality aware generative adversarial networks. In *Advances in Neural Information Processing Systems*, pp. 2948–2958, 2019.